Audio Engineering Society

Convention Paper

Presented at the 127th Convention
2009 October 9–12 New York NY, USA

Which Wideband Speech Codec? Quality Impact due to Room-acoustics at Send Side and Presentation Method

Alexander Raake[1], Marcel Wältermann[1], and Sascha Spors[2]

[1]*Quality & Usability Lab, Deustche Telekom Laboratories, TU Berlin, Germany*

Correspondence should be addressed to Alexander Raake (alexander.raake@telekom.de)

ABSTRACT

We report on two listening tests to determine the speech quality of different wideband (WB) speech codecs. In the first test, we have studied various network conditions, including WB–WB and WB–narrowband (WB–NB) tandeming, packet loss, and background noise. In addition to other findings, this test showed some codec quality rank-order changes when compared to the literature. To evaluate the hypothesis that secondary test factors lead to this rank-order effect, we conducted another speech quality listening test: Here, we simulated different source material recording conditions (room-acoustics and microphone positions), processed the material with different WB speech coders, and presented the resulting files monotically in one and diotically in another test. The paper discusses why and how these factors impact speech quality.

1. INTRODUCTION

Extending the transmitted frequency range of telephone speech from narrowband (NB, 300-3400 Hz) to wideband (WB, 50-7000 Hz) enables a considerably higher speech quality. In previous work, we have identified this advantage to be of approximately 30% [1, 2]. This quality improvement can be achieved already at equal or lower bitrates as in the case of NB telephony, by employing corresponding wideband speech codecs: In digital public-switched, narrowband telephone networks, the logarithmic Pulse Code Modulation being used operates at 64 kbit/s (G.711, [3]); in mobile networks, the voice channel is of approximately 12 kbit/s bandwidth (with the employed codecs being the AMR-NB or GSM-EFR [4, 5]). Consequently, WB codecs such as the AMR-WB [6] or the older G.722 [7] are candidates that can be employed in WB services: At 12.65-23.85 kbit/s (G.722.2) and 64 kbit/s (G.722) these codecs achieve a much higher quality than their NB-counterparts [1, 2, 8].

When a new speech codec is to be standardized, a set of listening tests is typically conducted to compare its

speech quality with one or more reference codecs under different operation conditions, with performance thresholds being set that need to be passed. Similarly, the selection of an appropriate codec during the network planning phase requires a quantitative measure of speech quality that is (ultimately) based on listening tests.

The research described in this paper is motivated by the analysis of an extensive listening test conducted to determine the speech quality of different wideband speech codecs [9, 10]. There were multiple reasons for conducting this test, among others to further develop a wideband version of the so-called E-model [11]. The E-model is a parameter-based quality prediction model recommended by the ITU-T for network planning. The model's quality scale is the Transmission Rating Scale (R-scale). The E-model relies on the assumption that different types of degradations are additive in terms of the perceptual impairment they cause. This is reflected by its basic formula:

$$R = R_0 - \sum_i I_i. \quad (1)$$

Here, R is the Transmission Rating, expressed on the model's quality scale that ranges from 0 to $R_{0,max}$. For NB speech, the bandwidth the model was developed for initially, $R_{0,max} = 100$. Based on our previous work, this maximum range has been extended to WB, with $R_{0,max} = 129$ [1, 2, 12]. In Equation (1), R_0 reflects the base-quality that is related to the basic signal-to-noise-ratio; I_i are so-called impairment factors, which reflect different types of impairments (simultaneous to the speech, delayed to the speech, or related with the codec or additional transmission errors such as VoIP packet loss). The impairment caused by a codec is typically referred to as Equipment Impairment Factor $I_{e,NB/WB}$; $I_{e,NB}$-values for different NB codecs (and hence to be used with the NB-version of the E-model [11]) can be found in [13]. In previous work, we had determined the quality impairment introduced by different WB and NB speech codecs when expressed on the extended WB-R-scale [2]. The resulting now standardized $I_{e,WB}$-values can be found in [8].

Now, when compared with this standard literature, the new test motivating the present research lead to surprising results. One observation is particularly interesting: While the G.722.2 codec (AMR-WB) at 23.05 kbit/s and 23.85 kbit/s showed consistently higher quality than the G.722 at 64 kbit/s in almost all tests compiled from the literature in Möller et al. [2], the new tests yield a statistically significant advantage of the G.722, and hence a reversed rank-order between these codecs. Also, the quality-degradation with decreasing bitrate of the G.722.2 is stronger than expected from previous work. Since this issue is of high general relevance for the case that a codec is to be compared with a reference codec during codec standardization, or to be selected for a given network, we started a more systematic analysis. It investigates whether one or both of the key differences between the standardization tests and our tests are responsible for the rank-order effects. These differences are summarized in Table 1: Our tests have been conducted with anechoic recordings using diotic presentation, while the majority of previous tests were conducted using slightly reverberant recordings and monotic listening.

	Source material	Presentation
Standard. tests $I_{e,NB/WB}$ [8, 2]	Slightly reverberant [14]	Monotic (majority of labs)
T-Labs codec test [9, 10]	Anechoic	Diotic

Table 1: Differences considered responsible for codec rank-ordering effects between standardized impairment factor values $I_{e,NB/WB}$ [8, 2] and T-Labs Test 1 results [9, 10].

The paper is structured as follows: Section 2 describes the initial listening test (Test 1) highlighting the rank-order differences when compared with previous results; in Section 3, we establish a set of working hypotheses as a result of Test 1; Section 4 describes a second listening test (Test 2) conducted to verify the working hypotheses, and Section 5 concludes the paper with a description of the respective consequences for future wideband codec tests.

2. LISTENING TEST 1: CODEC TEST

In the first test [9, 10], different conditions of WB and NB speech codecs were assessed, namely:

1. in single and tandem operation,
2. under IP packet loss,
3. in the presence of background noise at send side.

2.1. Test 1 Set-up

In total, 114 test conditions plus 11 reference conditions were tested, using anechoic source recordings from four speakers (two female, two male). The conditions included WB-codecs such as clean PCM, the AMR-WB (ITU-T Rec. G.722.2), the G.722, and the G.729.1 [6, 7, 15], and NB-codecs such as the G.711, the G.729A, and the G.726 [3, 16, 17]. 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 noise and car noise, each at two different levels.

For the majority of tested WB codecs, a number of conditions involved uniform packet loss, with loss-rates from the set $0, 1, 2, 4, 8\%$. The ITU processing tools were used [18]. As opposed to classical tests with samples from different speakers being assessed during one test session, we have used one set of listening sessions per each of the four speakers. In all other respects the tests were conducted according to [19]. As sentence material, shortened versions of the German EUROM sentence material were used [20]. 38 sentences were selected from the available 40. Each sentence from each speaker was processed with the 125 conditions, yielding a total of $38 \cdot 4 \cdot 125 = 19000$ files.

For each listener, the playlists (one per speaker and per listener) were created by – for each condition – randomly selecting one of the available 38 sentences. The speech files were presented diotically using Sennheiser HD 25 headphones. The test was administered using 6 separate laptop computers each equipped with RME HDSP Cardbus Cards and RME HDSP Mulitface II Soundcards. Each subject could listen to each of the sound files only once, and gave ratings on a slider-based version of the 5-point ACR-scale [19] using a test GUI. The diotic presentation-level was 73 dB SPL. 120 paid subjects took part in Test 1 (appr. 50% female, 50% male; age 17 to 80 years).

In the analysis of the results, only 100 subjects were retained: The initial goal of the test was to collect both quality ratings on WB-relevant channels, but also to provide insight into the relation between quality perception and user group aspects (IT experience, age, etc., see [9]). The audiometric screening conducted in this context indicated that the user group consisting of the 20 oldest users needed to be excluded to yield more fine-grained results.

2.2. Test 1 Results

The MOS-data were transformed onto the WB E-model R-scale [0, 129], following a similar procedure as in [2], i.e.:

1. MOS-data [1,5] were transformed to E-Model's NB R-scale [0,100] using the transformation given in ITU-T Rec. G.107 (2008).
2. The R,nb-values were linearly transformed to obtain R,wb-values, using: $R,wb = 1.29 \cdot R,nb$.
3. From the R,wb-values, preliminary wideband equipment impairment factor values Ie,wb were calculated using $Ie,wb = 129 - R,wb$.
4. The Ie,wb-values were linearly normalized following (within limits) the approach described in [21].

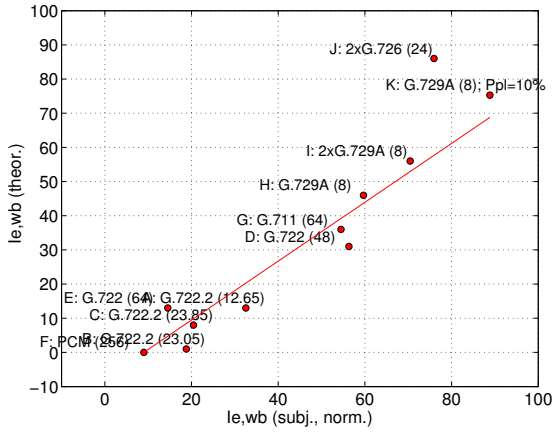

Fig. 1: Transformation to adjust the test results for conditions of known impairment to the values recommended in [13, 8]; the corresponding fitting-function is then applied to all test results to yield normalized values, following a method adopted from [21].

For Test 1, the results of step 4 are depicted in Figure 1, where the transformed test results are plotted in terms of preliminary WB equipment impairment factors $Ie,wb(\text{subj., norm})$ against the corresponding expected $Ie,wb(\text{theor.})$ taken from [8]. As can be seen from the figure, a straight line appears to be a good approximation of the relation between the two data sets. However, it can also be observed that there are some rank-order differences between the recommended values and the test data points, as will be discussed in more detail below. Note

that the G.726*G.726 tandem has been omitted from the transformation, since the large deviation from the general trend cast doubt on the additivity property or the recommended Ie-values in this case.

The plot in Figure 2 shows a comparison of the test results for single-coding conditions. The deviations of the transformed results from the R-values expected based on [12, 8] are highlighted: The blue boxes indicate the mean test results; the stacked boxes serve to indicate conditions yielding lower quality ratings than expected (pink), and conditions with higher quality than expected (purple).

A first observation is that the test confirms the quality advantage of WB over NB of more than 35 points on the 129-point WB R-scale; the E-model prediction for a clean G.711 channel is $R_{NB} = 93.2$.

The observation most relevant for the research reported in this paper 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 [22, 8]; see bars 4 and 9). In addition, for the G.722.2, the quality-decrease with decreasing bitrate is stronger than expected (bars 3-8). In contrast, the G.722 is rated better than expected, at least at the two higher bitrates (bars 9 and 10).

3. WORKING HYPOTHESES FOR CODEC RANK-ORDERING EFFECT

Wältermann et al. (e.g. [23]) have investigated the perceptual dimensions underlying the speech quality of transmitted speech. This and similar work by others indicates which perceptual features underly the quality judgments made by listeners. These multidimensional considerations together with an expert-listening to different speech samples lead us to the hypothesis, that different secondary factors (with the network conditions considered as the primary factors) may differently highlight certain of the perceptual features related with speech coding algorithms such as the G.722 and the G.722.2:

- Due to its ADPCM-type algorithm (sub-band adaptive pulse-code modulation), the G.722 introduces a low-level but audible wideband noise ("noisiness" dimension).
- The spectral distortion the G.722 introduces is less expressed than in case of the G.722.2 at 23.05 and 12.65 kbit/s, which sound less "full" than the G.722 ("coloration" dimension).
- Due to its ACELP-type algorithm (Algebraic Code-Excited Linear Prediction), the G.722.2 yields a certain non-linear distortion, which probably pertains, in perceptual terms, to the dimensions "bubbling" found for NB-codecs by [24] and "discontinuity" as mentioned in [23].

The decomposition of the codec-distortion into a linear and a non-linear component was initially introduced by Schüssler [25], and is the basis for several speech quality prediction algorithms. In the context of quantifying speech quality of wideband speech codecs in terms of E-model impairments, this decomposition was reconsidered by Wältermann and Raake [26]. It can be assumed that secondary factors such as the recording set-up may differently interact with the linear and non-linear components and corresponding perceptual features.

In a first, informal expert listening session comparing the test material used in former tests and the material used in our codec test, we have identified the following secondary factors as candidates for the above-mentioned interactions:

(A) The set-up chosen for the source-recordings determines, among other aspects, the room-information contained in the recordings. This is of relevance for our case, since the recordings used in the tests compiled to obtain the standardized values in [8] (see [2]) were mostly made in acoustically dry but audibly reverberant rooms, with microphone distances from the mouth equal to or greater than 14 cm (cf. [14]). In turn, our source files were recorded in a non-reverberant environment. In a previous study, [27] have reported a masking of audio codec distortions by the room acoustics at receive side; the question is whether a similar effect may occur for the case of speech and reverberation at send side.

(B) Monotic versus diotic presentation may differently highlight the above-mentioned perceptual dimensions provoked by the different codecs.

- For diotic presentation, speech quality of bandpass-filtered speech may be higher in case that all the low-frequency components are present in the signal, while a preference for less low-frequency components has been reported in case of monotic presentation. This probably reflects the fact that in ecologically valid situations low-frequency components are typically perceived more or less equally by the two

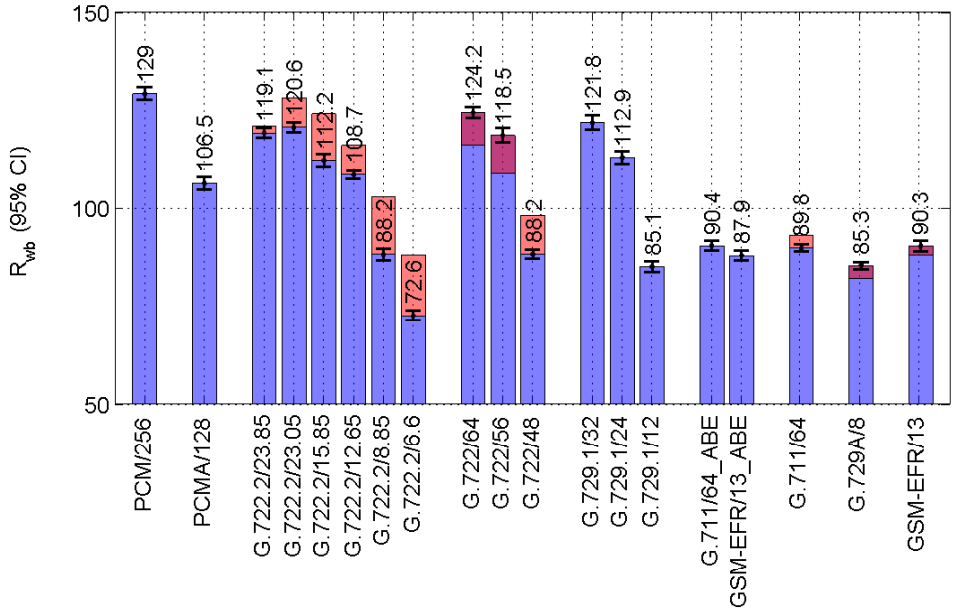

Fig. 2: Transformed and normalized results for error-free codecs. The blue boxes and respective errorbars show the test-results. The stacked pink and purple boxes indicate the deviation of the test results from the expected quality according to [8] and [12] (pink/errorbars below top of box: Test results lower than expected; purple/errorbars on top of box: test results higher than expected). Note: ABE indicates two NB conditions with artificial bandwidth extension.

ears, in contrast to high-frequency components (see e.g. [1], p. 89).

- The stream segregation process in human audition may render listeners to be more sensitive to additive noise in case of monotic listening than in case of diotic listening, in spite of the identity of the noise signals presented to the two ears in the latter case.

A first informal expert-listening session revealed a perceptually different impact of different wideband codecs when interacting with the room-acoustics present in the recordings. This appears plausible, since speech codecs are mainly based on models of the statistical properties of the source, typically considering some source–filter type of speech production model. This model does not account for reverberant speech, so that different algorithms are expected to differently deal with reverberation.

4. LISTENING TEST 2: IMPACT OF ROOM ACOUSTICS AT SEND-SIDE AND MONOTIC VS. DIOTIC PRESENTATION

In order to systematically investigate the hypothetical influences described above, we have conducted a second listening test. The test was subdivided into four randomly distributed sessions, where each session correspond to a combination of a given speaker (female, male) and a presentation mode (diotic, monotic). All other settings were identical between sessions, as outlined in the following.

4.1. Test 2 Set-up

In this test, we have investigated several factors and their combinations. An overview of all employed conditions is given in Figure 3. In order to study the impact of the room acoustics at send side, we have measured impulse responses of four different rooms using a speaking dummy head with the microphone positioned at three different distances (Mouth Reference Point (MRP), i.e. 2.5 cm; 15 cm; 30 cm). The non-reverberant source recordings described in Section 2.1 were convolved with the measured impulse responses. In an informal listening session, we have selected the room-distance-combinations considered to best cover the targeted effects; as a result, two of the four rooms and the two microphone distances of 15 and 30 cm were retained for further study. The rooms are two meeting rooms at

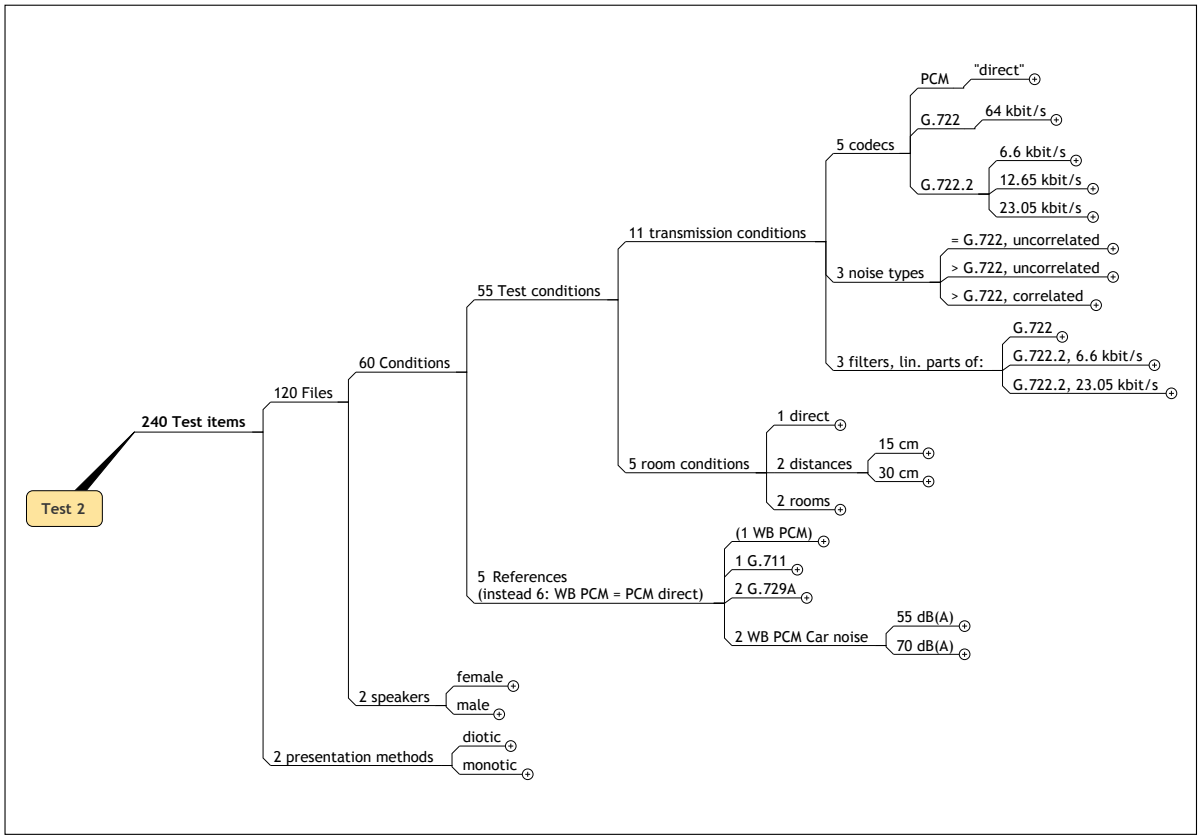

Fig. 3: Overview of Test 2 test conditions, see text for details.

our lab's premises, which are both moderately reverberant, but of very different dimensions (R1: 3.5 m x 3.5 m; R2: 9 m x 4.5 m).

The processed samples (no room, and convolved with the impulse responses for R1 and R2, each at 15 cm and 30 cm) were coded with the G.722 at 64 kbit/s, or the G.722.2 at the bitrates 23.05, 12.65 and 6.6 kbit/s.

In Section 3 we have identified the linear and the non-linear component as potential sources for interactions with the diotic versus monotic presentation mode or with the presence of a room in the source recordings. Reflecting the hypothesis that the noise introduced by the sub-band ADPCM of G.722 may be perceived differently under diotic and monotic listening conditions, we have inserted white noise at two different levels (-63 dBov and -50 dBov) into the uncoded speech. Note that the level of -63 dBov was perceptually adjusted to that of the G.722 by the authors. For the case of diotic presentation, auditory streaming mechanisms may favor speech-from-noise-separation when a non-coherent noise is presented in combination with the coherent speech signal (i.e. dichotic noise presentation, diotic speech presentation) [28]. To investigate this issue further, we have generated the files with -50 dBov noise both with a coherent and with a non-coherent (dichotic) version of the noise. Reflecting the hypothesis made in Section 3 that the linear (frequency) distortion introduced by a given codec may interact with the presentation mode or room acoustic conditions at send side, we have included conditions with codec-type linear filters. To this aim, we have estimated linear spectra for the G.722 at 64 kbit/s, and the G.722.2 at 23.05, 12.65 and 6.6 kbit/s (see Figure 5 for examples).

We have first conducted two formal expert-listening tests

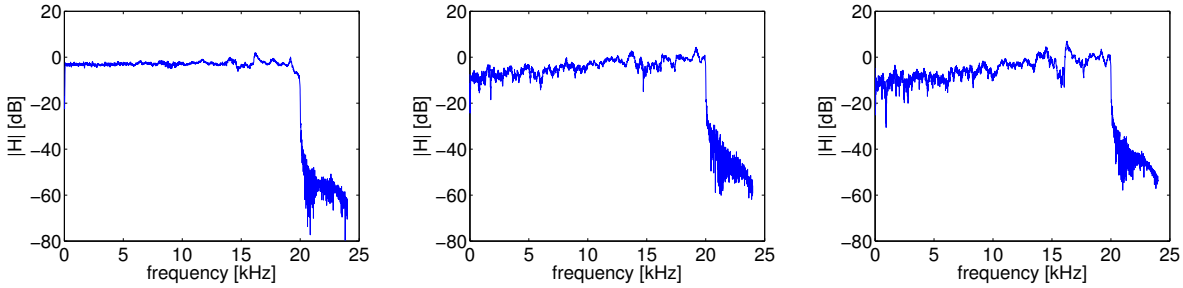

Fig. 4: Amplitude spectra for the Room R2 measurements, microphone at MRP (a), 15 cm (b), and 30 cm (c).

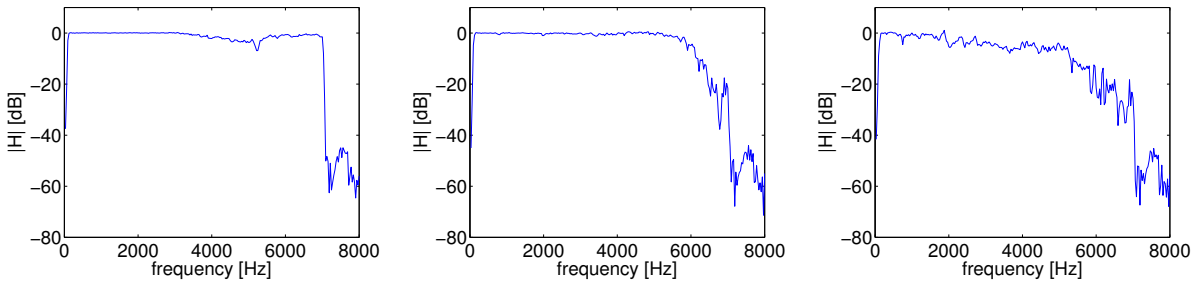

Fig. 5: Estimated amplitude spectra for the G.722 (64 kbit/s) (a), G.722.2 (23.05 kbit/s) (b), and G.722.2 (6.6 kbit/s) (c).

with 10 listeners each, one with monotic, and one with diotic presentation. The diotic test revealed a clear preference for the G.722 regardless of the room acoustics present in the source material. In turn, this preference was not found in the test using monotic presentation. To elucidate this question in more detail, we have conducted a formal listening tests with naive subjects, separated into 4 sessions according to the presentation type (diotic versus monotic) and the speaker (female, male). The same source recordings as described for Test 1 were used. In Test 2, every condition was paired with eight different sentences per speaker. For each subject, four individual playlists were generated (one per session), randomly selecting one of the eight available sentences per speaker-condition-combination, and then randomizing the resulting test items. For each subject, the four sessions were carried out one after the other in randomized session-order, with 10 min pauses between sessions and an overall duration of approximately 1 h. In order to prepare the subjects for the range of qualities and the perceptual dimensions to be expected during the test, each session was preceded by 6 training stimuli from the set of test stimuli (not used in the analysis of the results). The technical set-up was identical to that of Test 1, apart from the fact that only one subject took part in each session, and that the test was conducted in a small sound-proof booth instead of the larger sound-insulated studio room employed for Test 1. Monotic presentation was made to each subject's preferred ear. To this aim, the other earpiece of the employed Sennheiser HD25 headphone was flipped upwards to rest on the subject's head. 24 subjects participated in the test, who were recruited from the campus of Berlin Technical University (12 female, 12 male). The subjects were, to their own account, normal hearing.

4.2. Test 2 Results

At first, we have evaluated the correspondence of the ratings obtained from each subject per item with the average across all subjects, to validate the subject performance. Here, we found two subjects with a root mean squared deviation from the general mean (RMSD) greater than 1.

A detailed analysis of their ratings showed an erroneous selection of the respective playlists prior to the test. Consequently, 22 of the 24 subjects were used for the final data analysis.

Since we were not directly interested in the impact of the speaker, we have first averaged the results per condition and subject across the speakers. A repeated-measures mixed linear models ANOVA [29] of the ratings was performed, modeling the 'codec' (i.e. codec and bitrate considered together), the 'room', the 'distance' of the microphone and the 'presentation' (monotic vs. diotic) as fixed effects. Note that due to the incomplete design the conditions with linear distortion or additive noise were discarded here. The analysis reveals highly significant effects of the 'room' ($F = 22.815$, $p < 0.001$), of the microphone 'distance' ($F = 35.763$, $p < 0.001$), of the 'codec' ($F = 557.621$, $p < 0.001$), and of the interaction 'presentation'*'codec' ($F = 9.748$, $p < 0.001$). Less strong but still significant effects were found for the interactions 'room'*'distance' ($F = 4.651$, $p < 0.05$) and 'distance'*'codec' ($F = 3.544$, $p < 0.01$).

The effects can be interpreted as follows:

- 'Codec': This factor has the strongest impact on quality. In the light of the wide quality-range of the codecs being used, including both the higher quality G.722 and G.722.2 (23.05 kbit/s) as well as the lower-quality G.722.2 at 6.6 kbit/s, this result was to be expected.
- 'Room', 'distance': As already stated in [30, 1, 31], the room acoustic conditions at send side have a considerable impact on perceived naturalness and quality.
- 'Presentation'*'codec': The significant interaction observed between the employed codec and presentation is a proof of the hypothesis that the codec rank-order resulting from a test may actually depend on the employed presentation method.
- 'Distance'*'codec': There is a smaller but significant interaction between the employed codec and the room acoustic conditions at send side. This proofs our second hypothesis that different codecs are more or less affected by room reflections; here, the impact may range from masking of non-linear effects to a decreased codec performance due to interference with the employed predictions.

Two further repeated-measures mixed linear models ANOVA were performed: In the first, we have modeled the 'linear codec distortion' (with the nominal levels 'none', 'G.722', 'G.722.2, 23.05 kbit/s', 'G.722.2, 12.25 kbit/s', and 'G.722.2, 6.6 kbit/s',), the 'room', the 'distance' of the microphone and the 'presentation' as fixed effects. Note that due to the incomplete design, we excluded all conditions with real codecs for this analysis, as well as the conditions with codec-simulating noise. Likewise to the earlier analysis, significant effects were found for 'room' ($F = 27.657$, $p < 0.001$), and 'distance' ($F = 45.978$, $p < 0.001$). The only additional significant effect is that of the 'filter' in isolation ($F = 166.909$, $p < 0.001$), reflecting the very little distortion due to e.g. the G.722 and the strong bandwidth distortion of the G.722.2 at 6.6 kbit/s.

In the second repeated-measures mixed linear models ANOVA, we have modeled the codec-emulating 'noise' (excluding the additive car noise used as reference), the 'room', the 'distance' of the microphone and the 'presentation' as fixed effects. Note that we have excluded all conditions with real codecs here, as well as the conditions with additive noise. Due to the resulting limited number of conditions, and the fact that the noise effect is more dominant, a significant 'room' effect could not be found this time. Statistically significant effects were found for 'distance' ($F = 10.788$, $p = 0.001$), 'noise' ($F = 505.211$, $p < 0.001$), 'presentation' ($F = 5.38$, $p < 0.05$), and 'the interaction 'distance'*'noise' ($F = 5.781$, $p = 0.001$). None of these observations can directly be interpreted, but as we will see in the following, noise seems to explain the rank-ordering effects between the codecs G.722 and G.722.2 (23.05 kbit/s) for the two presentation modes.

From the analysis of the results and by informal listening it is revealed that the two rooms are perceptually quite similar in comparison with the no-room-condition. Also, there was no condition with microphone distance $d = 0$ or $d = 2.5$ *cm* being tested, so that the room effect is established in terms of the conditions 'no room', 'R1 or R2, 15 cm', and 'R1 or R2, 30 cm'. Hence, in the remainder of the paper, the two rooms R1 and R2 are considered together by averaging the results for the respective conditions.

In Figure 6, we show the results for linear PCM, the G.722 at 64 kbit/s, the G.722.2 at 23.05 kbit/s, and the G.711. The left part of the graph shows the results for the case of no room ('N') and monotic presentation

('M'); the middle part presents the results for the average over R1 and R2 at 15 cm ('R') and monotic presentation ('M'); the right part shows the results for no room ('N') and diotic presentation ('D'). Note that the graph shows a (however non-significant) disadvantage for the G.722.2 at 23.05 kbit/s over the G.722 for diotic listening, while it is equal to or better for the case of monotic listening.

However, as can be observed from the graph, there is a jump in quality for the G.722 when passing from the room condition under monotic presentation to the case of no room with diotic presentation. This case exactly corresponds to the difference between the tests summarized in the literature [2] and the ones we have conducted here (see Table 1). In turn, the results for all other cases are more or less identical between the different cases, which indicates that these are more or less invariant to presentation and perceivable room acoustics at send side. Since

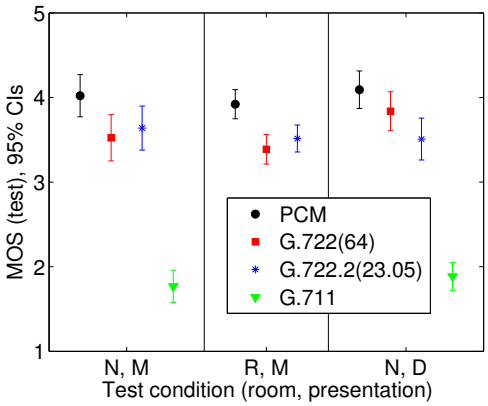

Fig. 6: Comparison of mean ratings for linear PCM, G.722 (64 kbit/s), G.722.2 (23.05 kbit/s) and G.711, when paired with the cases 'no room, monotic listening (N, M; left)', 'room R1 or R2 at 15 cm distance, monotic (middle; R, M)', and 'no room, diotic (right part; N, D)'. Note that we did not test the G.711 together with a room.

the G.722 is behaving differently from the other codecs, the hypothesis that noise may cause an interaction is a possible candidate for further explanation. To elucidate this issue, we have studied the case of additional noise in more detail. In this respect, Figure 7 shows the results for the case of no room (monotic presentation: left, diotic presentation: right) and of rooms R1/R2 at 15 cm under monotic presentation. Note that the results for diotic, R1/R2 were omitted here since they do not provide

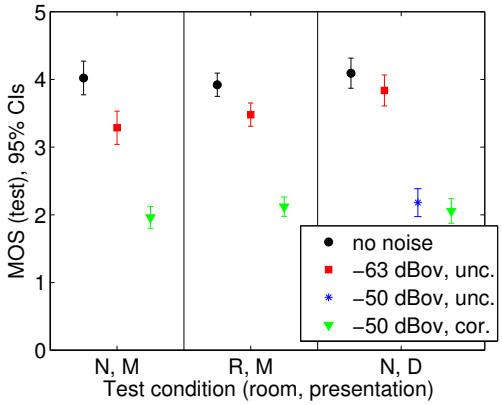

Fig. 7: Comparison of mean ratings for linear PCM without (no noise) and with additional codec-emulating noise (see Section 4.1) at different levels, and correlated as well as uncorrelated between the left and right channels. As in Figure 6, the left part represents the case of no room and monotic presentation, the mittle part that of room at 15 cm and monotic presentation, and the right part that for diotic presentation without room.

additional insight. The results show that there is a significant difference for the low-noise condition between monotic and diotic presentation, when no room acoustics are present (red error-bars in the left and right part of the graph, respectively). There is no difference between monotic and diotic presentation for the case of the higher noise level. An effect of the noise being correlated or uncorrelated cannot be proven fby our test.

Another finding can be observed from Figure 8. The ratings imply that the perception of non-linear distortions as they are introduced by the G.722.2 at low bitrates is – in tendany – worse for diotic than for monotic listening. This is in-line with some of the effects reported by [32]. In turn, the assumed improvement of quality for the G.722.2 when reverberant instead of non-reverberant recordings are employed, cannot be confirmed by our research. Instead, the reverberant conditions have a tendency to be slightly worse than the non-reverberant ones (left and middle port of Figure 8). Here, there seems to be a difference between the acoustic conditions at send and at receive side (compare Section 3 and [27]).

5. CONCLUSION

We have shown that the a diotic versus monotic presentation of coded speech has a significant impact on the

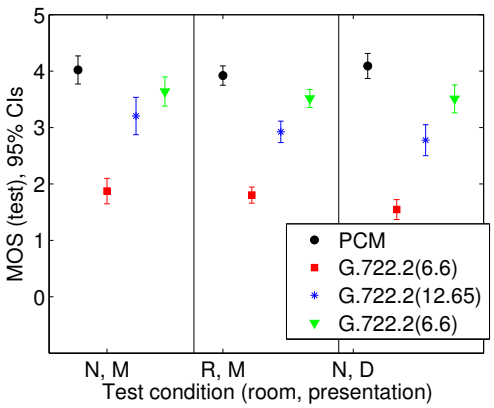

Fig. 8: Comparison of mean ratings for PCM, and the G.722 at 6.6, 12.65 and 23.05 kbit/s; left part: no room, monotic; middle part: R1/R2 at 15 cm, monotic; right part: no room, diotic.

quality rank-ordering of the investigated codecs. Our research suggests that the G.722 is – relative to the uncoded case of linear PCM – perceived significantly better under diotic listening than under monotic listening. The test results further suggest that this is due to the additive noise that the split-band ADPCM of the G.722 introduces. For higher noise levels, monotic and diotic presentation yield very similar quality. Our study could not support the hypothesis that certain room acoustics at send side may mask and thus attenuate the quality impact of the non-linear degradation as it is, for example, introduced by the G.722.2 at its lower bitrates. There was a significant effect shown in our statistical analysis, but it could not be linked with the comparison between certain test conditions. However, for the G.722 the combination of a presence of room acoustics at send side with a monotic presentation seems to increase the quality-difference to anechoic G.722-coded and diotically presented speech, which yields significantly higher quality.

The finding that there is an interaction between the presentation mode (monotic vs. diotic) and the degradation different codecs introduce is of particular importance when choosing a specific codec: Here, the presentation mode primarily employed by the users needs to be taken into consideration. Also for the characterization and selection of codecs in the standardization process, the listening mode is of high importance. Especially in the light of a migration from speech- or audio-only- towards speech-and-audio-applications, the observed rank-order effects need to be reflected in corresponding subjective tests.

6. REFERENCES

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

[2] Sebastian Möller, Alexander Raake, Nobuhiko Kitawaki, Akira Takahashi, and Marcel W altermann. Impairment factor framework for wideband speech codecs. *IEEE Trans. Audio Speech and Language*, 14(6):1969–1976, 2006.

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

[4] 3GPP TS 26.071. *Mandatory Speech Codec Speech Processing Functions; AMR Speech Codec; General Description*. 3rd Generation Partnership Project (3GPP), F–Sophia Antipolis, June 2004.

[5] ETSI EN 300 726 (GSM 06.60). *Digital Cellular Telecommunications System; Enhanced Full Rate Speech Transcoding*. European Telecommunications Standards Institute, F–Sophia Antipolis, 1996.

[6] 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.

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

[8] ITU–T Rec. G.113 Appendix IV. *Provisional Planning Values for the Wideband Equipment Impairment Factor Ie,wb*. International Telecommunication Union, CH–Geneva, June 2006.

[9] Alexander Raake, Sascha Spors, Hans-Joachim Maempel, Timon Marszalek, Simon Ciba, and Nicolas Côté. Speech quality of wide- and narrowband codecs: Object- and subject-oriented view. *In: Fortschr. Akust., 34. Jahrestagung für Akustik (DAGA 2008), D-Dresden*, pages 639–640, Dtsch. Ges. Akust. (DEGA), D–Berlin., 2008.

[10] Alexander Raake, Marcel Wältermann, Nicolas Côté, and Sebastian Möller. Speech quality of wideband Voip under packet loss. *In: Proc. 20th Conference Elektronische Sprachsignalverarbeitung (ESSV), to appear, DE-Dresden*, 2009.

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

[12] ITU–T Rec. G.107 Appendix II. *Provisional Impairment Factor Framework for Wideband Speech Transmission*. International Telecommunication Union, CH–Geneva, June 2006.

[13] ITU–T Rec. G.113 Appendix I. *Provisional Planning Values for the Equipment Impairment Factor Ie and Packet-Loss Robustness Factor Bpl*. International Telecommunication Union, CH–Geneva, May 2002.

[14] ITU–T Rec. P.501 Amendment 1. *Test signals for use in telephonometry; New Annexes A and B*. International Telecommunication Union, CH–Geneva, 2004.

[15] ITU–T Rec. G.729.1. *G.729 based Embedded Variable Bit-rate Coder: An 8-32 kbit/s Scalable Wideband Coder Bitstream Interoperable with G.729*. International Telecommunication Union, CH–Geneva, May 2006.

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

[17] ITU–T Rec. G.726. *40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)*. International Telecommunication Union, CH–Geneva, December 1990.

[18] ITU-T Rec. G.191. *Software Tools for Speech and Audio Coding Standardization*. International Telecommunication Union, CH–Geneva, Sept. 2005.

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

[20] Davydd 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.

[21] ITU–T Rec. P.833.1. *Methodology for the Derivation of Equipment Impairment Factors from Subjective Listening-only Tests for Wideband Speech Codecs*. International Telecommunication Union, CH–Geneva, May 2008.

[22] ITU–T Delayed Contribution D.151. *Towards a wideband E-model: R-scale extension and Impairment factors for wideband speech codecs*. Source: Federal Republic of Germany and Japan; authors: Möller, S., Raake, A., Kitawaki, N., Takahashi, A., Wältermann, M. International Telecommunication Union, CH–Geneva, June 2006.

[23] Marcel Wältermann, Alexander Raake, and Sebastian Möller. The sound character space of spectrally distorted telephone speech and its impact on quality. *In: Proc. 124th AES Convention*, May 17 - 20, NL–Amsterdam, 2008.

[24] Ville-Veikko Mattila. Descriptive analysis and ideal point modelling of speech quality in mobile communications. *In: Proc. 113th Audio Engineering Society (AES) Convention*, October 5-8, USA–Los Angeles, 2002.

[25] H.-W. Schüssler. An objective method for measuring the performance of weakly non-linear and noise systems. *Frequenz*, 41(6):147–154, 1987.

[26] Marcel Wältermann and Alexander Raake. Towards a new e-model impairment factor for linear distortion of narrowband and wideband speech transmission. *In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, March 30 - April 4, USA–Las Vegas:4817–4820, 2008.

[27] Daniel Schobben and Steven van de Par. The effect of room acoustics on mp3 audio quality evaluation. *In: Proc. 117th Audio Engineering Society (AES) Convention*, Oct. 28-31, USA–San Francisco, 2004.

[28] Al S. Bregman. *Auditory Scene Analysis*. The MIT Press, USA–Cambridge, 1990.

[29] Hugo Quené and Huub van den Bergh. On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, 43(1-2):103–121, 2004.

[30] Norman Gleiss. Desirable sending frequency response of telephone sets. *TELE (English edition)*, 1/89:18–23, Swedish Telecommunications Administration, S–Stockholm, 1989.

[31] Marc Brüggen. *Klangverfärbung durch Rückwürfe und ihre auditive und instrumentelle Kompensation*. dissertation.de, `www.dissertation.de`, D–Berlin, 2001.

[32] Arnault Nagle, Catherine Quinquis, Aurélien Sollaud, and Anne Battistello. Quality impact of diotic versus monaural hearing on processed speech. *In: Proc. 123rd Audio Engineering Society (AES) Convention*, Oct. 5-8, USA–New York, 2007.