Metrics for the Evaluation of Audio Quality

Magnus Schäfer, Lars Thieling, Lukas Vollmer
HEAD acoustics GmbH, 52134 Herzogenrath, Deutschland, Email: telecom@head-acoustics.de

Abstract

The reliability and performance of an instrumental assessment method for audio systems hinges on the choice of appropriate metrics that quantify the quality of the system in a hearing-adequate manner. The presented approach for instrumental assessment of audio systems is based on binaural recordings of real music signals as well as measurement signals. The music signals were also used for auditory assessment of audio systems in earlier contributions.

Recently, research results revealed that the perceived overall quality of an audio system can be predicted to a high degree from three attributes: timbre, distortions and immersion. Besides a common preprocessing, each attribute requires specific analyses. For example, the analyses for distortions are not based on any binaural cues while these are of paramount importance for assessing immersion.

This contribution presents an overview of the assessment system along with an explicit description of example analyses with their resulting metrics that are the foundation of the instrumental quality prediction. It describes the relation between the metrics and the quality perception, and makes a comparison with auditory results.

Introduction

The perceived quality of an audio system is influenced by a multitude of technical (and non-technical) aspects. Some methods have been proposed for quantifying *audio quality*, e.g., in [1] and [2]. These approaches mainly address the degradation that is introduced by lossy coding of audio signals. Thus, the auditory basis for the models are degradation tests according to [3]. These methodologies focus on the basic audio quality without further analysis of additional attributes.

A completely separate task is the assessment of different audio systems. This comprises, e.g., the comparison of loudspeakers, amplifiers, spatial audio rendering schemes or even complex applications like car audio systems. A listening test of such systems was presented in [4] using an absolute category rating paradigm for a single attribute: overall quality. The test and its later extension in [5] also provide a comparison of different test environments leading to the conclusion that tests in a listening laboratory mostly lead to the same results as tests in a car. The only systematic differences were observed for the case of driving in a driving simulator while assessing the audio quality (and even then only for the lower end of the rating scale).

Some auditory investigations into additional attributes for assessing audio quality can be found in [6], [7] and [8]. A listening test methodology with four quality attributes was introduced in [9] that targets the comparative assessment of different audio systems.

This contribution briefly describes the listening tests from [9] that provide the underlying perceptual data for developing an instrumental model before providing a basic overview of a possible structure for an instrumental model that inherently exploits one important finding from the listening tests. Subsequently, the requirements for good metrics are discussed and examples for two quality attributes are given. These examples are described and the relation between the metrics and the listening test results is illustrated.

Auditory Assessment of Audio Quality

The assessment methodology that was introduced in [9] consists of a comparison category rating listening test (adapted from [10]). It simultaneously assesses four quality attributes: timbre, distortion, immersion and overall quality. The test inherently provides information on the consistency of the ratings that were given by the individual test subjects. An analysis of circular triads was carried out in [9] and a procedure for quantifying the consistency was proposed along with thresholds for deciding which results have to be discarded. The general result was that while some cases of inconsistent voting occurred, the test can be conducted with naïve test subjects.

A second investigation in [9] focused on the relation between the individual quality attributes and the overall quality. A simple linear regression formula was derived from the auditory results that connects the attributes and the overall judgement:

$$\text{Overall quality} = 0.5 \cdot \text{T} + 0.23 \cdot \text{D} + 0.46 \cdot \text{I} \quad (1)$$

This simple relation is capable of predicting the overall quality almost perfectly as illustrated in Figure 1. The correlation between the auditory results and the prediction is very high and there are no outliers.

Structure of an Instrumental Model

An overview of the basic structure of an instrumental model is given in Figure 2. There are two inputs to the model: Besides the music signals that were also used in the listening test (cf. [9]), sweep measurements were conducted with all tested audio systems which can be utilized by the signal analysis stages. Both inputs actually represent input groups consisting of the reference signals

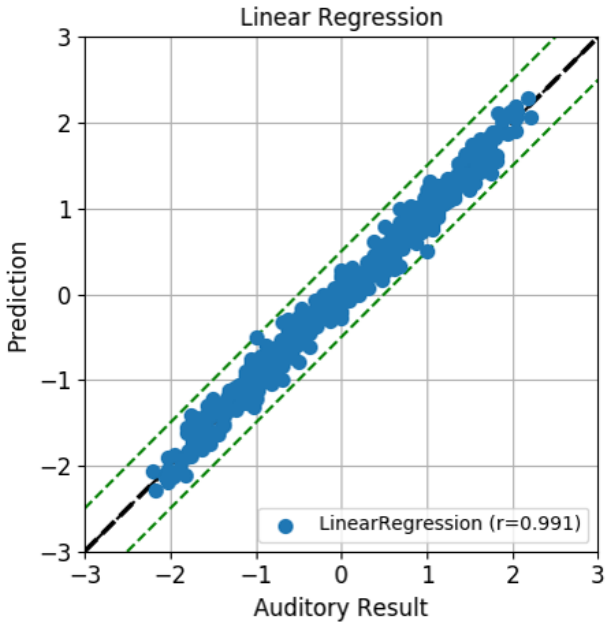

Figure 1: Comparison of auditory results for overall quality and prediction of overall quality from the other three attributes (figure from [9])

that were fed into the audio systems and the signals that were recorded binaurally.

The two input groups are fed into a primary preprocessing stage that asserts that, e.g., there is no difference in the sampling rates between the reference and the recorded signals. The subsequent analysis stages for the three individual quality attributes each consist of a group of metric calculations (examples of which are presented in later sections) and a regression stage that combines the metric results and maps them to a mean opinion score (MOS) scale for the attribute.

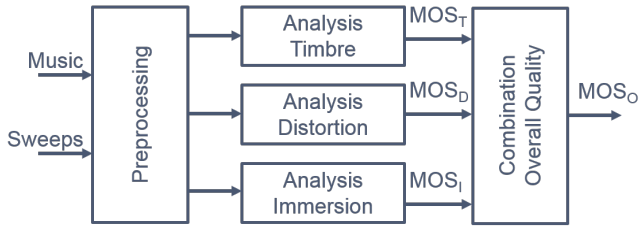

Figure 2: Block diagram of an instrumental model

Due to the results of the auditory assessment, there is no individual analysis stage for the overall quality. The overall quality score is calculated from the individual attribute scores, e.g., by Equation 1.

Analyses and Metrics

The specific analyses for the individual quality attributes are the link that connects the audio signals to the regression stage which then condenses the analysis results into a MOS score. The analyses results therein are scalar values which should ideally exhibit significant correlation to the auditory results. Every quality attribute will require multiple analyses as no single metric will be able to explain all the variance in the auditory data.

It can be expected from the sound attributes that, e.g., the spectral properties of the audio system have an impact on the rating for timbre, nonlinearities are a major influence on the rating for distortion and that the perceived source position and width are important for the rating for immersion. It is also expectable that there is some overlap between the attributes. As an example, it is known that different frequencies contribute differently to spatial perception [11]. Accordingly, an analysis of the frequency response of the audio system might be meaningful both for predicting timbre and immersion.

Possible metrics for immersion that can be derived from a binaural hearing model [12] were already tested in [13] for another set of auditory data. In the next two sections, two examples of possible analyses for timbre and distortion are described. As an additional aspect, it is investigated how the result of an analysis of sweep signals can be related to the audio signals that were used in the listening test.

Timbre Metric – Target Frequency Response

As mentioned before, the spectral properties of the audio system are presumably an important aspect when assessing timbre. The frequency response of the audio system is a possibility to quantify these properties. Assuming sufficient auditory data is available, an average target for the frequency response can be derived from the auditory results. It should be noted that there is a certain amount of variance around this average frequency response due to a spread of the personal preferences depending on various conditions like music taste or cultural background. Different targets, e.g., flat response or amplification of low frequencies, are possible for the frequency response.

The metric that shall be derived is determined from the difference $D(f)$ between the frequency response, calculated from the sweep signals, and the chosen target curve, both in third octave bands. A higher frequency resolution does not lead to better results based on the available auditory data. The difference is depicted as the blue dashed curve in the upper subplots of Figures 3 and 4. It can be seen that the audio system does match the target curve fairly well in the mid frequencies but slightly overemphasizes the high frequencies and is clearly lacking in low frequencies.

$D(f)$ does not contain any relation to the individual music signals that were assessed by the test subjects in the listening test. As the listening test was designed to contain very different types of music signals, the properties of the signals are very different as well. This leads to the observation that some differences to the target curve for the frequency response are more relevant for some music signals than for others. Two examples are shown in the lower subplots of Figures 3 and 4. A spectrum of a signal with a lot of energy in the low frequencies is shown

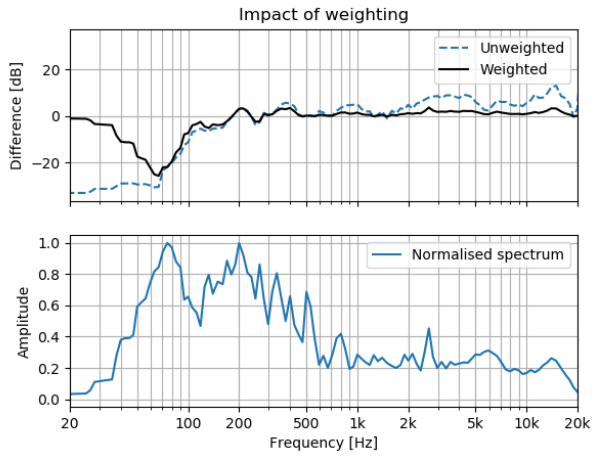

Figure 3: Weighted difference to target frequency response – signal with strong low frequency components

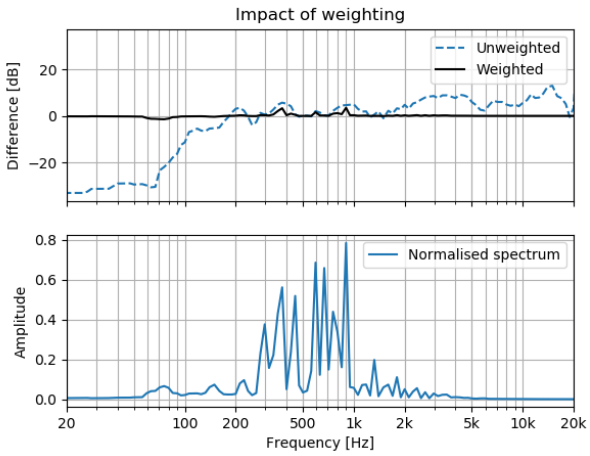

Figure 4: Weighted difference to target frequency response – signal with strong mid frequency components

in Figure 3. For this signal, the overemphasized high frequencies are not important at all – there is barely any signal energy in that frequency region anyway. The large difference for the low frequencies, however, will have a strong impact on the perception of this signal. Another signal with most energy in the mid frequencies is shown in Figure 4. This signal will be reproduced very well by the given audio system (at least from the spectral properties) as there is almost no signal energy in the regions where the frequency response of the audio system deviates from the target curve.

The relation between $D(f)$ and the music signals can be made by weighting $D(f)$ with the normalised signal spectrum $\bar{S}(f)$ according to

$$D_W(f) = D(f) \cdot \bar{S}(f). \tag{2}$$

The normalization of the signal spectrum is done by dividing the spectrum by the largest value of the spectrum individually per channel of the reference signal and then averaging over the channels. The weighted differences $D_W(f)$ are shown as the solid black lines in the upper subplots of Figures 3 and 4. It can be seen that the weighting leads to the desired behaviour: Even though the audio system is identical, the final weighted difference curves are very different due to the spectral content of the two music signals.

Calculating, e.g., the arithmetic mean of the absolute value of $D_W(f)$ provides a scalar value that can be used as an input to the regression stage for the prediction of the quality with respect to timbre. It should be noted that there are different possibilities to condense the information in $D_W(f)$ into a single value, e.g., using the geometric mean to emphasize the impact of larger deviations or considering positive and negative values in $D_W(f)$ differently.

Distortion Metric – Analysis of Harmonics

An established value that is commonly determined in loudspeaker measurements is the total harmonic distortion (THD), the ratio of powers of all harmonics to the power of the fundamental frequency:

$$\mathrm{THD} = \frac{P_{2\ldots\infty}}{P_1} \tag{3}$$

A related metric was derived from an analysis of the listening test results which showed that there is a certain group of harmonics that have the highest impact on the perception of distortion. The partial harmonic distortion can be calculated by relating the power of a subset of the harmonics to the power of the fundamental frequency:

$$\mathrm{PHD} = \frac{P_{4\ldots11}}{P_1} \tag{4}$$

Results

The relation between the values of the two metrics and the corresponding auditory results can be illustrated by scatter plots of the results for each audio system (i.e., averaged over the six music signals that were used). The plot for the arithmetic mean of the absolute value of the weighted difference to the target frequency response and the auditory results for timbre is shown in Figure 5.

It can be observed that there is a clear correlation between metric and auditory result. The correlation coefficients are Pearson's $r = -0.81$ and Spearman's $\rho = -0.82$. The negative sign of the coefficients makes sense: Less deviation from the target curve is better.

Figure 6 presents the relation between the partial harmonic distortion and the auditory results for distortion. The plot exhibits some similarities to Figure 5: Again, there is a clear correlation between the metric and the auditory results. The correlation coefficients are slightly lower at $r = -0.77$ and $\rho = -0.79$. Less distortion is better, hence the negative signs. One important difference concerning the auditory data is the smaller range of values. This should be addressed in the choice of signals for future listening tests to test if the proposed metric also quantifies larger amounts of perceived distortion correctly.

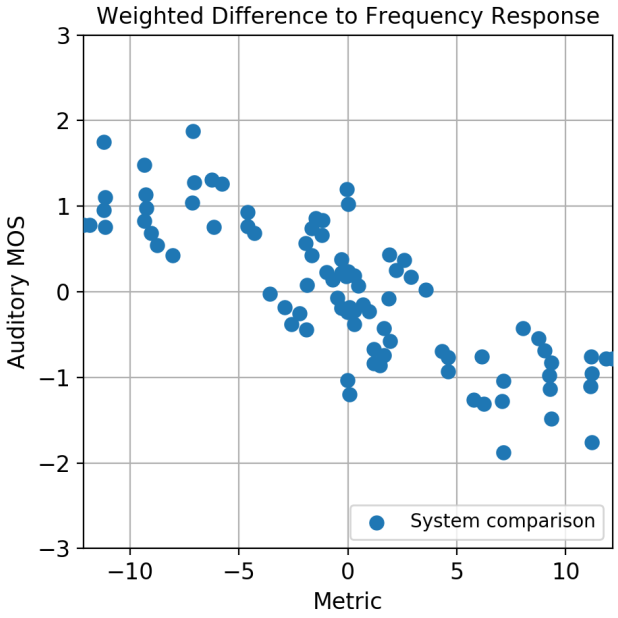

Figure 5: Relation between the timbre metric and the auditory results

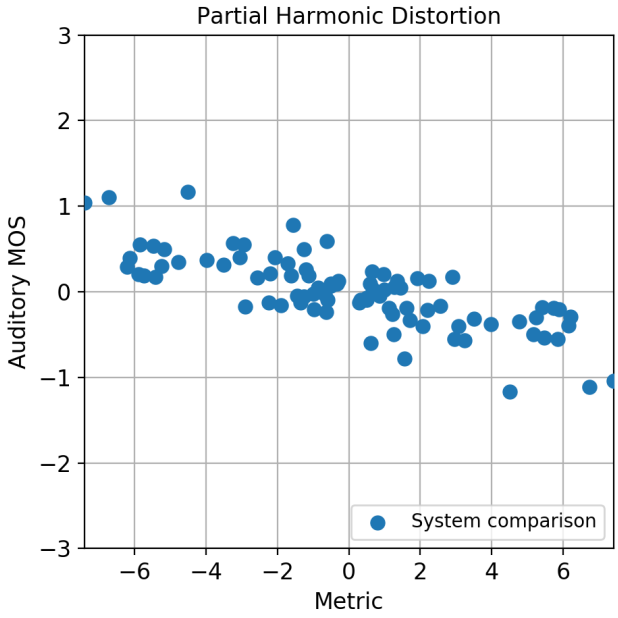

Figure 6: Relation between the distortion metric and the auditory results

Conclusions

An overview of the assessment of audio quality was presented with a focus on the specific analysis requirements for an instrumental assessment and how they can be derived from the auditory results. A structure for the instrumental assessment was described that inherently utilizes the outcome of the listening tests that the overall quality can be predicted from the individual quality attributes to a high degree. Two example metrics for quantifying timbre and distortion were described and their correlation to the auditory results illustrated. Both metrics show clear correlation to the auditory results and can be a valuable component for modeling the complex perception process of audio quality.

References

[1] Rainer Huber and Birger Kollmeier. PEMO-Q – A new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on audio, speech, and language processing*, 14(6):1902–1911, 2006.

[2] ITU-R Recommendation BS.1387-1. *Method for objective measurements of perceived audio quality*, November 2001.

[3] ITU-R Recommendation BS.1116-3. *Methods for the subjective assessment of small impairments in audio systems*, February 2015.

[4] Jan Reimes, André Fiebig, Thomas Deutsch, and Michael Oehler. Comparison of Auditory Testing Environments for Car Audio Systems. In *Fortschritte der Akustik - DAGA 2017*. Berlin, 2017.

[5] Magnus Schäfer, Jan Holub, Jan Reimes, and Tomáš Drábek. Subjective testing of car audio systems with and without parallel task. In *Fortschritte der Akustik - DAGA 2018*. DEGA e.V., Berlin, 2018.

[6] Paulo Marins, Francis Rumsey, and Slawomir Zielinski. Unravelling the relationship between basic audio quality and fidelity attributes in low bit-rate multichannel audio codecs. In *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[7] Catherine Colomes, Sarah Le Bagousse, and Mathieu Paquier. Families of sound attributes for assessment of spatial audio. In *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[8] Sarah Le Bagousse, Mathieu Paquier, and Catherine Colomes. Assessment of spatial audio quality based on sound attributes. In *Acoustics 2012*, 2012.

[9] Magnus Schäfer. Auditory assessment of multichannel audio systems. In *Speech Communication; 13th ITG-Symposium*, pages 1–5, October 2018.

[10] ITU-T Recommendation P.800. *Methods for subjective determination of transmission quality*, August 1996.

[11] Johannes Raatgever. *On the binaural processing of stimuli with different interaural phase relations*. PhD thesis, Technische Hogeschool Delft, The Netherlands, 1980.

[12] Magnus Schäfer, Mohammad Bahram, and Peter Vary. Improved Binaural Model for Localization of Multiple Sources. In *10. ITG Symposium on Speech Communication*, Braunschweig, Germany, September 2012.

[13] Magnus Schäfer. An approach for instrumental quality evaluation of car audio systems. In *Fortschritte der Akustik - DAGA 2017*. DEGA e.V., Berlin, 2017.