PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

A system for instrumental evaluation of audio quality

Magnus Schäfer

HEAD acoustics GmbH, Germany, telecom@head-acoustics.de

Abstract

An approach for instrumental assessment of audio systems is presented in this contribution, which is based on binaural recordings of real music signals as well as measurement signals. Any system for instrumental assessment of audio quality should ideally be able to replicate the perception of a human test subject. An auditory test design based on music signals was recently devised that was shown to lead to reliable test results even with naïve test subjects. The test utilizes three individual attributes (timbre, distortions and spaciousness) along with a judgement on overall quality to quantify the perceived quality. Several auditory tests were conducted with the proposed design to collect training material for the system that is presented in this contribution. The system consists of two main components: An analysis stage that contains specific components for the individual attributes and a trained regression that utilizes the results of the conducted listening tests to establish a relation between the analyses and human perception. This contribution presents an overview of the assessment system, highlights the interaction between analyses and quality perception, and makes a comparison with auditory results.

Keywords: Audio Quality, Instrumental Assessment

1 INTRODUCTION

Instrumental assessment of audio quality is a very challenging task due to the different components (technical and non-technical) which influence the quality perception. The known instrumental approaches (e.g., [3] and [5]) are based on degradation tests according to [4]. They focus on the degradation that is introduced by lossy coding of audio signals and quantify only the basic audio quality without further analysis of additional attributes.

A completely separate task is the assessment of different audio systems including the acoustic setup. This comprises, e.g., the comparison of loudspeakers, amplifiers, spatial audio rendering schemes or even complex applications like car audio systems. Such systems were auditorily assessed in a listening test (presented in [11] and extended in [15]) for a single attribute: overall quality. The test results in different test environments were compared – the outcome being that tests in a listening laboratory are mostly equivalent to tests in a car.

Some auditory investigations into additional attributes for assessing audio quality can be found in [2], [8] and [9]. A listening test methodology for the comparative assessment of different audio systems using four quality attributes was introduced in [13].

This contribution briefly describes the listening tests, which provide the underlying perceptual data for developing an instrumental model before providing an overview of a structure for an initial instrumental model and outlining the analyses for the different quality attributes. The similarities and differences between the chosen analyses are described and example prediction results are given.

2 AUDITORY ASSESSMENT

A comparison category rating (CCR) listening test (adapted from [7]) was already proposed in [13] for the auditory assessment of multichannel audio systems. The four quality attributes in the test are timbre, distortion, immersion and overall quality. The reliability of the ratings that were given by the individual test subjects can be quantified by performing an analysis of circular triads in the result data. [13] also describes a procedure for this analysis along with thresholds for deciding which results can be used in the subsequent analysis and modeling stages. The general outcome of the investigation was that it is (in general) possible to conduct this

type of listening test with naïve test subjects. Only a few cases of inconsistent voting occurred and these can be identified (and removed) by the described analysis approach.

Additionally, the relation between the individual quality attributes and the overall quality was investigated in [13] as well. The connection between the attributes and the overall judgement was described by a simple linear regression:

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

The auditory results for the overall quality and the results according to Equation 1 are depicted in Figure 1. It can be observed that the correlation between the auditory results and the prediction is very high and there are no outliers.

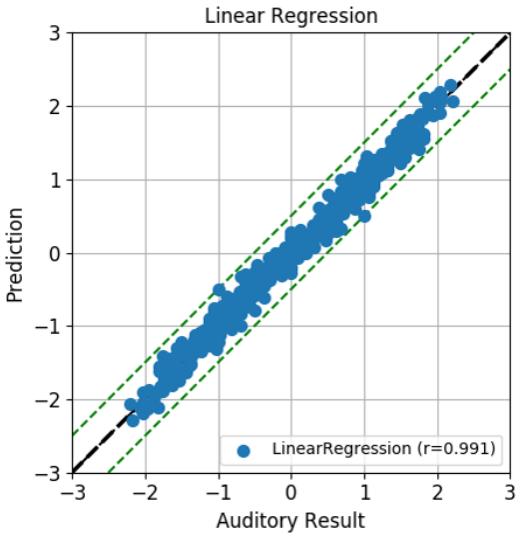


Figure 1. Comparison of auditory results for overall quality and prediction of overall quality from the other three attributes (figure from [13])

3 STRUCTURE OF THE PREDICTION SYSTEM

A very brief overview of the structure for an instrumental model was given in [16]. The structure is shown in Figure 2. The preprocessing stage ensures that the signals are fed into the analysis stages in a defined manner, e.g., regarding the sampling rate.

In the following, the analyis stages are described in more detail. It is shown that there are some fundamental classes of analyses which could be utilized and arguments against one of the classes are given.

Note that this contribution only considers the prediction of the individual quality attributes and there is no separate analysis stage for the overall quality: If the quality attributes are predicted well, Equation 1 will yield a good prediction of overall quality.

4 ANALYSES STAGES

The analysis stages aim to quantify the perceived quality of the music signals $m_{a,n}(k)$ (with a denoting the audio system and k as the discrete time index) that were also used in the listening test, i.e, the stages have to provide R different analysis results $r_{1\ldots R,,(T|D|I)}(a,n)$ for each music piece n (the subscript indicating the

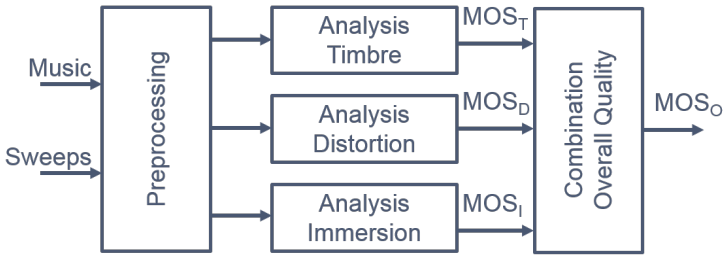


Figure 2. Simple block diagram of an instrumental model (figure from [16])

attribute that shall be quantified: *T*imbre, *D*istortion and *I*mmersion). As an additional information to these music signals, sweep recordings $s_a(k)$ for the audio system a must be provided as well for certain analyses as will be described in the following. The listening test uses six music signals for each audio system (cf. [13]), but there is only one sweep recording for each audio system.

There are three fundamental possibilities how these signals could be used to quantify the behaviour of audio system a for the music piece $m_{a,n}(k)$:

- Analysing only the music signal $m_{a,n}(k)$,
- Analysing only the sweep signal $s_a(k)$,
- Analysing both signals ($m_{a,n}(k)$ and $s_a(k)$).

Depending on the specific analysis, there are cases where analysing the music signal alone or analysing both signals makes more sense. Analysing the sweep signal alone, though, is practically never useful for two reasons.

First and foremost, the analysis result would be independent from the actual piece of music that was played. This is in clear contrast to the everyday experience that an audio system can not play all types of music (or all songs) similarly well. For example, a very small audio system is usually not capable of reproducing low frequencies. While this is very perceivable (and detrimental) for music pieces that contain a lot of bass, it is only a minor issue for other pieces that contain, e.g., vocal performances.

Secondly, this is also an issue from an algorithmic point of view: The mapping between the analysis results $r_{1\ldots R,(T|D|I)}(a,n)$ and the instrumental quality values $\mathrm{MOS}_{(T|D|I)}$ is done by a trained regression approach (e.g., a neural network or a random forest). The critical aspect here is the training of the regression which connects the analysis results with the auditory results: Analyses that only consider the sweep recording lead to a situation where (possible vastly) different auditory results (one for each music signal $m_{a,n}(k)$) have to be derived from identical analysis results (for $s_a(k)$).

In the following, five analyses for each quality attribute are briefly described. The results of these analyses are combined to get a quality prediction for each attribute.

4.1 TIMBRE ANALYSES

A first example for an analysis that is useful for predicting the quality attribute was already given in [16]: A comparison of the frequency response $\mathrm{FR}_a(\mu)$ of the audio system with a target frequency response $\mathrm{FR}_{\mathrm{target}}(\mu)$ with μ as the discrete frequency index. As the frequency response is a property of the audio system that does not consider the actual music signal in any way, a weighting was already introduced in [16] that uses a normalized signal spectrum $M_n(\mu)$. This allows to get the first analysis result for timbre by calculating the mean absolute value (across the N_μ frequency bins) of the weighted difference according to:

$$r_{1,T}(a,n) = \frac{1}{N_\mu} \sum_{\mu} \left| M_n(\mu) \cdot \left(\mathrm{FR}_a(\mu) - \mathrm{FR}_{\mathrm{target}}(\mu) \right) \right| \tag{2}$$

It was observed when analysing the auditory results for the other quality attributes, that this analysis is reasonably correlated with the other two attributes as well. Accordingly, it is used in all analysis blocks:

$$r_{1,D}(a,n) = r_{1,T}(a,n) \quad (3)$$

$$r_{1,I}(a,n) = r_{1,T}(a,n) \quad (4)$$

Three other analyses for timbre also focus on the transmission characteristics of the audio system by comparing the reference signals $x_n(k)$ that were fed into the audio systems with the recorded signals $m_{a,n}(k)$ in different spectral representations.

Two analyses calculate percentiles of differences of averaged blockwise Fourier transforms (denoted $X_n(\mu)$ and $M_{a,n}(\mu)$ in dB, respectively) in specific frequency ranges that were identified as particularly important for the judgement of the test subjects. The Fourier transform uses a block length of 4096 samples with a Hann window and 50 % overlap. The two frequency bands are denoted as *bass* from 60 Hz to 250 Hz and *upper midrange* from 2000 Hz to 4000 Hz. The two analysis are formulated with $\mathscr{P}_\xi$ as the ξ-th percentile across frequency as

$$r_{2,T}(a,n) = \mathscr{P}_5\left(X_{n,\text{bass}}(\mu) - M_{a,n,\text{bass}}(\mu)\right) \quad (5)$$

$$r_{3,T}(a,n) = \mathscr{P}_{90}\left(\left|X_{n,\text{upper midrange}}(\mu) - M_{a,n,\text{upper midrange}}(\mu)\right|\right) \quad (6)$$

The hearing model from [17] is used to calculate hearing model spectra (denoted $\mathscr{X}_n(\mu_\text{hm})$ and $\mathscr{M}_{a,n}(\mu_\text{hm})$, respectively) and the mean absolute difference between the two spectra is used as the fourth metric for timbre.

$$r_{4,T}(a,n) = \frac{1}{N_{\mu_\text{hm}}} \sum_{\mu_\text{hm}} \left|\mathscr{X}_n(\mu_\text{hm}) - \mathscr{M}_{a,n}(\mu_\text{hm})\right| \quad (7)$$

Finally, one analysis, the spectral flux according to [1], is used as $r_{5,T}(a,n)$ to roughly quantify the temporal behaviour of the audio system.

4.2 DISTORTION ANALYSES

In addition to the aforementioned frequency response metric $r_{1,D}(a,n)$, the analyses for distortion mostly aim at quantifying non-linearities. One example metric was also already introduced in [16]. It is closely related to the total harmonic distortion but uses only the power of a select few harmonics in relation to the power of the fundamental:

$$r_{2,D}(a,n) = \frac{P_{4\ldots11}(a,n)}{P_1(a,n)} \quad (8)$$

The next two analyses are based on an analysis of the spectrogram of the sweep measurement with an image processing approach. A detailed description of the methodology is beyond the scope of this overview. The analyses quantify the amplitude and the spread of the harmonics separately for low ($r_{3,D}(a,n)$) and high ($r_{4,D}(a,n)$) frequencies.

In a similar vein to the timbral analyses, there is also one analysis for distortion that mainly considers temporal characteristics. Modulation spectra $X_{\text{mod},n}(\lambda,\mu_\text{mod})$ and $M_{\text{mod},a,n}(\lambda,\mu_\text{mod})$ (with λ as the index of the analysis frame and μ_mod as the modulation frequency) are calculated and the mean value of the difference between the spectra across the N_λ analysis frames and the N_{μ_mod} modulation frequencies is used according to

$$r_{5,D}(a,n) = \frac{1}{N_\lambda \cdot N_{\mu_\text{mod}}} \sum_\lambda \sum_{\mu_\text{mod}} X_{\text{mod},n}(\lambda,\mu_\text{mod}) - M_{\text{mod},a,n}(\lambda,\mu_\text{mod}) \quad (9)$$

4.3 IMMERSION ANALYSES

It is known that there is a fairly strong relation between the spectral structure of a signal and the spatial perception with, e.g., different frequency bands contributing differently to localisation cues (cf. [10]). Accordingly, it is not surprising that there is a strong overlap between the analyses for timbre and immersion.

In fact, there are two analyses that are identical and one that is very similar. As denoted before in Equation 4, the frequency response analysis is used here as well. The same holds for the spectral flux according to [1].

$$r_{2,I}(a,n) = r_{5,T}(a,n) \tag{10}$$

The third analysis is a variation of Equation 6 – the only difference is that the difference between the averaged Fourier transforms is used instead of the *absolute* value of the difference.

$$r_{3,I}(a,n) = \mathscr{P}_{90}\left(X_{n,\text{upper midrange}}(\mu) - M_{a,n,\text{upper midrange}}(\mu)\right) \tag{11}$$

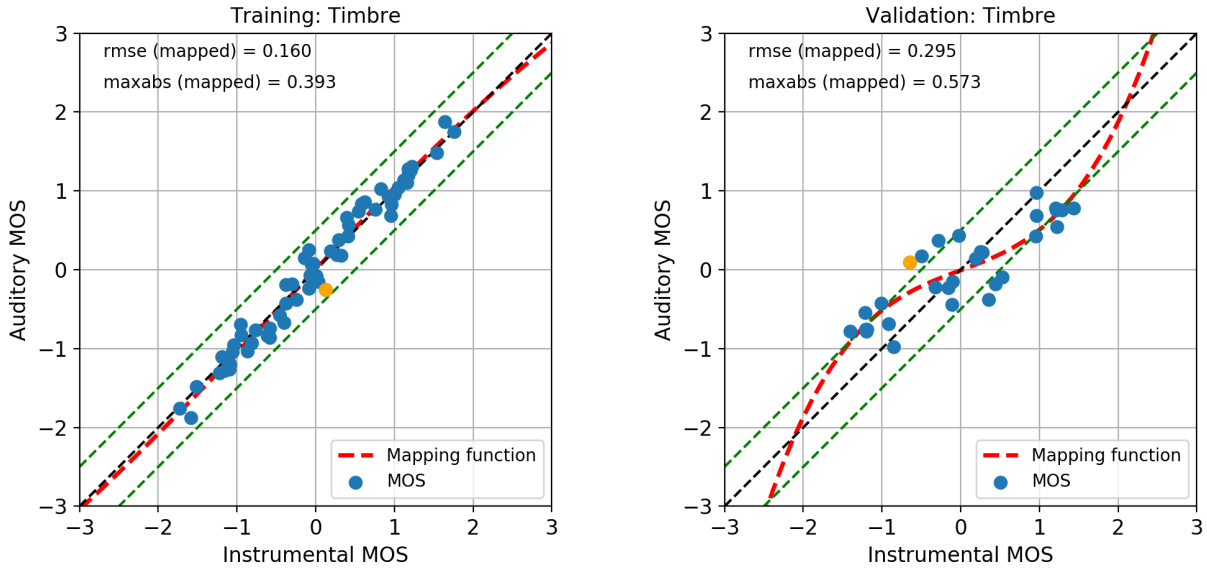
The final two analyses are very specific for immersion. They are based on a binaural hearing model [14] that provides information about the spatial distribution of sound sources in the signals in the form of correlograms $X_{\text{corr},n}(\lambda,\tau)$ and $M_{\text{corr},a,n}(\lambda,\tau)$ with τ as one of the N_τ considered lateralizations. Differences between the correlograms are then used to quantify changes in the spatial representation (cf. [12]). The percentile is determined across the entire correlogram difference, i.e., across all frames and all lateralizations.

$$r_{4,I}(a,n) = \mathscr{P}_5\left(X_{\text{corr},n}(\lambda,\tau) - M_{\text{corr},a,n}(\lambda,\tau)\right) \tag{12}$$

$$r_{5,I}(a,n) = \frac{1}{N_\lambda \cdot N_\tau}\sum_\lambda\sum_\tau\left(X_{\text{corr},n}(\lambda,\tau) - M_{\text{corr},a,n}(\lambda,\tau)\right) \tag{13}$$

5 PREDICTION RESULTS

Since the listening tests compare two signals, the differences between the analyis results for two signals are used as the input for a random forest regression (50 trees, at least 3 samples per leaf). Three different auditory tests according to the procedure described in [13] were carried out – two of these are used for training the random forests, the remaining test is used as a validation data set.



(a) Training data

(b) Validation data

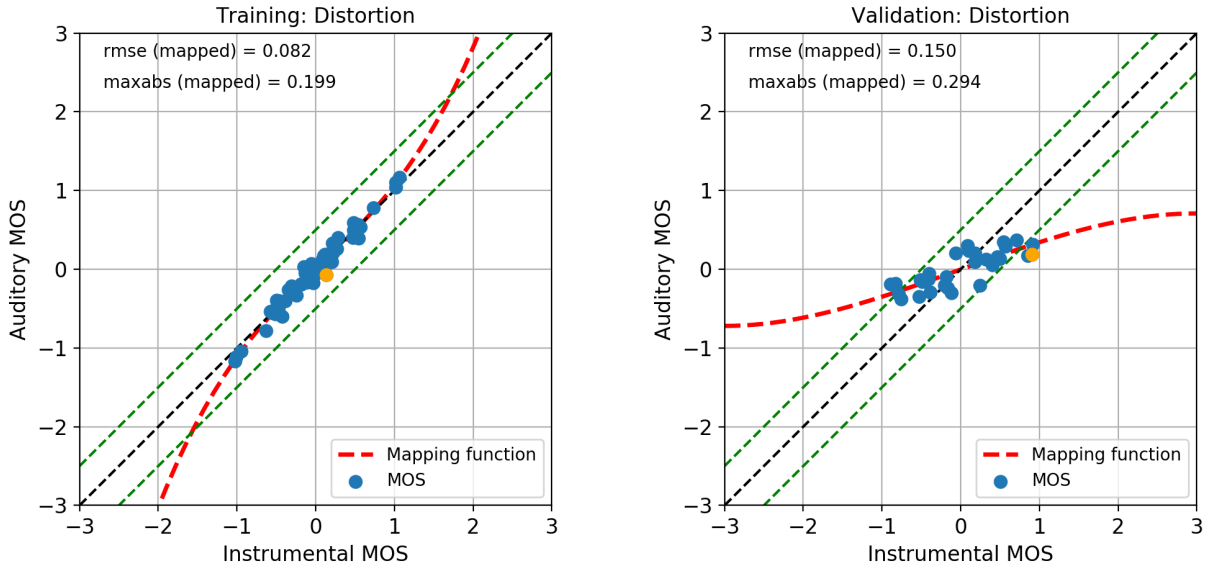
Figure 3. Prediction results for timbre

All result figures are scatter plots of the auditory mean opinion score (MOS) against the instrumental MOS. The left subfigures contain the results for the training data while the right subfigures give the results for the

validation data. The results are averaged across the different songs that were used. Thus, the blue points always represent comparisons between two audio systems. For a perfect prediction, instrumental and auditory results would be identical and all points would be on the main diagonal. In particular for the validation data, this is not realistic as, e.g., the context of the listening test was different. To compensate for this, a monotonic third order mapping function (cf. [6]) is determined that shifts the instrumental results towards the main diagonal. This mapping function is depicted as the dashed red line in the figures and root mean square error (rmse) as well as the largest absolute difference (maxabs) between the auditory results and the mapped instrumental results is given. These two values quantify both the average and the worst case performance of the instrumental assessment.

The results for the quality attribute timbre are shown in Figure 3. Looking at the results for the training data, it can be observed that the utilized regression algorithm is capable of reproducing the auditory results from the results of the chosen analyses. There is a small residual error due to the parameterisation of the random forest – having more than one sample per leaf effectively removes the capability of the regressor to perfectly memorize the training data.

The validation data set can be predicted satisfactorily, the instrumental assessment models the auditory results well. The error is unsurprisingly higher than for the training data. There is also a visible tilt in the data points: The instrumental results cover a range of approximately 3 points on the rating scale while the auditory results only cover a range of slightly more than 2 points. The instrumental assessment slightly overestimates the perceptual differences in this data set.



(a) Training data

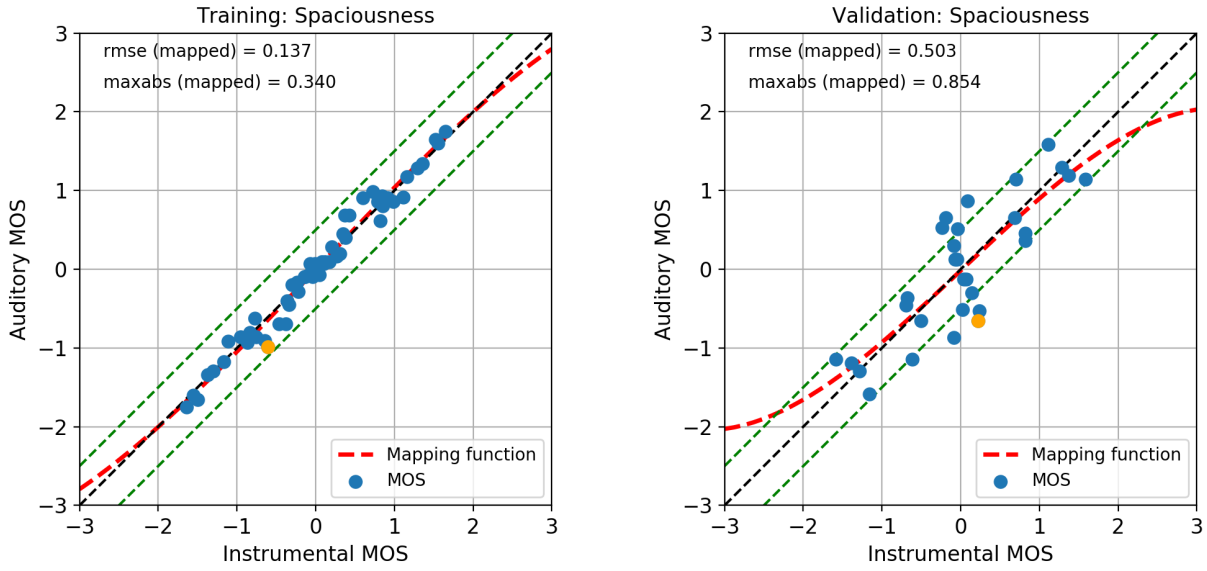
(b) Validation data

Figure 4. Prediction results for distortion

The results for the distortion attribute are depicted in Figure 4. As before, the training data is reproduced very well, both the average and the maximum error are even smaller than for timbre. At least part of the reason for this better performance is the limited range of values that is present in the auditory data for this attribute: There are barely any data points beyond ± 1.

This is even more pronounced for the validation data: Only a very limited range of ratings was observed for these audio systems. Nevertheless, the instrumental assessment is capable of distinguishing between the audio

systems quite well. Again, the differences are overestimated compared to the auditory results leading to a shallow slope of the mapping curve.



(a) Training data

(b) Validation data

Figure 5. Prediction results for immersion

Lastly, the results for immersion are given in Figure 5. The results for the training data are very similar to the results for timbre. The average and maximum error are slightly better than timbre but worse than distortion. The range of values in the auditory results is ± 2.

The performance for the validation data is quite different from the other to attributes. There is no overestimation, the mapping curve is very close to the main diagonal in the range where any results are located (i.e., ± 1.7). The average and maximum error are clearly the highest, though. A particular issue is the vertical group of data points around an instrumental result of 0. The instrumental assessment is not capable of distinguishing between these systems but the test subjects perceived a quality difference.

In general, the assessment system is capable of approximately predicting the auditory results for all three quality attributes well. There are some cases of overestimations for timbre and distortion and a slightly inferior prediction performance for immersion. The reason for these is certainly the limited amount of auditory data that was available for training the regression stages.

6 CONCLUSIONS

An overview of an instrumental assessment system for audio quality was presented and the necessary analyses were described. Similarities and differences between the analyses for the three quality attributes timbre, distortion and immersion were highlighted. The system utilizes five analyses for each attribute.

Example results of the assessment system confirm that the chosen analyses provide meaningful information concerning the perceived quality of the audio systems. The performance of the system for the training and the validation data was discussed separately for the quality attributes showing that while the performance in general is very satisfactory, there are some differences in detail that require further investigation and additional auditory tests.

REFERENCES

[1] V. Alluri and P. Toiviainen. Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception: An Interdisciplinary Journal*, 27(3):223–242, 2010.

[2] C. Colomes, S. Le Bagousse, and M. Paquier. Families of sound attributes for assessment of spatial audio. In *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[3] R. Huber and B. 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.

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

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

[6] ITU-T Recommendation P.1401. *Statistical analysis, evaluation and reporting guidelines of quality measurements*, Jul. 2012.

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

[8] S. Le Bagousse, M. Paquier, and C. Colomes. Assessment of spatial audio quality based on sound attributes. In *Acoustics 2012*, 2012.

[9] P. Marins, F. Rumsey, and S. Zielinski. Unravelling the relationship between basic audio quality and fidelity attributes in low bit-rate multi-channel audio codecs. In *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

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

[11] J. Reimes, A. Fiebig, T. Deutsch, and M. Oehler. Comparison of Auditory Testing Environments for Car Audio Systems. In *Fortschritte der Akustik - DAGA 2017*. DEGA e.V., Berlin, 2017.

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

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

[14] M. Schäfer, M. Bahram, and P. Vary. Improved Binaural Model for Localization of Multiple Sources. In *10. ITG Symposium on Speech Communication*, Braunschweig, Germany, Sept. 2012.

[15] M. Schäfer, J. Holub, J. Reimes, and T. Drábek. Subjective testing of car audio systems with and without parallel task. In *Fortschritte der Akustik - DAGA 2018*. DEGA e.V., Berlin, 2018.

[16] M. Schäfer, L. Thieling, and L. Vollmer. Metrics for the evaluation of audio quality. In *Fortschritte der Akustik - DAGA 2019*. DEGA e.V., Berlin, 2019.

[17] R. Sottek. A hearing model approach to time-varying loudness. *Acta Acustica united with Acustica*, 102(4):725–744, Jul / Aug 2016.