AN EXTENSION OF THE PEAQ MEASURE BY A BINAURAL HEARING MODEL

Magnus Schäfer, Mohammad Bahram, and Peter Vary

Institute of Communication Systems and Data Processing (ind), RWTH Aachen University
{schaefer,bahram,vary}@ind.rwth-aachen.de

ABSTRACT

Instrumental evaluation of the perceived audio signal quality is an important tool for the development of audio signal enhancement and transmission systems. There are various single channel measures which can be used for different application scenarios. Binaural signals have not received much focus so far and no sophisticated model of spatial perception is utilized in the available measures. In this contribution, an extension to Perceptual Evaluation of Audio Quality (PEAQ) is presented which makes use of a recently proposed binaural hearing model. It is shown that the inclusion of spatial information into the instrumental quality measurement leads to a strongly increased correlation between the instrumental measure and a listening test.

Index Terms— Instrumental Quality Measure, Binaural, PEAQ

1. INTRODUCTION

When designing speech and audio signal transmission or enhancement systems, it is an important task to evaluate the perceived quality. The gold standard for this are listening tests [1] which are very flexible as they can be tailored to the system under test (SUT). There are tools available to conduct listening tests (e.g., [2]). During algorithm development however, continuously conducting listening tests for each minor modification is cumbersome and time consuming.

Hence, the instrumental assessment of the perceived quality is a topic that has been receiving continuous interest. An overview on quality assessment in general can be found in [3, 4]. Approaches for the evaluation of multi-channel signals in particular are considered in [5, 6].

This paper presents an extension to PEAQ that is based on a binaural auditory and perceptual model we presented in [7] for the calculation of binaural model output variables (MOVs) which are fed to a neural network (NN) together with the result of PEAQ.

1.1. Relation to prior work

For the evaluation of single channel signal processing systems, PESQ [8, 9] and POLQA [10] are established for speech enhancement and transmission systems. For generic audio signals, PEAQ [11] is a known and reliable measure. Therefore, it is regularly used in different areas of acoustic signal processing [12, 13]. The fundamental principle of PEAQ is the calculation of so-called MOVs of a monaural hearing model, comparing these MOVs of the reference (input) signal and the degraded (output) signal of the SUT and feeding these differences into a NN that is trained based on the known results of numerous listening tests.

PEAQ does offer the possibility to evaluate stereo signals as well. In this case, two monaural hearing models are used in parallel ([11]: "... in the case of stereo signals all computations are performed in the same manner and independently of one for the left and right channel."), i.e., no inter-channel cues are taken into consideration. The MOVs of the two channels are then averaged before the NN and an overall quality for the stereo signal processing system results. The effect of this averaging before the NN is very similar to just using PEAQ separately for the two channels and averaging the final objective difference grade (ODG) values.

It will be shown in this paper that this leads to a fairly poor matching between the estimated quality and the perceived subjective spatial quality for many audio signals. Many lossy transmission systems (e.g., speech and audio codecs) are not able to exactly preserve the positions of the various sources within the auditory scene or they may even discard the spatial information altogether if the available data rate is too small. Since the fidelity of this spatial information is not explicitly considered in PEAQ, this may lead to faulty quality estimates.

2. EXTENDED PEAQ MEASURE FOR BINAURAL SIGNALS

The proposed approach is an *add-on* to the PEAQ model which consists of five additional parameters that are derived from a binaural model and a subsequent clustering approach to represent the spatial properties of the signals. This setup allows to exploit the capabilities of PEAQ while simultaneously improving the performance for use cases in which the spatial scenario has a significant impact on the perceived quality. The approach is illustrated schematically in the Figure 1. As can be seen from the figure, the non-linear mapping of the PEAQ output and the spatial parameters onto an overall quality measure is realized by means of a trained NN using an extensive listening test. The newly developed measurement will be denoted as advanced ODG (AODG) in the following.

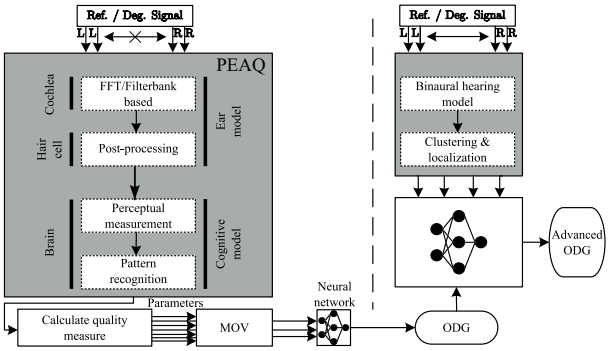

Fig. 1. System overview of the novel quality measure

2.1. Binaural model and clustering approach

A short review of the content of [7] is given here to introduce the necessary quantities for the following considerations. The binaural hearing model is based on the model of [14] and features the extensions of [15, 16] along with a novel frequency and delay weighting to calculate the correlogram $\psi(\lambda, \tau)$ for every audio signal. The correlogram is the output of the hearing model part and a representation of the temporal structure of the correlation between the two ear signals. The two dimensions of the correlogram are the frame index λ and the internal delay τ which represents the index of the interaural crosscorrelation.

Local maxima in the two-dimensional correlogram are grouped by a k-means clustering algorithm to identify clusters (i.e., sources) and a new refinement step was proposed in [7] to reduce the probability of overestimation of the number of sources P. The clusters consist of multiple points in the λ-τ-plane grouped around the centroid $\mu_i(\lambda, \tau)$ of the cluster i. Both an estimate Q_i of the spatial position of the source and a regression curve $q_i(\lambda)$ are calculated for every cluster. The regression curve is a representation of the movement of the source over the length of the signal.

2.2. Spatial quality parameters

While many parameters can be calculated based on the hearing and cognitive model [7], most of the spatial information can already be extracted by a few well-chosen parameters. Having too many parameters also increases the risk of overfitting the quality measure to the available training data which is always limited due to practical constraints. Through analysis of a preliminary listening test, five parameters could be identified which provide a good representation of the spatial properties:

- **Mean difference of correlograms** The average difference between the correlograms of the reference signal and the degraded signal is determined according to:

$$p_1 = \mathsf{E}_\lambda \left\{ \mathsf{E}_\tau \left\{ |\psi_{\mathrm{ref}}(\lambda, \tau) - \psi_{\mathrm{deg}}(\lambda, \tau)| \right\} \right\} \tag{1}$$

- **Mean difference of regression curves** This parameter is calculated as the average of the absolute values of the difference between the regression curves for sources in the reference signal and regression curves for sources in the degraded signal.

$$p_2 = \mathsf{E}_i \left\{ \mathsf{E}_\lambda \left\{ |q_{i,\mathrm{ref}}(\lambda) - q_{i,\mathrm{deg}}(\lambda)| \right\} \right\} \tag{2}$$

- **Mean difference of estimated source positions** The average of the absolute values of all the differences between the estimated spatial source positions in the reference signal and the degraded signal is used as the third spatial parameter:

$$p_3 = \mathsf{E}_i \left\{ |Q_{i,\mathrm{ref}} - Q_{i,\mathrm{deg}}| \right\} \tag{3}$$

- **Average difference of cluster centroids** The average of the absolute values of the differences between the cluster centroids of the reference signal and the degraded signal is calculated as follows:

$$p_4 = \mathsf{E}_\lambda \left\{ \mathsf{E}_\tau \left\{ |\mu_{i,\mathrm{ref}}(\lambda, \tau) - \mu_{i,\mathrm{deg}}(\lambda, \tau)| \right\} \right\} \tag{4}$$

- **Difference between the widths of the auditory events** This parameter takes the difference in the width of the sources between the reference signal and the degraded signal into account. The width B_i of a source is determined from the internal delays τ_i that belong to this source (i.e., cluster i) as follows:

$$B_i = |\max(\tau_i) - \min(\tau_i)| \tag{5}$$

The final spatial parameter is determined as the average of the changes in width of the sources:

$$p_5 = \mathsf{E}_i \left\{ B_{i,\mathrm{ref}} - B_{i,\mathrm{deg}} \right\} \tag{6}$$

From these parameters, the spatial degradation in comparison to the reference signal can be measured instrumentally. Increasing values for these parameters indicate quality degradation.

3. MAPPING PARAMETERS TO AODG

Even though the target of the proposed method is to remove listening tests from the development process, it is of great importance that instrumental measures correctly include human perception. Hence they are trained based on the results of a suitable listening test.

3.1. Design of the listening test

The main focus during the development of AODG was the overall audio quality while specifically taking degradation with respect to the spatial signal properties into account. In order to have a usable instrumental measure, it is necessary to tune the model parameters to human perception. A listening test as recommended in [1] was conducted which is described in the following.

The test that was used in this development is a degradation category rating (DCR) test. In this test type, every participant gets to hear two signals:

- a reference signal of high quality and
- a degraded signal.

It is known to the participant which signal is the reference and the degraded signal, respectively. The test material is composed of speech and music signals containing fixed as well as moving sources. Different types of degradations (e.g., various codecs or a complete removal of all spatial properties by downmixing) were used to generate a meaningful set of test items.

The rating scale for this test consists of five rating levels according to [1] which can be found in Table 3.1. Since standard PEAQ

Rating level r_{DCR}	Degradation is
5	inaudible
4	audible but not annoying
3	slightly annoying
2	annoying
1	very annoying

Table 1. Rating scale for the DCR test

utilizes a rating scale of 0 to -4, the rating levels are adjusted by $r = r_{\mathrm{DCR}} - 5$.

The listening test was conducted in a quiet studio booth with very little reverberation. The test signals were reproduced by a calibrated combination of a digital equalizer (Head Acoustics PEQ V) and a headphone (Sennheiser HD 600). In total, twenty listeners participated in the listening test that consisted of 50 test items per participant. A preliminary training phase with signals similar to the test signals was included before the test started.

3.2. Model Calibration

The crucial part in any instrumental quality measure is the mapping between parameters that are calculated from the audio signals and

the quality estimate. This mapping is realized (as in standard PEAQ) by an NN which consists of a feed-forward structure with an input layer with six neurons for the five spatial parameters p_1 to p_5 (cf. 2.2) and the output of standard PEAQ, one hidden layer consisting of ten neurons, a single output neuron and two bias units.

All neurons are connected by weighted edges so that every neuron can be characterized by its input and output edges and by its activation behaviour. This activation behaviour is modeled as a symmetrical sigmoid function on the sum of the weighted input values:

$$\text{tansig}\,(x) = \frac{2}{1+e^{-2x}} - 1 \tag{7}$$

The reason for this choice is the fact that this function can be easily differentiated which is a necessary prerequisite for the applicability of many learning rules [17].

The NN needs to be trained first, this is done in a supervised manner by the Levenberg Marquardt method which is a very efficient method for small networks that converges comparatively quickly. It is described extensively in [18]. The training process can be summarized by the following steps:

1. Initialize the weighting factors randomly.
2. Pick 34 of the 50 signals from the listening test randomly and in random order. The results of the listening test were averaged over all participants before training.
3. Let the NN calculate the output for the current weighting factors.
4. Modify the weighting factors and bias units by means of the learning algorithm in order to minimize the mean squared error between the output of the NN and the results of the listening test.
5. Control the learning process by the associated validation algorithm continuously and stop the learning process if no further gains are to be expected. This control inherently also helps to minimize the risk of overfitting.

The 16 signals that were not chosen for training are used for a later evaluation of the performance of the model.

This training regime has a disadvantage that has to be mentioned: The *best* neural network can only be determined after all $\binom{50}{34} = 4.92 \cdot 10^{12}$ possible combinations of training and evaluation signals are tested. As a reasonable compromise between quality and complexity, 2000 different sets of combinations were used to train 2000 NNs. Out of these NNs, the network with the highest correlation with the results of the listening test and the lowest mean square error was chosen.

Overfitting is an issue that can arise when training neural networks with limited amounts of training data. Due to the maximum length of a listening test that does not lead to major discomfort for the participants, the training data in this setup is limited to 50 signals of which a third can not be used for training since it is necessary for evaluation purposes. An additional criterion is introduced that is specifically tailored to this application and should help in minimizing the risk of overfitting: Already in the development of standard PEAQ, a certain confidence interval was specified to define the allowed deviation between estimated and true quality, cf. Fig. 2. The distance between quality estimates that are outside of the confidence interval and the confidence interval itself shall be minimized as well.

All input parameters are normalized to values between -1 and $+1$ for the training process. The normalization factor is determined based on more than 1300 simulations for different signals and different signal processing systems. The maximum value for every parameter was calculated along with the 90$^{\text{th}}$ percentile ($Q_{.90}$). These values are collected in Table 2.

Parameter	Maximum	$Q_{.90}$
p_1	$12,05$	$4,61$
p_2	$1,25$	$0,39$
p_3	$1,36$	$0,41$
p_4	$1,26$	$0,29$
p_5	$1,37$	$1,14$

Table 2. Maximum and 90$^{\text{th}}$ percentile of the input parameters

Every parameter is then normalized according:

$$f\,(p_i) = \begin{cases} \frac{p_i}{Q_{i,.90}}, & |p_i| < Q_{i,.90} \\ \text{sign}(p_i) \cdot Q_{i,.90}, & |p_i| \geq Q_{i,.90} \end{cases} \tag{8}$$

The clipping to the 90$^{\text{th}}$ percentile reduces the impact of outliers and leads to a more generalized model.

After determining the best NN according to the presented targets, it can then be evaluated with the remaining 16 signals that were not used for training. This evaluation is done in the next section.

4. EVALUATION OF THE PROPOSED QUALITY MEASURE

The comparison of the novel quality measure with PEAQ, the basis for the development, can be done based on different criteria:

- The correlation ρ between the quality estimate and the results of the listening test
- The mean square error RMSE of the estimation compared to the results of the listening test.
- The coefficient of determination R^2 is a measure for the ability of the model to generalize and approximate the true relationship between the input parameters and the estimated quality. Possible values for this measure are between 0 and 1, with higher values indicating a higher quality. The coefficient of determination is calculated from the results $r(i)$ of the listening test, their average $\bar{r}(i)$, and the quality estimates $\hat{r}(i)$ as
$$R^2 = 1 - \frac{\sum_{i=1}^{50} \left(r(i) - \hat{r}(i)\right)^2}{\sum_{i=1}^{50} \left(r(i) - \bar{r}(i)\right)^2} \tag{9}$$
- The number of outliers $N_{d_{out}}$ with respect to the previously defined confidence interval.

With these quality measures, a comparison of PEAQ and the new quality measure can be carried out. This comparison is done with those signals that were not used for training the NN. In the diagrams, all 50 signals are depicted to get a better impression of the performance of PEAQ in these cases but the signals that were used for training are clearly marked to also allow for a quick overview on the performance of the new quality measure in cases that were not included in the training process.

In the Fig. 2, the results when using standard PEAQ can be seen. A perfect instrumental quality measure would lead to having all individual data points on the dashed main diagonal. The positive that can be taken from this is the fact that most of the markers are

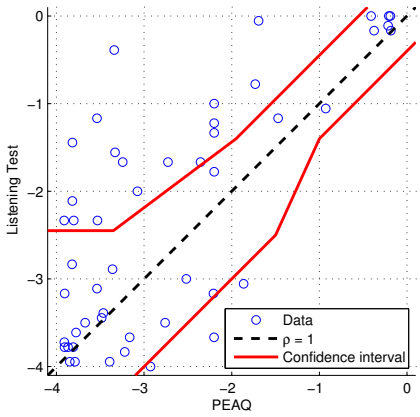

Fig. 2. Scatter plot of listening test results and ODG values according to PEAQ

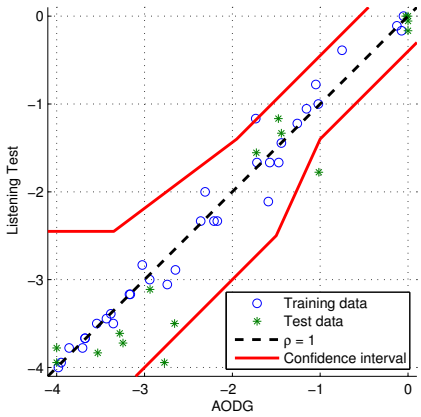

Fig. 3. Scatter plot of listening test results and AODG values according to the proposed quality measure

within the confidence interval and that there are no major outliers to the right of the confidence interval, i.e., there are no cases for which PEAQ strongly overestimates the quality of the signals. On the other hand, there are numerous cases of strong underestimation of the signal quality which can be seen as the various points in the top left part of the diagram.

The results for the proposed AODG measure that explicitly takes spatial properties into account are depicted in Fig. 3. It can clearly be seen that including the spatial parameters leads to a significantly stronger correlation between the results of the listening test and the quality estimates of the instrumental measure. The number of outliers outside of the confidence interval is very small and even these outliers are very close to the confidence intervals.

As a more formal and clearer comparison of the performance of both standard PEAQ and the proposed instrumental measure, the aforementioned criteria are calculated for both measures and collected in Tab. 3.

Criterion \ Measure	PEAQ	AODG	AODG (Test Data Only)
ρ	0.704	0.971	0.954
RMSE	1.067	0.322	0.497
R^2	0.179	0.942	0.898
$N_{d_{out}}$	40%	6%	12%

Table 3. Comparison between PEAQ and the proposed instrumental measure

All criteria illustrate the improved performance of the proposed instrumental measure compared to standard PEAQ. Especially the column of values that are calculated only for the signals that were not used for training the model is important for quantifying the performance of AODG. These values and the high coefficient of determination clearly indicate that the measure will perform accordingly for other test cases and signals.

As an example, a binaural piece of music (consisting of a piano, a violin and a trumpet playing from different directions) is transmitted by MPEG-1 layer 3 (MP3) [19] and parametric stereo [20]. Both systems are used at configurations that will not lead to a really good transmission quality, but the decisive difference between the two coding systems is that parametric stereo explicitly transmits and reconstructs spatial parameters. This difference leads to a more natural and consistent spatial impression. This example was also included in the listening test, where the quality for the transmission with the MP3 system in this configuration was identified as quite bad while the transmission with parametric stereo is still acceptable for this signal. The results for both transmission systems and both instrumental quality measures are collected in Tab. 4 along with the results of the listening test.

Measure \ Transmission system	MP3	parametric stereo
PEAQ	-3.5	-3.8
AODG	-2.9	-1.4
Listening Test	-3.1	-1.4

Table 4. Instrumental quality measures and listening test results for different transmission systems

It is obvious that while both standard PEAQ and the AODG correctly indicate that the transmission quality of the MP3 system in this configuration is not good at all, only AODG is able to correctly identify the improved subjective quality of the transmission with parametric stereo.

5. CONCLUSION

An extension to PEAQ was presented which makes use of five parameters that are derived from a recently proposed binaural hearing model. These parameters provide a compact description of the signal properties that are important for spatial perception. The extension follows the basic principle of PEAQ by calculating these parameters and then mapping them onto the quality measure by means of a neural network. The inclusion of spatial information into the instrumental quality measurement leads, in contrast to PEAQ, to a consistently high correlation between the instrumental measure and a listening test.

6. REFERENCES

[1] ITU, *Methods for subjective determination of transmission quality (ITU-T Recommendation P.800)*, International Telecommunications Union, Aug. 1996.

[2] M. Schäfer, C. Schnelling, B. Geiser, and P. Vary, "A Listening Test Environment for Subjective Assessment of Speech and Audio Signal Processing Algorithms," in *Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, 2011.

[3] 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 Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1890–1901, 2006.

[4] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality.*, ser. T-Labs Series in Telecommunication Services. Springer, 2011.

[5] S. George, S. Zielinski, and F. Rumsey, "Feature extraction for the prediction of multichannel spatial audio fidelity," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1994–2005, 2006.

[6] S. Zielinski, F. Rumsey, R. Kassier, and S. Bech, "Development and initial validation of a multichannel audio quality expert system," *Journal of the Audio Engineering Society*, vol. 53, no. 1/2, pp. 4–21, 2005.

[7] M. Schäfer, M. Bahram, and P. Vary, "Improved Binaural Model for Localization of Multiple Sources," in *Proceedings of 10. ITG Symposium Speech Communication*, 2012.

[8] ITU, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs (ITU-T Recommendation P.862)," International Telecommunication Union, 2001.

[9] ——, "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs (ITU-T Recommendation P.862.2)," International Telecommunication Union, ITU-T Rec. P.862.2, 2005.

[10] ——, "Perceptual objective listening quality assessment (ITU-T Recommendation P.863)," International Telecommunication Union, 2011.

[11] ——, *Method for Objective Measurements of Perceived Audio Quality (ITU-R Recommendation BS.1387-1)*, International Telecommunications Union, 2001.

[12] J.-M. Valin, T. B. Terriberry, C. Montgomery, and G. Maxwell, "A high-quality speech and audio codec with less than 10-ms delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 58–67, 2010.

[13] M. Schäfer and P. Vary, "Hierarchical multi-channel audio coding based on time-domain linear prediction," in *Proceedings of European Signal Processing Conference (EUSIPCO)*. EURASIP, Aug. 2012, pp. 2148–2152.

[14] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psych.*, vol. 41, pp. 35–39, 1948.

[15] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals," *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1608–1622, 1986.

[16] ——, "Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front," *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1623–1630, 1986.

[17] T. Vogl, J. Mangis, A. Rigler, W. Zink, and D. Alkon, "Accelerating the convergence of the back-propagation method," *Biological Cybernetics*, vol. 59, pp. 257–263, 1988, 10.1007/BF00332914. [Online]. Available: http://dx.doi.org/10.1007/BF00332914

[18] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *Neural Networks, IEEE Transactions on*, vol. 5, no. 6, pp. 989–993, nov 1994.

[19] K. Brandenburg and G. Stoll, "Iso-mpeg-1 audio: A generic standard for coding of high-quality digital audio," in *Audio Engineering Society Conference: Collected Papers on Digital Audio Bit-Rate Reduction*, 5 1996. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=7134

[20] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, p. 561917, 2005. [Online]. Available: http://asp.eurasipjournals.com/content/2005/9/561917