6th Congress of Alps-Adria Acoustics Assosiation
16.-17. October 2014
Graz, Austria

Austrian Acoustics Association
Österreichische Gesellschaft für Akustik

Alps Adria Acoustics Association

ISBN: 978-3-200-03921-6

AUDIO QUALITY: COMPARISON OF PEAQ AND FORMAL LISTENING TEST RESULTS

Markus Zaunschirm*, Matthias Frank*, Alois Sontacchi*, Paolo Castiglione†

* Institute of Electronic Music and Acoustics (zaunschirm@iem.at)
† AKG Acoustics GmbH (Harman International)

Abstract: *PEAQ (Perceptual evaluation of audio quality) is an international standard for quality prediction of wide-band audio codecs (coder - decoder) according to ITU-R BS.1387, developed by an international consortium of leading audio quality experts in 1999. The commercially available implementation of PEAQ offers two analysis models (basic and advanced) with different algorithmic complexity. This contribution compares predictions obtained from PEAQ with quality evaluation results from previous listening experiments employing a trained panel of 39 expert listeners. The evaluated stimuli were generated using a proprietary sub-band ADPCM (Adaptive Differential Pulse Code Modulation) codec for digital wireless transmission with multiple settings and various audio signals (bass, orchestra, speech, triangle, trumpet, and vocals). In addition to two commercial PEAQ implementations, the comparison includes an open-source implementation of ITU-R BS.1387. While PEAQ predictions can be used to identify tendencies in the lower quality range, they seem to be less sensitive to quality differences at the upper-end of the grading scale.*

Keywords: audio quality assessment, audio codec, perceptual evaluation of audio quality

1. INTRODUCTION

Lossy audio compression is used in digital communication systems or audio applications, such as wireless microphones and headphones. Such lossy audio coding algorithms usually take advantage of the properties of the human auditory system and the reduction in bit-rate is achieved by removing redundant and perceptually irrelevant data from the audio signal. Thus, the introduced coding error (difference between decoded and reference signal) is physically present but may be inaudible due to its spectro-temporal distribution. As conventional measures, such as signal-to-noise ratio (SNR) or average spectral distortion do not reliably relate to the perceived audio quality, the need of conducting formal listening tests is evident. In the ITU-R recommendation BS-1116 [1] a procedure for subjective assessment of small impairments in audio systems is suggested. It recommends a double-blind triple-stimulus with hidden reference test. During the test, subjects are asked to rate the differences of the test stimuli compared to a reference signal on a five grade impairment scale ranging from very annoying to imperceptible.
However, formal listening tests are generally time consuming, expensive and require a large enough listening panel. Thus, a technical measurement method that models the actual perception is preferred, especially during the development process of a codec (encoder-decoder).
More than a decade ago, the perceptual evaluation of audio quality (PEAQ, [2] [3]) was approved by the ITU-R committee as a technical method for assessing the perceived audio quality that is based on a model of the human auditory system. The proposed method outputs the objective difference grade that should relate to the results obtained from a formal listening test. But how do the results correlate and are there limitations when using PEAQ? These questions may be answered by comparing audio quality ratings obtained from a previously conducted formal listening test [4] against ratings from three different PEAQ implementations, where two of them are commercially available [5] and one is open-source [6].
This paper is arranged as follows: Section 2 gives a summary of audio quality assessment methods (formal listening tests and PEAQ). In the subsequent section, we compare data from a listening test against predictions from the PEAQ implementations. Results are shown for different audio samples as well as for a combined sample pool consisting of 159 items overall. Finally, the paper is summarized and concluded in section 4.

2. AUDIO QUALITY ASSESSMENT

Besides computational complexity and overall latency, the achievable audio quality is the key factor when evaluating the performance of an audio codec. Methods for assessing the perceived audio quality include formal listening tests or a technical measure that predicts the perceived audio quality by modelling the human auditory system. Both methods compare the perception of a processed signal against an unprocessed reference signal (see Fig. 1) and give the overall audio quality as output measure.

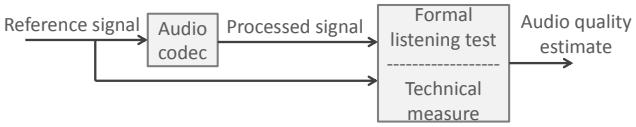


Fig. 1. Concept of audio quality evaluation.

2.1. Listening experiments

The recommendations in [1] address the selection of test materials and the listening panel, as well as the test method. A double-blind triple-stimulus with hidden reference method has been found to be most sensitive to small quality differences yielding robust results. Thereby, the reference signal is always available as stimulus REF, while the hidden reference and the test signal are randomly assigned to stimulus A or B. During the test, subjects can switch between the three presented stimuli (REF, A, B) and are asked to assess the quality differences of stimulus A or B compared to REF, respectively. The quality assessment of the hidden reference stimulus allows for simple screening of intra-subject reliability. A five-grade continuous impairment scale form *imperceptible/very good* (5) to *very annoying/very bad* (1) is used for quality ratings and any perceived differences shall be interpreted as a decrease of audio quality. An exemplary graphical user interface (GUI) is depicted in Fig. 2.

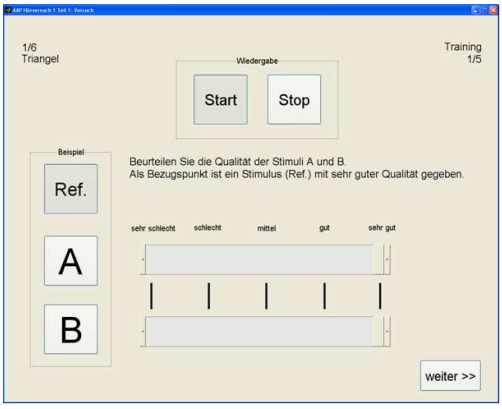


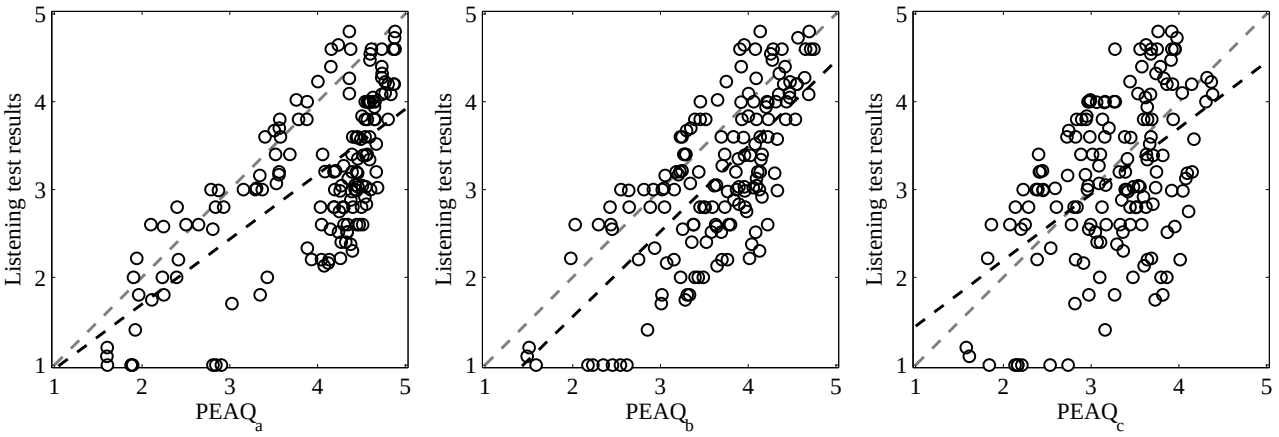
Fig. 2. GUI from listening experiment: Double-Blind Triple-Stimulus with Hidden Reference method on a 5-point continuous scale from *very bad* ("sehr schlecht") to *very good* ("sehr gut"): Rate the quality of the stimuli A and B. As reference, there is a stimulus with very good quality given.

2.2. PEAQ - Perceptual Evaluation of Audio Quality

The PEAQ recommendation incorporates two different versions [2]. The basic version is intended for cost-efficient real-time applications, whereas the advanced version is intended to yield the highest possible accuracy. Both versions generate multiple model output variables (MOVs) that are fed into a neuronal network that computes the final quality measure.
The MOVs used in the basic version are computed using an FFT-based ear model and overall 11 different MOVs are used for the prediction of the perceived audio quality. The advanced version uses both MOVs calculated from a filter-bank-based and an FFT-based ear model, but only 5 MOVs are considered to compute a single quality estimate. The neuronal network that maps the MOVs to a final prediction of audio quality was calibrated using data from various formal listening tests that were conducted in accordance to the ITU-R BS-1116 methodology. Validation tests showed a slightly higher accuracy for predictions of the advanced algorithm than for the basic version. Overall, correlation between results from listening tests and PEAQ predictions is in the range of 80% [3].
For further reading and exact definitions of MOVs, settings of FFT-based and filter-bank-based ear models, as well as the used cognitive model the reader is referred to [2].

3. COMPARISON OF LISTENING TEST RESULTS AND PEAQ PREDICTIONS

In order to find out whether the PEAQ measures can be used by a developer of audio codecs to reliably compare the performance of different codecs or different settings for one codec, we compare the audio quality ratings obtained from a formal listening test (abbreviated as LTR) [4] against the ratings obtained from three different PEAQ implementations (abbreviated as $PEAQ_a$, $PEAQ_b$, $PEAQ_c$). $PEAQ_a$ and $PEAQ_b$ are commercially available [5] implementations of the advanced and basic PEAQ algorithms, and $PEAQ_c$ is an open source implementation [6] of the basic algorithm, respectively.
Performance metrics are computed using PEAQ predictions and the median value of the subject ratings and they include the coefficient of determination (R^2), the mean squared error (MSE) and the slope of the regression line (β). For visual evaluation, the LTR (median of subject ratings) are plotted against the PEAQ predictions.
The conducted listening test [4] evaluated the audio quality of a codec that incorporates a proprietary sub-band ADPCM (Adaptive Differential Pulse Code Modulation) algorithm and achieves an end-to-end latency below 5 ms. Internally, the sub-bands are quantized independently with adjustable resolutions. Thus, the aim of the listening experiment was to determine the most efficient combinations of resolutions in all frequency bands for each sound, i.e. the smallest overall bit rate necessary to obtain a certain degree of audio quality.



(a) $R^2 = 0.49, MSE = 1.14, \beta = 0.74$ (b) $R^2 = 0.52, MSE = 0.65, \beta = 0.97$ (c) $R^2 = 0.25, MSE = 0.67, \beta = 0.67$

Fig. 3. Relationship between LTR and PEAQ results for entire sample pool consisting of 159 items. Gray dashed line indicates perfect correspondence.

Six different sounds from the EBU SQAM recordings [7] were tested: bass, orchestra, speech, triangle, trumpet and vocals. In accordance to the ITU-R BS.1116 [1], a double-blind triple-stimulus with hidden reference method with ratings on a 5-point continuous scale (see Fig. 2) was used for the second experiment. The stimuli were presented over STAX SR4040 II headphones at a level of 70 dB Leq(A). A preliminary test allowed for a limitation of the trial number to 159 and yielded a duration of the experiment of approximately one hour. The listening test was carried out with a trained expert listening panel of 39 subjects [8, 9, 10]. In only 0.9% of the trials, the original reference sound was not rated with the highest possible quality. A detailed description of the listening test can be found in [4].

3.1. All Sound Samples

In this section the correspondence between LTR and PEAQ predictions is analyzed for the entire sample pool consisting of all 159 items (all 6 different sound samples). Fig. 3 depicts the median values of quality ratings in dependence of the corresponding PEAQ values. Clearly, perfect correspondence between PEAQ predictions and LTR is not obtained by any implementation, as not all data points fall on the main diagonal (marked as gray dashed line). In detail, squared correlations between LTR and the three PEAQ predictions are 0.49, 0.52 and 0.25, respectively.

It can be seen in Fig. 3(a), that predictions from $PEAQ_a$ (advanced algorithm) tend to overrate audio quality. This tendency is underlined by the distribution of the quality ratings, cf. Fig. 4: Approximately 65% of ratings are clustered between 4 and 5, whereas experienced listeners rated only about 18% of audio items as having good to very good audio quality.

Approximately, 80% of predictions form $PEAQ_b$ (basic algorithm) are concentrated at the upper half of the grading scale (see Fig. 4). Thus, audio quality tends to be overrated again. However, when examining Fig. 3(b) there are no clustered outliers apparent as predictions are evenly scattered around the regression line.

Results obtained from implementation $PEAQ_c$ (basic algorithm, open-source) show the smallest correspondence to the listening test results. It is also striking when examining Fig. 3(c) that ratings are concentrated around the middle, with very few ratings at the upper and lower end of the grading scale. Moreover, predictions between $3 \ldots 4$ are obtained for audio items that were rated between $2 \ldots 5$ by the expert listeners. Thus, results from implementation $PEAQ_c$ are highly scattered.

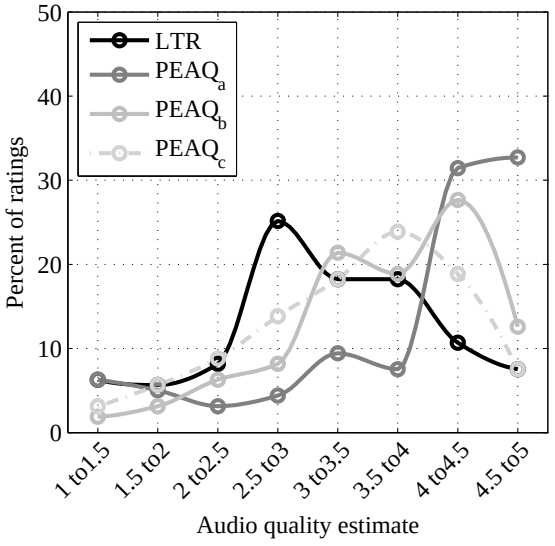
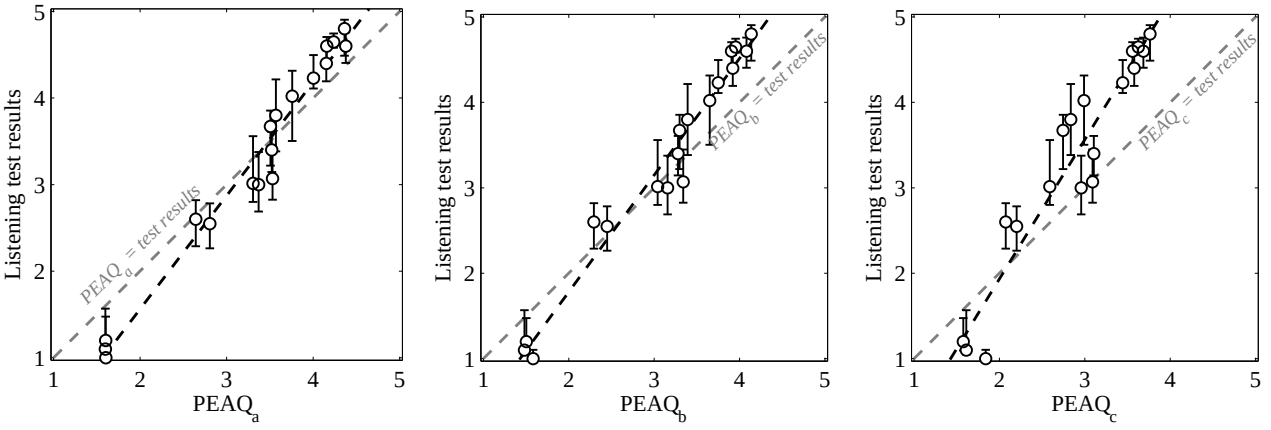


Fig. 4. Distribution of quality ratings for different PEAQ implementations and listening test results.

High-quality Settings

Generally, state-of-the-art audio codecs aim at yielding excellent audio quality at moderate data rates and complexity. Accordingly, fine-tuning of codec parameters will most



(a) $R^2 = 0.97, MSE = 0.14, \beta = 1.31$ (b) $R^2 = 0.97, MSE = 0.21, \beta = 1.37$ (c) $R^2 = 0.91, MSE = 0.62, \beta = 1.63$

Fig. 5. Relationship between LTR and PEAQ results for a speech sample coded with 18 different resolution settings. Gray dashed line indicates perfect correspondence.

likely result in minor changes of audio quality at the upper end of the grading scale. Anyway, audio items coded with low quality are easily identified by informal listening and thus, neither formal listening tests nor PEAQ predictions are necessary in this case. In contrast, a formal assessment of audio quality is essential to resolve small differences of audio quality for codecs that obtain good audio quality ratings. The ability of PEAQ to predict audio quality in the range of interest is evaluated by considering only the audio items of the sample pool that were rated between $4 \ldots 5$ by the expert listening panel (35 items out of the overall 159). The obtained values for R^2 between LTR and the different PEAQ implementations decreased from $[0.49, 0.52, 0.25]$ for the entire data set to $[0.03, 0.06, 0.07]$ for items that are rated between good and very good.

3.2. Single-Sample Subset

The global PEAQ predictions using the pool of all samples appear to be unsatisfying and thus, this section analyzes the performance of PEAQ algorithms for each sample subset, separately. As performance depends on the chosen audio sample (see Tab. 1), we focus on the best (speech) as well as the worst case (triangle) out of the sample pool.
Fig. 5 shows the relation between LTR and PEAQ results for a speech sample coded with 18 different resolution settings. The correlation between LTR and PEAQ predictions is in a range from $0.98 \ldots 0.95$ across the different implementations. By tendency, the predicted values do not exploit the lower and higher end of the grading scale, yielding an increased occurrence probability around the center.

	bass	orch	spch	tria	trmp	vox
$PEAQ_a$	0.93	0.80	0.97	0.60	0.65	0.87
$PEAQ_b$	0.63	0.81	0.97	0.82	0.77	0.70
$PEAQ_c$	0.10	0.86	0.90	0.51	0.51	0.49

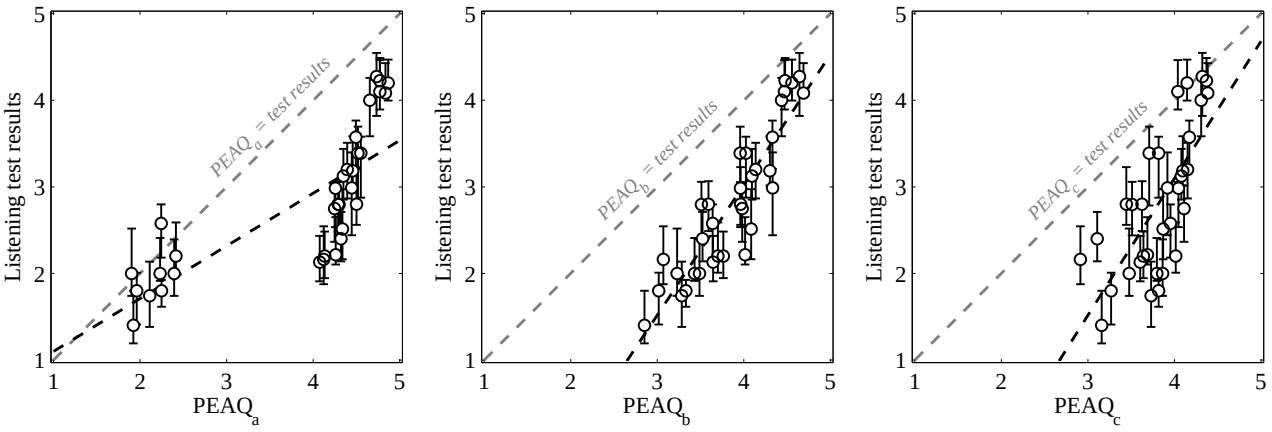
Table 1. Determination coefficient R^2 for different audio samples.

Overall the different PEAQ implementations yield satisfying reliability for predicting the audio quality of the processed speech samples, with $PEAQ_a$ yielding the best match.
The relation between PEAQ predictions and LTR for a triangle sample coded with overall 33 different coding settings is depicted in Fig. 6. It can be seen that the performance for the triangle sample is highly degraded when compared against the performance for the speech sample (see Tab. 1). The results from $PEAQ_a$ are clustered around 2.5 and 4.5 and $PEAQ_b$ and $PEAQ_c$ predict the audio quality of all samples between $3 \ldots 5$ without any ratings at the low end of the scale. Furthermore, for the triangle samples, PEAQ rates the audio quality higher as the expert listeners, what is in contrary to the predictions for the speech sample (cf. slope of regression line). Hence, performance of PEAQ seems to be highly dependent on signal characteristics (bandwidth and temporal resolution) and thus, an evaluation for a combined sample pool is not recommended. Moreover, PEAQ predictions must be interpreted on a relative scale as they do not give a rating of audio quality on an absolute scale.

High-quality Settings
Again the correspondence between PEAQ predictions and LTR for audio items that are rated between good and very good by the expert listening panel is examined. The resulting determination coefficients are listed in Tab. 2. While correspondence for orchestra, speech and vocal samples is comparable to the overall results, the correspondence for bass, triangle and trumpet samples is highly decreasing when compared against results that include the entire ratings.

	bass	orch	spch	tria	trmp	vox
$PEAQ_a$	0.09	0.91	0.87	0.11	0.05	0.78
$PEAQ_b$	0.21	0.56	0.87	0.11	0.16	0.75
$PEAQ_c$	0.30	0.86	0.84	0.01	0.16	0.60

Table 2. Determination coefficient R^2 for items rated between good and very good.



(a) $R^2 = 0.60, MSE = 1.53, \beta = 0.60$ (b) $R^2 = 0.82, MSE = 1.36, \beta = 1.50$ (c) $R^2 = 0.51, MSE = 1.47, \beta = 1.58$

Fig. 6. Relationship between LTR and PEAQ results for a triangle sample coded with 33 different resolution settings. Gray dashed line indicates perfect correspondence.

4. CONCLUSION

We compared the audio quality ratings from a previously conducted listening experiment (39 expert listeners) against audio quality predictions form three different PEAQ implementations. The entire data set included 159 audio items that were generated from 6 different samples (bass, orchestra, speech, triangle, trumpet and vocals) using a sub-band low-delay audio codec with different bit-rate settings.

If the entire sample pool is considered, the determination coefficients (R^2) between listening test results and PEAQ predictions lie between $0.25 \ldots 0.52$. Furthermore, we showed that PEAQ prediction performance varies depending on the chosen audio sample and as the three implementations overrate or underrate (cf. slope of regression line) the perceived audio quality depending on the sample, PEAQ ratings must be interpreted on a relative scale.

Differences between audio samples that are rated between good and very good by an expert listening panel are not reliable reproduced by PEAQ predictions, at least not for all sample subsets (cf. triangle, trumpet). Consequently, formal listening tests are essential for identifying differences between high-quality audio codecs. However, PEAQ predictions may be used during development process to identify rough relative changes of audio quality.

ACKNOWLEDGEMENT

This work was supported by the projects AAP and ASD, which are funded by Austrian ministries BMVIT, BMWFJ, the Styrian Business Promotion Agency (SFG), and the departments 3 and 14 of the Styrian Government. The Austrian Research Promotion Agency (FFG) conducted the funding under the Competence Centers for Excellent Technologies (COMET, K-Project), a program of the above-mentioned institutions.

REFERENCES

[1] International Telecommunication Union: ***ITU-R BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems***, 1997.

[2] International Telecommunication Union: ***ITU-R BS.1387: Method for objective measurements of perceived audio quality***, 1998.

[3] T. Thiede, W. Treurniet, and R. Bitto: **PEAQ-The ITU standard for objective measurement of perceived audio quality**, *Journal of the Audio Engineering Society*, 48(1/2), 2000, 3-29.

[4] M. Frank and A. Sontacchi: **Subjective sound quality evaluation of a codec for digital wireless transmission**, in *Audio Engineering Society Convention 132*, 2012.

[5] OPTICOM Gmbh: **Technical Specification for the OPERA Software Suite V3.0 OPR-000-XXX-S**, Tech. Rep., 2001.

[6] P. Kabal: **An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality**, *TSP Lab Technical Report*, 2002.

[7] European Broadcasting Union: ***EBU Tech 3253 - Sound Quality Assessment Material recordings for subjective tests***, 2008.

[8] A. Sontacchi, H. Pomberger, and R. Höldrich: **Recruiting and evaluation process of an expert listening panel**, *NAG/DAGA*, Rotterdam, 2009.

[9] M. Frank, A. Sontacchi, and R. Höldrich: **Training and guidance tool for listening panels**, in *DAGA*, Berlin, 2010.

[10] M. Frank and A. Sontacchi: **Performance review of an expert listening panel**, in *DAGA*, Darmstadt, 2012.