Subjective and Objective Assessment of Perceived Audio Quality of Current Digital Audio Broadcasting Systems and Web-Casting Applications

Peter Počta and John G. Beerends

***Abstract*—This paper investigates the impact of different audio codecs typically deployed in current digital audio broadcasting (DAB) systems and web-casting applications, which represent a main source of quality impairment in these systems and applications, on the quality perceived by the end user. Both subjective and objective assessments are used. Two different audio quality prediction models, namely Perceptual Evaluation of Audio Quality (PEAQ) and Perceptual Objective Listening Quality Assessment (POLQA) Music, are evaluated by comparing the predictions with subjectively obtained grades. The results show that the degradations introduced by the typical lossy audio codecs deployed in current DAB systems and web-casting applications operating at the lowest bit rate typically used in these distribution systems and applications seriously impact the subjective audio quality perceived by the end user. Furthermore, it is shown that a retrained POLQA Music provides the best overall correlations between predicted objective measurements and subjective scores allowing to predict the final perceived quality with good accuracy when scores are averaged over a small set of musical fragments ($\mathbf{R = 0.95}$).**

***Index Terms*—Perceived audio quality, audio coding, digital audio broadcasting (DAB) systems, audio web-casting applications, subjective audio quality assessment, objective audio quality assessment, Perceptual Evaluation of Audio Quality (PEAQ), Perceptual Objective Listening Quality Assessment (POLQA) Music.**

I. Introduction

DEPLOYMENT and popularity of digital audio broadcasting systems and web-casting applications is growing steadily due to an evolution of digital radio and communication systems. As digital audio broadcasting systems, like DAB (Digital Audio Broadcasting) [1], [2], DRM (Digital Radio Mondiale) [3], [4] and T-DMB (Terrestrial Digital Multimedia Broadcasting) [2], [5], are currently replacing conventional analog radio systems around the world and web-casting applications are allowing to listen to an audio content everywhere and every time due to a proliferation of ubiquitous Internet access and mobile hand-held devices the audio quality provided by these systems and applications becomes an important issue for broadcasters, service providers as well as end users.

Manuscript received November 29, 2014; revised March 10, 2015; accepted April 10, 2015. Date of publication May 8, 2015; date of current version September 2, 2015.

P. Počta is with the Department of Telecommunications and Multimedia, FEE, University of Žilina, Žilina SK-01026, Slovakia (e-mail: pocta@fel.uniza.sk).

J. G. Beerends is with TNO, The Hague NL-2509 JE, The Netherlands.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2015.2424373

TABLE I
CODECS AND TYPICAL BIT RATES USED IN CURRENT DIGITAL AUDIO BROADCASTING SYSTEMS AND WEB-CASTING APPLICATIONS (ADOPTED FROM [6] AND UPDATED)

Codec	Typically used bit rates (kbps)	Digital audio broadcasting systems and web-casting applications deploying this codec
MP2	64-192	DAB
HE-AACv2	24-128	DAB+, DRM, DRM+, DMB, Wimp, Web streaming, Deezer
MP3	64-320	Web streaming, Deezer
Ogg Vorbis	64-320	Web streaming, Spotify
AAC-LC	32-320	Web streaming, Wimp, ISDB, iTunes
Opus	32-320	Web streaming
ALAC	≥ 700	iTunes, Wimp (iOS)
FLAC	≥ 700	Wimp, Online music catalogs

As current digital audio broadcasting systems and web-casting applications are designed to be very resilient to errors introduced by a error-prone transmission channel, due to deploying highly efficient channel coding schemes (broadcasting systems) or highly reliable transmission method based on TCP protocol (web-casting applications), the audio codec is the main source of quality degradation in these systems and applications. Table I summarizes codecs and typical bit rates used in current digital audio broadcasting systems and web-casting applications. As we can see from Table I, 6 lossy coding schemes, i.e., MP2, HE-AACv2, MP3, Ogg Vorbis, AAC-LC and Opus, and 2 lossless coding schemes, i.e., FLAC and ALAC, are deployed in current digital audio distribution systems and applications. The bit rate used by the lossy codecs varies from 24 to 320 kbps. On the other hand, a bit rate higher than 700 kbps is mostly used when the lossless codecs, like ALAC and FLAC, are deployed.

The audio quality perceived by the end user is directly influenced by the operating bit rate. Broadcasters and service providers are forced to use a lowest bit rate possible in order to cope with a crowded and expensive broadcast radio spectrum or to allow the end users to use the web-casting

applications at remote places where good 4G or 3G coverage is mostly rare. In parallel to keep the bit rate as low as possible, they want to offer them the best quality possible at a particular bit rate deployed for the service. Therefore, it is of crucial importance for broadcasters and service providers to measure/monitor the quality perceived by the end user in the digital audio broadcasting systems and web-casting applications as users migrating from analog systems to these digital systems/applications expect higher quality than experienced in the analog systems.

To quantify the perceived audio quality of a transmission chain or audio codec, a number of measurement techniques have been developed during the past decades. In the subjective domain the most widely used for assessing small impairments is ITU-R Recommendation BS.1116-1 [7]. In this approach the original and degraded signals are usually synchronously looped and subjects are free to switch between a hidden reference, the original and degraded signal. Subjects listen to the signals using high-quality headphones and vote on a 5-grade continuous opinion scale focusing on impairments, see [17, Table I]. The resultant quality score is termed SDG (Subjective Difference Grade). In the objective domain, several intrusive models were developed during the early 90s and were submitted to a competition run by ITU-R from 1994 to 1996 [8]–[13]. After a competitive phase the ITU standardized two models, known as the Basic and Advanced Version of the Perceptual Evaluation of Audio Quality (PEAQ, ITU-R Recommendation BS.1387-1 [13], [14]). It should be noted here that all the models mentioned above were trained on subjective scores obtained from the tests based on ITU-R Rec. BS.1116-1, with emphasis on distortions that are inaudible or just noticeable to expert listeners. In [15], improvements to PEAQ including new distortion parameters and a new cognitive model have been proposed. Moreover, a limitation of PEAQ to only deal with single channel distortions has been addressed by the development of an expert system to assist with an optimization of multi-channel audio systems [16]. A modified version of PEAQ model allowing a perceptual evaluation of high-quality multi-channel audio codecs has been proposed in [17]. In [18], Huber and Kollmeier proposed a new model for audio quality assessment called PEMO-Q representing an expansion of their speech quality measure [19]. This model is based on a psychoacoustically validated quantitative model of the "effective" peripheral auditory processing defined by Dau in [20]. The model performs slightly better than PEAQ on a major subset of its training data and shows a more robust prediction performance on completely new data. It is worth noting that the model is not suitable for predicting the impact of linear distortions on the quality perceived by the end user. Moreover a model called CASP-Q was proposed in [21]. This model is based on PEMO-Q model but an auditory front end of PEMO-Q was replaced by the recent model described by Jepsen *et al.* in [22] offering nonlinear auditory filters. It offers a bit better accuracy than PEMO-Q on data used in an experiment presented in [21]. In addition to that, POLQA (Perceptual Objective Listening Quality Assessment) model [23], [24] originally designed for assessing a speech quality of narrowband, wideband and super wideband speech signals transmitted over telecommunication networks and standardized in ITU as ITU-T Rec. P.863 [25] is currently being extended towards an audio quality assessment by Beerends. A working title for this model is POLQA Music. For more information about subjective and objective assessment of perceived audio quality, we refer the interested reader to [26] and [27].

Some work has been carried out to study the perceived audio quality of digital radio systems and the performance of PEAQ model for several perceptual audio codecs. In [28], Lee *et al.* evaluated the perceived audio quality of four digital radio systems, namely DAB, DAB+, HD Radio and T-DMB, using a test methodology defined in ITU-R Rec. BS.1534-1. The evaluation process consisted of two phases. In the first phase, the audio quality of each system was evaluated as a function of the bit rate and a relationship between the audio quality and channel capacity of each system was obtained. In other words, the authors ran a subjective test comparing an audio quality provided by different digital audio broadcasting systems, namely DAB, DAB+, HD Radio and T-DMB for different bit rates to obtain information on the relationship between the audio quality and channel capacity of each system. In order to check evaluation consistency, the audio quality perceived by the end user as a function of the systems was assessed in second phase of this study. Two subgroups comprising mid and high quality bit rates of all the investigated systems respectively were determined for this phase from the test cases involved in the first phase to evaluate the audio quality as a function of the systems. The evaluation consistency was confirmed, which means that the results obtained in the first phase of the study are valid. As the authors ran their tests on real digital radio equipments, the results can be used for benchmarking purposes. Moreover, the results can also be used as a reference to determine an optimal channel capacity of a digital radio service for a given target perceived audio quality.

In [6], Berg *et al.* reviewed audio quality criteria for radio broadcasting systems and investigated how the perceived audio quality of different broadcasting systems complies with the defined criteria. Two subjective tests were run in this study. The first test used ITU-R Rec. BS.1116-1 to evaluate the audio quality of the HE-AACv2 codec at bit rates between 96 and 192 kbps. The investigated conditions represent typical DAB+ conditions in terms of the codec settings. The second test used ITU-R Rec. BS.1534-1 to evaluate different DAB+ system configurations with bit rates ranging from 48 to 192 kbps and FM systems. Both broadcasting systems involved in the second test were setup to mimic as much as possible a realistic broadcasting signal chain including commonly used dynamic processors. The results showed that the currently highest available sub-channel bit rate for DAB+, namely 192 kbps, was insufficient for attaining perceptually transparent quality for critical items, whereas it provided a quality comparable to or in some cases even better than that offered by a modern FM system. Extrapolation of data indicates that critical items may need to be coded at even higher bit rates to reach perceptually transparent quality. The study concludes that when making decisions on broadcasting systems,

it is important to have well-founded and clearly defined criteria for a minimum acceptable quality and/or perceptually transparent quality. In [29], Treurniet and Soulodre compared predictions provided by the PEAQ model for several perceptual audio codecs with subjective scores obtained from a subjective listening test run according to ITU-R Rec. BS.1116-1 and described in more detail in [30]. The results showed that the predicted quality ratings for the codecs correlate well with the subjective ratings obtained from the listening test. Furthermore, the objective quality measurements averaged over audio items were statistically indistinguishable from similarly averaged subjective quality ratings. Finally, mean objective codec quality measurements generated from a larger number of pre-selected audio items were found to be similar to the subjective ratings and objective quality measurements for a few critical items chosen for the listening test.

As already mentioned, the audio codec is the main source of quality degradation in current digital audio broadcasting systems and web-casting applications and it is important for broadcasters and service providers to measure/monitor the quality perceived by the end user in these systems and applications in order to be competitive in the market. In order to be able to measure/monitor the perceived quality we have to make use of a perceptual measurement approach. In this paper we investigate the performance of two perceived audio quality prediction models based on the perceptual measurement approach, namely PEAQ and POLQA Music, for lossy codecs and bit rates typically deployed in current digital audio broadcasting systems and web-casting applications (see Table I). Two lossless codecs listed in Table I, namely FLAC and ALAC, are excluded from our study as their deployment in the web-casting applications is still rare and as they are designed to be fully transparent. In order to cover all range of coding impairments introduced by the corresponding codec in the context of the broadcasting systems and web-casting applications deploying that codec, boundary bit rates summarized in Table I are used in this study. To the best of our knowledge a study comparing the perceived quality of these codecs in typical conditions for current digital audio broadcasting systems and web-casting applications on the same critical signals is not available at this moment.

In the first part of this study we assess the perceived quality of six lossy audio codecs, operating at the boundary bit rates typically deployed in current digital audio broadcasting systems and web-casting applications, using ITU-R Rec. BS.1116-1. The performance is evaluated on six critical signals selected from the publicly available EBU SQAM database [31] created by broadcasting experts involved in EBU (European Broadcast Union). Some of them were also used in the two studies conducted by the Audio Perception Lab of the Communications Research Centre mentioned above [29], [30]. In the second part of this study, the performance of PEAQ and POLQA Music models is assessed by comparing the predictions with SDG values obtained from the test described in this paper. The aim of this study is threefold: first, we would like to know to what extent degradations introduced by the lossy audio codecs deployed in current digital audio broadcasting systems and web-casting applications have an impact on the audio quality perceived by the end user. Secondly, we would like to compare the performance of the lossy codecs listed in Table I for typical boundary bit rates deployed in current digital audio broadcasting systems and web-casting applications on the same critical signals from the perceived quality perspective. Thirdly, we would like to see whether the PEAQ and POLQA Music models are able to provide valid predictions of perceived audio quality for the given application domain.

The remaining of the paper is organized as follows. Section II describes the subjective test carried out within this study and its results. In Section III, the experimental results obtained from the prediction models are compared with the subjective data presented in this paper and discussed. Section IV provides the final conclusions.

II. Subjective Test

The first aim of this study is to see how degradations introduced by the lossy audio codecs deployed in current digital audio broadcasting systems and web-casting applications impact the audio quality perceived by the end user. The second one is to benchmark the lossy codecs listed in Table I for boundary bit rates typically deployed in current digital audio broadcasting systems and web-casting applications and on same critical signals from the perceived quality perspective. To this end, a subjective test has been carried out. The following sections provide a description of this test and the results that are obtained.

A. Experiment Description

The subjective listening test was performed in accordance with ITU-R Rec. BS.1116-1. In all experiments one listener was seated in a small listening room (acoustically treated) with a background noise below 20 dB SPL (A). Altogether, 23 listeners (15 male, 8 female, 16–46 years, mean 27.48 years) with an appropriate expertise, as defined in ITU-R Rec. BS.1116-1, participated in the test. All involved subjects were pre-screened for detecting small impairments in a training phase of the test as advised by the particular ITU-R recommendation. Two subjects (one male and one female) were rejected during a post-screening phase due to an incorrect identification of the hidden reference in the test (a high quality test case) and the associated scores were removed from the data set, so grades from 21 listeners (14 male, 7 female, 16–46 years, mean 27.33 years) were used for the analysis. The subjects were remunerated for their efforts. The audio samples were played out using high quality studio equipment in a random order and binaurally presented (Sennheiser HD 455 headphones, presentation level: 79 dB SPL (A)) to the test subjects. The presentation level was derived experimentally before the subjective test as the average of preferred playback levels of 5 expert listeners for all the audio samples involved in the test. The absolute grades, ranging from 1 (very annoying) to 5 (imperceptible), given by the subjects in the test were transformed to difference grades (a grade for the degraded sample minus the grade given to the hidden reference) and averaged to obtain SDG values for each audio signal.

TABLE II
AUDIO TEST SIGNALS USED IN THE SUBJECTIVE TEST

Name	Description	Duration (s)	Source
dbass	Double bass arpeggio	14	*EBU SQAM CD (track 11/Index 1) [31]
flute	Flute arpeggio	15	EBU SQAM CD (track 13/Index 1) [31]
bascl	Bass clarinet arpeggio	10	*EBU SQAM CD (track 17/Index 1) [31]
casta	Castanets single tone, rhythm	8	EBU SQAM CD (track 27/Index 1) [31]
harps	Harpsichord arpeggio	10	*EBU SQAM CD (track 40/Index 1) [31]
eddie	Eddie Rabbitt	13	EBU SQAM CD (track 70/Index 1) [31]

*Audio signals used in [30]

TABLE III
CODECS AND LOWEST AND HIGHEST BIT RATES USED IN THE SUBJECTIVE TEST

Codec	Bit rate (kbps)	
	Lowest	Highest
MP2	64	192
AAC-LC	32	320
Opus	32	320
MP3	64	320
HE-AACv2	24	128
Ogg Vorbis	64	320

The selection of critical audio test material is an important step in assessing the quality of lossy audio codecs since they do not perform uniformly over all audio signals. A given codec may be effective at coding some audio signals but it may perform poorly for other signals. Therefore, a goal of this test was to evaluate the performance of the codecs under worst-case conditions from a signal perspective. We did our best to find audio signals that would best reveal the limitations of each codec. As a starting point, we used a selection of the test signals made by Soulodre *et al*. in [30], as some of the investigated codecs in their study are the same as the codecs deployed in this study. Moreover, as one of the intentions of this study is to benchmark the investigated codecs on the audio signals coming from a publicly available database, the EBU SQAM database created by broadcasting experts involved in EBU was deployed in this study. The stereo audio test signals listed in Table II were found to be critical for the investigated codecs by the experimenters in a pre-test and were used in this subjective test. More detail about the used audio signals can be found in [32].

The codecs listed in Table III together with the defined lowest and highest bit rates, representing typical lowest and highest bit rates for current digital audio broadcasting systems and web-casting applications, were used in the subjective test. It should be noted here that a sampling rate of 48 kHz, which is commonly used in the digital audio distribution systems and applications, was deployed.

The six critical audio test signals listed in Table II were encoded by the corresponding codecs and bit rates listed in Table III to yield to 72 test items. The 72 test items were divided to two test sessions balanced from a size and degradation perspective in order to avoid subject fatigue as advised in ITU-R Rec. BS.1116-1.

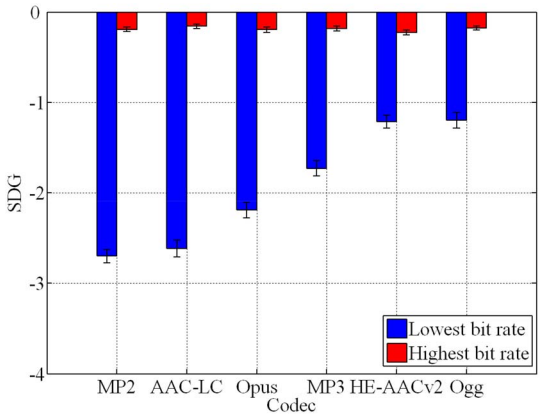

Fig. 1. Effect of codecs and bit rates (see Table III for more information about the exact lowest and highest bit rates used for a particular codec) on average SDG values. The vertical bars show 95% CI computed over 126 SDG values (21 subjective scores per audio signal for six audio signals).

B. Experimental Results

In Fig. 1, we summarize the results of the subjective test averaged over the 6 different audio signals involved in the experiment, as described in Section II-A. As can be clearly seen from this figure, the investigated codecs are ranked downwardly from a SDG obtained for the lowest investigated bit rate perspective. It is worth noting that a SDG value near 0 indicates a high quality, whereas a SDG value near -4 indicates a very poor quality. Moreover, the SDG scale is divided into four equal sections, namely "perceptible but not annoying" (ranging from -0.1 to -1), "slightly annoying" (ranging from -1 to -2), "annoying" (ranging from -2 to -3) and "very annoying" (ranging from -3 to -4). Half of the investigated codecs, namely MP2, AAC-LC and Opus, have SDG values obtained for the lowest investigated bit rate situated in the "annoying" section of the SDG scale. The SDG values obtained for the rest of the investigated codecs, i.e., MP3, HE-AACv2 and Ogg Vorbis, are located in a "slightly annoying" section of the SDG scale.

It is also apparent from Fig. 1 that the performance of each codec improves with a higher bit rate, as clearly expected. It should be also noted here that the SDG values reported for the highest bit rate oscillate around -0.19 for all the investigated codecs. In other words, all the SDG values obtained for all the highest investigated bit rates and involved codecs are placed in a "perceptible but not annoying" section and are even very close to "imperceptible" section of the scale ranging from 0 to -0.1. So, we can conclude that all the investigated codecs provide very good, almost excellent, quality for the highest investigated bit rate and a quality difference between them is negligible in this case. Moreover, the results of the HE-AACv2 codec, are in line with those presented in [6], the codec operating at 128 kbps was unable to attain a perceptually transparent quality represented by the "imperceptible"

section of the scale for all the critical items used in this study.

When comparing four generations of MPEG codecs from the lowest investigated bit rate perspective, we see a consistent improvement of the perceived quality. More specifically, the MP2 codec has the poorest performance while the MP3 offers an intermediate quality. The AAC-LC codec is very close to the MP2 at half of the bit rate. Finally, HE-AACv2 provides the best quality even for the lowest bit rate involved in our test, namely 24 kbps. The results confirm the results presented in [30], involving only MP2, MP3 and AAC-LC codecs, where the MPEG codecs where ranked according to their evolution, i.e., MP2, MP3, AAC-LC.

When comparing the perceived audio quality provided by the investigated codecs from the most critical condition point of view, i.e., the lowest bit rate, it can be seen from Fig. 1 that the best quality is provided by Ogg Vorbis codec at 64 kbps representing web-casting applications, which is very closely followed by HE-AACv2 codec at 24 kbps characterizing broadcasting systems. It should be noted here that the bit rate of 32 kbps was used in our study for Opus and AAC-LC codecs dominantly representing web-casting applications.

One three-way analysis of variance (ANOVA) test was conducted on the subjective results using codec, signal and bit rate as fixed factors (Appendix, Table IV). The effect of bit rate was found to be highly statistically significant (highest F-ratio of 3093, $p < 0.001$). Furthermore, the effect of codec was also highly statistically significant with $F = 72$, $p < 0.001$. The last factor investigated in the ANOVA test was the signal factor and it turns out to have a weaker effect on quality than the previous factors on their own ($F = 46$, $p < 0.001$) but was also highly statistically significant. Regarding interactions of all the involved factors, the results show that all of them were highly statistically significant. Moreover, the highest F-ratio was reported for an interaction of codec and bit rate ($F = 77$, $p < 0.001$), followed by an interaction of signal and bit rate ($F = 34$, $p < 0.001$) and codec and signal ($F = 5.4$, $p < 0.001$). To summarize, the results of the ANOVA test revealed that subjects were more sensitive to the bit rates than to all the investigated codecs and signals, and highly statistically significant interactions between all the investigated factors were found.

III. Objective Tests

In this section the subjective results are compared to the predictions made with two different models, PEAQ (Basic and Advanced Version) and POLQA Music. The basic approach in both methods is the same, the original and degraded signals are mapped onto an internal representation using a perceptual model after which the difference in this representation is interpreted by a cognitive model to predict the perceived audio quality of the degraded signal (see Fig. 2). This perceived listening quality is expressed in terms of an Objective Difference Grade (ODG) in the case of PEAQ [8], [13], [14] and in terms of an Objective Mean Opinion Score (OMOS) in the case of POLQA Music [23]–[25]. The internal representations that are used by the cognitive model to predict the perceived audio

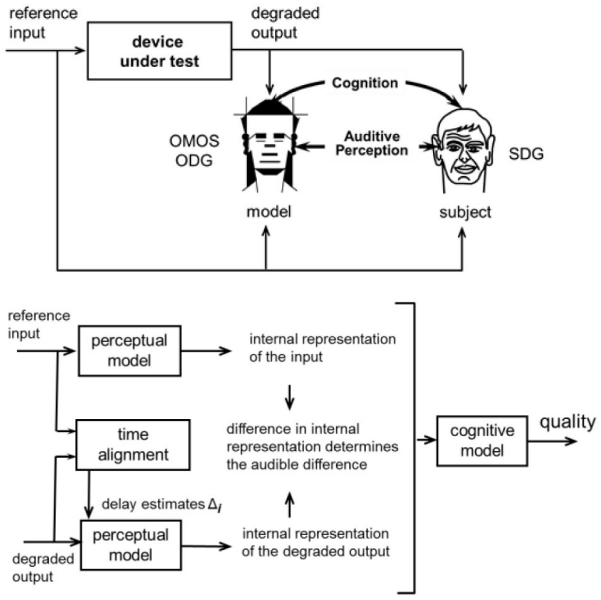

Fig. 2. Overview of the basic philosophy used in PEAQ and POLQA. A model of the subject, consisting of a perceptual and a cognitive model, is used to compare the degraded output with the reference input, using alignment information as derived from the time signals in the time alignment module. The subjective MOS is referred to as subjective difference grade (SDG), the objective score as objective difference grade (ODG) in the case of PEAQ and objective mean opinion score (OMOS) in the case of POLQA.

quality are calculated on the basis of time-frequency-intensity representations (windowed FFT) that use the psychophysical equivalents of frequency (pitch measured in Barks) and intensity (loudness measured in Sones).

Details of the PEAQ model can be found in [13] while POLQA Music is still under development. Two versions of POLQA Music are evaluated, both using the ideas as given in [13] and other relevant audio quality literature but then applied to POLQA [23], [24]. The major differences between POLQA and POLQA Music are:

- The idealization process in which noise in the reference speech signal is suppressed is removed. Speech quality experiments often use recordings with slight audible recording noise (either from the recording equipment or from breathing like sounds and body movements) which is sometimes suppressed in the device under test, making it necessary to suppress the noise in the calculation of the internal representation of the reference signal.
- The suppression of steady state noise in the degraded signal is reduced. In music steady state noise also has less impact on the perceived quality than time varying noise, however the amount of suppression needed in POLQA Music is significantly less. Steady state noise has more impact on the perceived quality of music signals than on the perceived quality of speech signals.
- It turned out that the impact of linear frequency response distortions is more severe with speech quality evaluation, probably because subjects use a priori knowledge of the ideal timbre of speech signals. A music fragment that contains a severe linear frequency response distortion still sounds acceptable as long as the overall timbre is

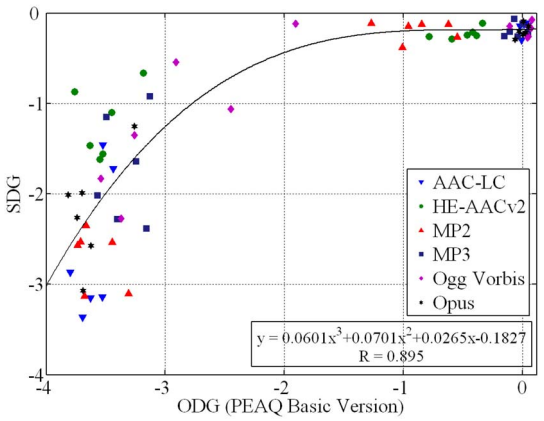

Fig. 3. Correlation between the subjective results (SDG values) and PEAQ Basic Version predictions (ODG values).

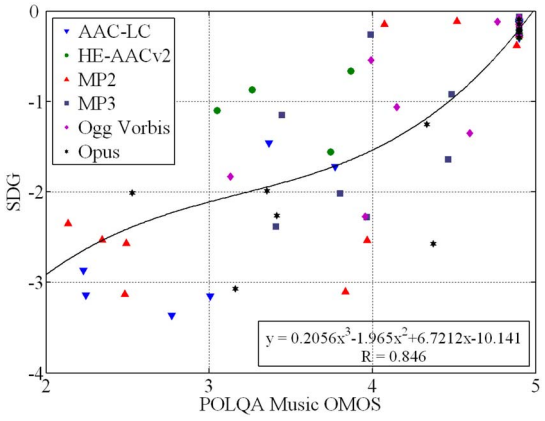

Fig. 5. Correlation between the subjective results (SDG values) and POLQA Music predictions (POLQA Music OMOS values).

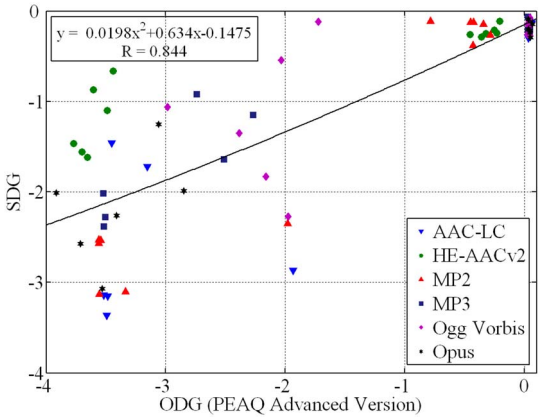

Fig. 4. Correlation between the subjective results (SDG values) and PEAQ Advanced Version predictions (ODG values).

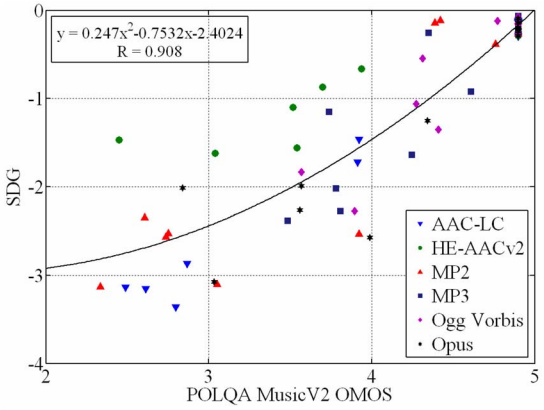

Fig. 6. Correlation between the subjective results (SDG values) and POLQA MusicV2 predictions (POLQA MusicV2 OMOS values).

maintained. In POLQA Music the partly linear frequency response distortions compensation, which suppresses linear frequency response distortions, is thus stronger than in POLQA.

- POLQA partly compensates amplitude variations as introduced by automatic level control mechanisms in speech communication devices. In POLQA Music the amplitude variations are needed to be fully compensated. Apparently (small) amplitude variations do occur in music coding but they are not interpreted as degradations.
- The voice activity detector as used in POLQA is still used in POLQA Music in the form of a simple level detector that detects soft intervals (loudness below 60 dBA) which are used in the noise level estimation processes.

The first version of POLQA Music is trained on the same audio material as PAQM [8] while the second version was retrained on all available data including the data of the current experiment. A retraining of PEAQ was not carried out because of the neural network that is used in PEAQ, they are too easily overtrained. POLQA only uses a basic perceptual model that suffers less from overtraining. All models were run using a sampling rate of 48 kHz.

Figs. 3–6 compare the SDG values per audio fragment obtained from the test described in Section II with the PEAQ (Basic and Advanced Version) and POLQA Music predictions respectively. In order to model the experimental context of a particular experiment, 2nd/3rd order monotonic regressions were used in the calculation of the correlation between the SDG and predicted ODG (Objective Difference Grade) or POLQA Music OMOS values. This context dependent mapping was also used in the standardization process of the PEAQ and POLQA models and allows to ignore bias and context effects which are caused by experimental context. More detail about the context dependent mapping can be found in [33].

It can be observed from Fig. 3 that the Basic Version of PEAQ model performs on an acceptable level with a correlation between the SDG and ODG values of 0.895. The performance of the Advanced Version of PEAQ model depicted in Fig. 4 is a bit worse with a correlation of 0.844.

Figs. 5 and 6 show that POLQA Music correlates on about the same level as PEAQ with the best overall performance for the retrained POLQA MusicV2. As the POLQA Music is still under development, the correlations obtained in this experiment are quite promising, which makes POLQA Music a good candidate for predicting the quality perceived by the end user in the digital audio broadcasting systems and web-casting applications.

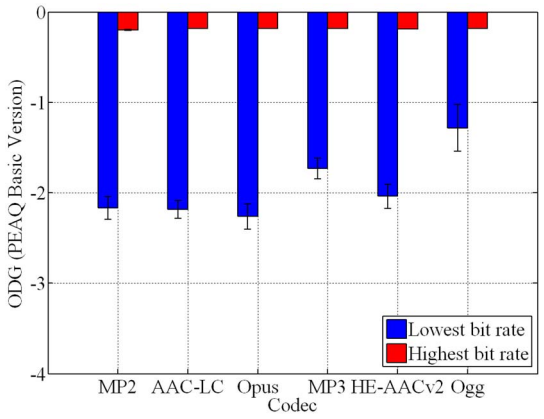

Fig. 7. Effect of codecs and bit rates (see Table III for more information about the exact lowest and highest bit rates used for a particular codec) on average ODG values predicted by PEAQ Basic Version. The vertical bars show 95% CI computed over six ODG values (six audio signals used in the test).

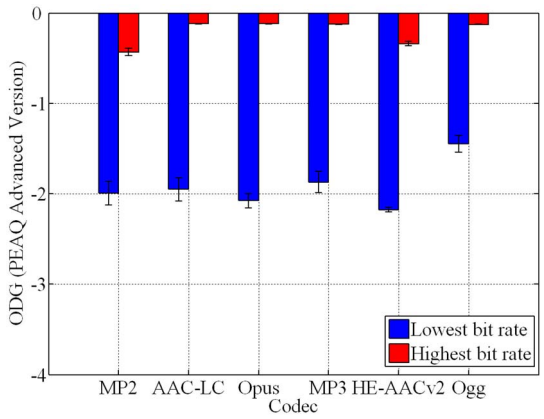

Fig. 8. Effect of codecs and bit rates (see Table III for more information about the exact lowest and highest bit rates used for a particular codec) on average ODG values predicted by PEAQ Advanced Version. The vertical bars show 95% CI computed over six ODG values (six audio signals used in the test).

Comparison of the regression curves show that POLQA Music shows the smoother behavior over the complete range of coding distortions and that the Basic version of PEAQ has less discriminating power in the lower quality range. The strong non-linear behavior of PEAQ Basic makes it more difficult to retrain this model for correct behavior on the PEAQ databases as well as on our DAB database.

As differences between the reported correlations for all the investigated models are small, the corresponding statistical significance test (see [33]) was performed to specify the significance of the differences between them. It was found that all the differences are statistically not significant. It means that all the investigated models produce statistically equivalent results.

When we use the optimal regression curves to calculate the context dependent scores per codec the correlations increase significantly because the objective models no longer need to model the impact of degradations that are specific to a certain musical fragment. Figs. 7–10 provides the result for all four models using the same graphical representation

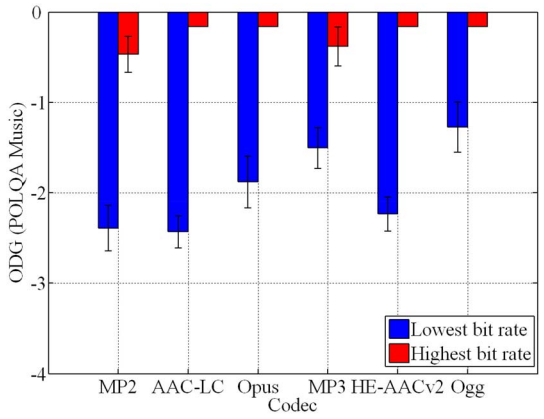

Fig. 9. Effect of codecs and bit rates (see Table III for more information about the exact lowest and highest bit rates used for a particular codec) on average ODG values predicted by POLQA Music. The vertical bars show 95% CI computed over six ODG values (six audio signals used in the test).

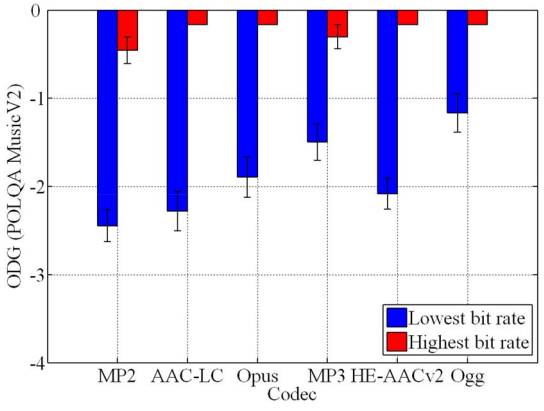

Fig. 10. Effect of codecs and bit rates (see Table III for more information about the exact lowest and highest bit rates used for a particular codec) on average ODG values predicted by POLQA MusicV2. The vertical bars show 95% CI computed over six ODG values (six audio signals used in the test).

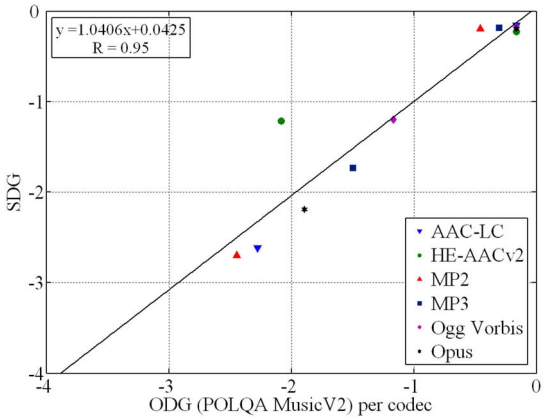

Fig. 11. Correlation between the subjective results (SDG values) and ODG values per codec predicted by POLQA MusicV2.

as Fig. 1 while Fig. 11 provides the per codec scatterplot for the retrained POLQA Music (POLQA MusicV2). When comparing Figs. 7–10 with Fig. 1, we can clearly see that a ranking provided by POLQA MusicV2 best matches a ranking obtained from the subjective test,

TABLE IV
SUMMARY OF ANOVA TEST CONDUCTED ON THE SDG VALUES

Effect	SS	df	MS	*F*	*p*
Codec	136.19	5	27.24	72.53	0.00
Signal	85.87	5	17.17	45.74	0.00
Bit rate	1161.48	1	1161.48	3093.12	0.00
Codec*Signal	50.9	25	2.04	5.42	0.00
Codec*Bit rate	144.61	5	28.92	77.02	0.00
Signal*Bit rate	64.6	5	12.92	34.41	0.00
Error	550.11	1465			
Total	2193.75	1511			

especially when comes to the lowest investigated bit rate (the most critical condition). Moreover, Fig. 11 shows that POLQA MusicV2 is capable of predicting the codec performance very accurately except for the low bit rate version of the HE-AACv2 codec. Current investigations are focused on improving this behavior of the POLQA Music model.

IV. CONCLUSION

In this paper the impact of degradations introduced by six lossy audio codecs typically deployed in current digital audio broadcasting systems and web-casting applications on the audio quality perceived by the end user was exhaustively investigated by a subjective testing approach defined in ITU-R BS.1116-1. Furthermore the performance of two audio quality prediction models, PEAQ and POLQA Music, were checked by comparing the predictions with the SDG values obtained from the subjective test described in this paper.

A first conclusion is that the degradations introduced by the lossy audio codecs typically deployed in current digital audio broadcasting systems and web-casting applications operating at the lowest investigated bit rate have a serious impact on the perceived audio quality. A second conclusion is that there is a negligible difference between all the investigated codecs at the highest investigated bit rate in terms of perceived audio quality, all codecs provide a near transparent audio quality. When it comes to the lowest investigated bit rates, the best quality is provided by Ogg Vorbis codec at 64 kbps representing web-casting applications, very closely followed by HE-AACv2 codec at 24 kbps characterizing broadcasting systems. It is worth noting here that the lowest and highest investigated bit rates correspond to those typically deployed in current digital audio broadcasting systems and web-casting applications. A third conclusion is that a retrained version of POLQA Music allows reasonable accurate predictions of the perceived audio quality of the codecs typically deployed in current digital audio broadcasting systems and web-casting applications, especially when scores are averaged over a small set of musical fragments.

APPENDIX

Table IV shows the results of the ANOVA test carried out on the subjective data (Dependent variable: SDG) described in more detail in Section II-B.

REFERENCES

[1] *Radio Broadcasting Systems; Digital Audio Broadcasting (DAB) to Mobile, Portable and Fixed Receivers*, ETSI Standard EN 300 401, 2006.

[2] W. Hoeg and T. Lauterback, *Digital Audio Broadcasting; Principles and Applications of DAB, DAB+ and DMB*, 3rd ed. Chichester, U.K.: Wiley, 2009.

[3] *Digital Radio Mondiale (DRM); System Specification*, ETSI Standard ES 201 980, 2012.

[4] F. Hofmann, C. Hansen, and W. Schafer, "Digital radio mondiale (DRM) digital sound broadcasting in the AM bands," *IEEE Trans. Broadcast.*, vol. 49, no. 3, pp. 319–328, Sep. 2003.

[5] S. Cho *et al.*, "System and service of terrestrial digital multimedia broadcasting (T-DMB)," *IEEE Trans. Broadcast.*, vol. 53, no. 1, pp. 171–178, Mar. 2007.

[6] J. Berg, C. Bustad, L. Jonsson, L. Mossberg, and D. Nyberg, "Perceived audio quality of realistic FM and DAB+ radio broadcasting systems," *J. Audio Eng. Soc.*, vol. 61, no. 10, pp. 755–777, 2013.

[7] *Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems*, ITU-R Recommendation BS.1116-1, Int. Telecommun. Union, Geneva, Switzerland, 1997.

[8] J. G. Beerends and J. A. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 40, no. 12, pp. 963–974, 1992.

[9] B. Paillard, P. Mabilleau, S. Morisette, and J. Soumagne, "PERCEVAL: Perceptual evaluation of the quality of audio signals," *J. Audio Eng. Soc.*, vol. 40, pp. 21–31, Feb. 1992.

[10] C. Colomes, M. Lever, J. B. Rault, and Y. F. Dehery, "A perceptual model applied to audio bit-rate reduction," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 233–240, 1995.

[11] T. Thiede and E. Kabot, "A new perceptual quality measure for bit rate reduced audio," in *Proc. 100th Audio Eng. Soc. Conv.*, Copenhagen, Denmark, 1996.

[12] T. Sporer, "Objective audio signal evaluation—Applied psychoacoustics for modelling the perceived quality of digital audio," in *Proc. 103rd Audio Eng. Soc. Conv.*, New York, NY, USA, 1997.

[13] T. Thiede *et al.*, "PEAQ—The ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, pp. 3–29, Feb. 2000.

[14] *Method for Objective Measurements of Perceived Audio Quality*, ITU-R Recommendation BS 1387-1, Int. Telecommun. Union, Geneva, Switzerland, 1999.

[15] J. Barbedo and A. Lopes, "A new cognitive model for objective assessment of audio quality," *J. Audio Eng. Soc.*, vol. 53, pp. 22–31, Feb. 2005.

[16] S. Zielinski, F. Rumsey, R. Kassier, and S. Bech, "Development and initial validation of a multichannel audio quality expert system," *J. Audio Eng. Soc.*, vol. 53, pp. 4–21, Feb. 2005.

[17] J.-H. Seo, S. B. Chon, K.-M. Sung, and I. Choi, "Perceptual objective quality evaluation method for high-quality multichannel audio codecs," *J. Audio Eng. Soc.*, vol. 61, pp. 535–545, Jul. 2013.

[18] R. Huber and B. Kollmeier, "PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception," *IEEE Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[19] M. Hansen and B. Kollmeier, "Objective modeling of speech quality with a psychoacoustically validated auditory model," *J. Audio Eng. Soc.*, vol. 48, no. 5, pp. 395–409, 2000.

[20] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[21] N. Harlander, R. Huber, and S. D. Ewert, "Sound quality assessment using auditory models," *J. Audio Eng. Soc.*, vol. 62, no. 5, pp. 324–336, 2014.

[22] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *J. Acoust. Soc. America*, vol. 124, no. 1, pp. 422–438, 2008.

[23] J. G. Beerends *et al.*, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I—Temporal alignment," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 366–384, 2013.

[24] J. G. Beerends *et al.*, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part II—Perceptual model," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 385–402, 2013.

[25] *Perceptual Objective Listening Quality Assessment*, ITU-T Recommendation P.863, Int. Telecommun. Union, Geneva, Switzerland, 2011.
[26] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application*. Chichester, U.K.: Wiley, 2006.
[27] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality-technology and applications," *IEEE Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1890–1901, 2006.
[28] S. Lee *et al.*, "An audio quality evaluation of commercial digital radio systems," *IEEE Trans. Broadcast.*, vol. 57, no. 3, pp. 629–635, Sep. 2011.
[29] W. C. Treurniet and G. A. Soulodre, "Evaluation of the ITU-R objective audio quality measurement method," *J. Audio Eng. Soc.*, vol. 48, no. 3, pp. 164–173, 2000.
[30] G. A. Soulodre, T. Grusec, M. Lavoie, and L. Thibault, "Subjective evaluation of state-of-the-art two-channel audio codecs," *J. Audio Eng. Soc.*, vol. 46, no. 3, pp. 164–177, 1998.
[31] EBU SQAM CD. (2008). *Sound Quality Assessment Material Recordings for Subjective Tests*. [Online]. Available: https://tech.ebu.ch/publications/sqamcd
[32] *Sound Quality Assessment Material Recordings for Subjective Tests: Users' Handbook for the EBU SQAM CD*, EBU Document TECH 3253, Eur. Broadcast. Union, Geneva, Switzerland, 2008.
[33] *Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*, ITU-T Recommendation P.1401, Int. Telecommun. Union, Geneva, Switzerland, 2012.

Peter Počta received the M.Sc. and Ph.D. degrees in telecommunication from the Faculty of Electrical Engineering, University of Őilina, Őilina, Slovakia, in 2004 and 2007, respectively. He was an Erasmus Student with the Department of Electrical Engineering and Information Technology, Chair of Telecommunications, Dresden University of Technology, Germany, for three months, where he collaborated on testing principles over ADSL access lines. He was with Alcatel-Lucent, Researcha and Development Center, Network Integration department, Stuttgart, Germany, where he investigated an impact of some settings of WiMAX system on speech quality. He is currently an Associate Professor with the Department of Telecommunications and Multimedia, University of Őilina, involved with International Standardization through the ETSI TC STQ as well as ITU-T SG12. His current research interests include speech, audio and video quality assessment, software-defined networking, and cross-layer QoE management. He has published over 30 peer-reviewed papers in international journals and conferences including Acta Acustica united with Acustica, *Computer Standards and Interfaces* (Elsevier), *AEÜ—International Journal of Electronics and Communications* (Elsevier) and MESAQIN and QoMEX conferences. He was a recipient of a number of fellowships. He serves as an external reviewer for the *Journal of Systems and Software* (Elsevier), *Computer Standards and Interfaces* (Elsevier), *Speech Communication* (Elsevier), *Telecommunication Systems* (Springer), and several conferences in area of speech quality and communication networks.

John G. Beerends received a degree in electrical engineering from the HTS (Polytechnic Institute) of The Hague, The Netherlands, in 1975, the M.Sc. degree from the University of Leiden in 1984. In 1983 he was awarded a prize of DFl 45000 by Job Creation for the further development of his patented asymmetric loudspeaker enclosure design. From 1984 to 1989 he worked at the Institute for Perception Research where he received a Ph.D. from the Technical University of Eindhoven in 1989. The main part of his doctoral work, which deals with pitch perception, was published in the Journal of the Acoustical Society of America. The results of this work led to a patent on a pitch meter by the N.V. Philips Gloeilampenfabriek. From 1986 to 1988 he worked on a psycho-acoustically optimized loudspeaker system for the Dutch loudspeaker manufacturer BNS. The system was introduced at the Dutch consumer exhibition FIRATO in 1988. In 1989 he joined the KPN Research where he worked on audio and video quality assessment, audio-visual interaction, and on audio coding (speech and music). This work led to several patents and two measurement methods for objective, perceptual, assessment of audio quality which he developed together with Jan Stemerdink. The first one dealt with telephone-band speech and was standardized in 1996 as ITU-T Recommendation P.861 (Perceptual Speech Quality Measure, PSQM), the second one with wide-band audio and was integrated into ITU-R Rec. BS.1387 (1998, Perceptual Evaluation of Audio Quality, PEAQ). Most of the work on audio quality (speech, music and audiovisual interaction) was published within the Audio Engineering Society and the ITU. From 1996 to 2002 he worked with Andries Hekstra on the objective measurement of the quality of video and speech. The work on speech quality, partly carried out with researchers from British Telecom, was focussed on improving PSQM and was standardized in 2001 as ITU-T Rec. P.862 (Perceptual Evaluation of Speech Quality, PESQ). The work on video quality led to several patents and a measurement method for objective, perceptual, assessment of video quality, standardized in 2008 by the ITU-T as Rec. J.247 (Perceptual Evaluation of Video Quality, PEVQ). In January 2003 he joined TNO, which took over the research activities from KPN, where he worked on the objective measurement of speech intelligibility, (super) wideband speech quality, degradation decomposition, hearing aid quality, videophone quality and data chirping techniques. The main focus was on speech quality and intelligibility assessment for the normal hearing and hearing impaired. In the period 2006-2007 he worked on the development of the Perceptual Hearing Aid Quality Measure (PHAQM) which proved to be the best predictor of speech quality in a benchmark carried out by a consortium of leading hearing aid manufacturers. In the period 2003-2010 he worked on the development of the follow up of PESQ P.862. In a joint effort with OPTICOM and SwissQual this work resulted in ITU-T Rec. P.863 (Perceptual Objective Listening Quality Assessment, POLQA) in 2011. Currently he is working on extending the perceptual measurement approach towards acoustic domain measurements (loudspeaker reproduction quality, including the impact of the reproduction room) and on the glass box modelling of audio, speech, video, data services. John Beerends is the (co-) author of more than 100 (conference) papers/ITU contributions and 35 patents. In 2003 he received an AES fellowship award for his work on audio and video quality measurement.