HEAD acoustics

APPLICATION NOTE

About this document

Content

This document is an application note on how to create sound quality metrics. It describes what such a metric is, how to use it and what advantages it has. The document also provides guidance on what to consider when creating sound quality metrics.

Target group

The following text is addressed to acoustic engineers working on automated product noise evaluation, especially to (potential) users of ArtemiS SUITE who want to use the Metrics Project.

Questions?

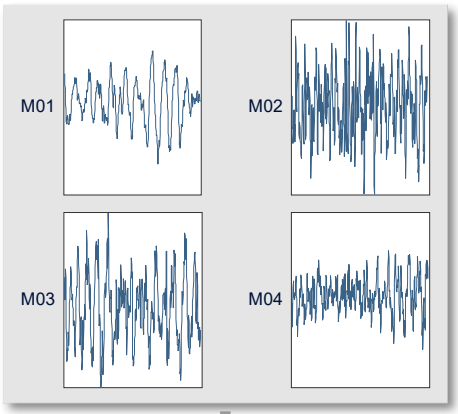
Do you have any questions? Your feedback is appreciated!
For questions on the content of this document: Imke.Hauswirth@head-acoustics.com
For technical questions on our products: SVP-Support@head-acoustics.com

Creating metrics – objective and procedure

1. Introduction

Meaningful sound evaluation

Human perception of sound is very complex and cannot be represented by a single technical measurement parameter such as sound pressure level. If you want to define a more meaningful quality index for your sounds, you are well advised to use a calculation rule that links various parameters instead. Such calculation rules, also called metrics, use the results of different technical analyses, for example, and thus determine a characteristic single value for your product sounds. Linking relevant analyses allows different sound aspects to be taken into account and incorporated into the final result. This allows, for example, not only the noise level to be included in the evaluation but also proportions of high frequencies, expression of tonal components as well as contributions of other disturbing noise patterns.



Sounds	M01	M02	M03	M04
Sound Quality Metric	3,8	8,9	7,9	5,4

Using a sound quality metric for numerical evaluation of the sound quality

Advantages of a sound quality metric

A good sound quality metric helps you to

- evaluate the acoustic quality of your products reliably and in a time-saving manner,
- identify strengths and weaknesses of your products, and
- reliably derive target sounds.

Metrics can be developed based on the results of a jury test, for instance. In this process, the results of the jury tests are mapped by measurement-based analysis results. A metric ascertained in this way will subsequently allow you to determine the perceived sound quality of your products in a time-saving manner without the need for further jury tests.

HEAD acoustics products for developing metrics

The development process of a sound quality metric involves several steps, for which HEAD acoustics provides you with various tools:

- **Binaural recording:** Use the HMS binaural artificial head measurement system to record your sounds, for example. The artificial head measurement system correctly reproduces all acoustically relevant components of the human outer ear and documents the sound situation holistically.

Binaural recording with the HMS artificial head measurement system

- **Aurally-accurate playback:** Using a labP2 playback system, you can play back a binaural recording in an aurally-accurate manner. This enables a valid perceptual evaluation of your product sounds.
- **Perceptual evaluation:** The SQala jury testing module allows you to design and conduct a jury test in just a few steps. At the end, you will receive a clear summary of the noise ratings.

Jury test with SQala

- **Technical measurement analyses:** The analysis software ArtemiS SUITE provides you with a wide range of analyses. In addition to well-known standardized methods, such as Level calculation, Octave analysis and the determination of psychoacoustic parameters, special analysis methods, such as Relative Approach or Tonality (hearing model) are available.

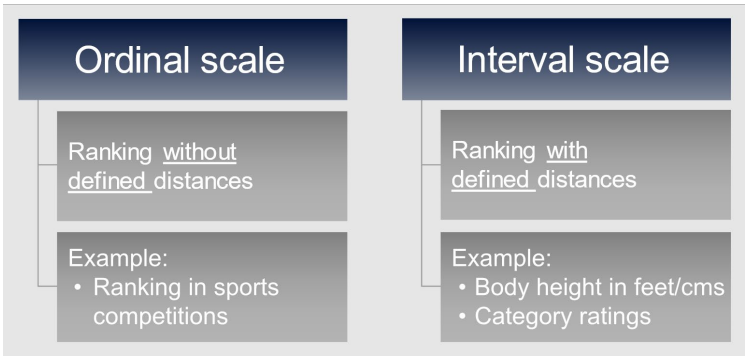
- **Defining metrics:** The Metrics Project of ArtemiS SUITE determines the correlation between perceptual judgments, e.g., from a jury test, and the single values of analysis results. Linking several differently weighted single values from different technical analyses in a linear regression model results in a calculation rule that can subsequently be used for a numerical evaluation of your sounds.

2. Suitable data for creating metrics

Before creating a metric, it must be checked whether the input data are actually suitable for metric determination.

Ordinal scale ↔ Interval scale

Jury test results to be used for metric creation should be interval-scaled. Interval-scaled results can for example be obtained from a jury test with categorical evaluation. Results from a ranking test or paired-comparison are usually ordinal-scaled and are not readily usable. This is because these tests only ask for information about the ranking order but not about the perceptual distance between the individual sounds. Thus, it is unknown whether the distance between the first and second rank is equal to the distance between the second and third rank. If ordinal-scaled results are simply translated into numerical values, this suggests an equidistant distribution that may not correspond to reality. Yet, if the numerical values misrepresent the jury test results, the basis of the correlation analysis is incorrect. Although statistical tools exist to convert ordinal-scaled data into interval-scaled data (e.g., Bradley-Terry-Luce (BTL) models), these are to be calculated with care and in many cases are not recommended to be used.



Differences ordinal scale ↔ interval scale

In order to obtain a robust metric that can actually replace performing jury tests, please consider the following guidance:

Appropriate experimental design

- If metric creation is based on jury test results, the jury tests must be designed with an appropriate understanding of noise perception. If results from an inappropriately designed jury test are used, the calculated sound quality metric will be based on a poor foundation and will not provide valid results. Only if the basis, i.e., the results of the jury test, are actually meaningful and adequately represent the perception of the sounds, can the resulting metric provide reasonable predictions for further sounds. Example: You conduct a jury test and ask the participants to evaluate the quality of sounds made by seat adjustment motors. The sounds used in the jury test were acquired using different recording systems in different recording situations so that the sounds differ not only from

motor to motor, but also due to the recording equipment used and the environment chosen. This leads the participants to not only include the actual motor sound quality in their evaluations but also the recording quality. Thus, the jury test results will not reflect the actual subject of the study. A metric created on the basis of these jury test results can therefore not provide a convincing prediction for further seat adjustment motors.

Meaningful static evaluation

- When evaluating jury test results based on statistics, care must be taken to ensure that valuable evidence is not simply "averaged away". Example: You have conducted a jury test and, despite all possible care taken in formulating the task, the participants have difficulties in evaluating certain sounds. This will result in one group of participants rating these sounds very well and the other group rating them very poorly. If you simply average the results at this point, these sounds will receive a median rating. However, this median rating will not reflect the participants' evaluation. A metric created on the basis of these scores will not give a good prediction of the sound quality. In such a case, you have to estimate which results are to be considered for your sounds and must not include the evaluations of the other group in the averaging. You may need to revise your test design and conduct another jury test as a check, or ask about and document the reasons for the subjects' conflicting ratings by conducting an appropriate interview, for example.

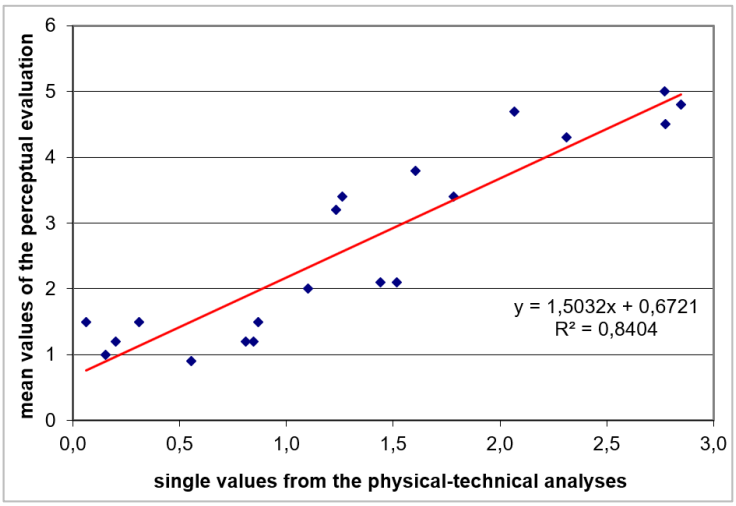
Care in statistical evaluation of jury test results ensures good results.

Further information on planning and evaluating jury tests can be found on our website Application Notes Sound & Vibration (head-acoustics.com) in section „Listening tests".

3. Important information on creating metrics

ArtemiS SUITE Metric Project

The Metric Project in ArtemiS SUITE calculates a metric based on the correlation of two data series, e.g., the single values from physical-technical analyses and the perceptual evaluations from a jury test. Both manual input and semi-automatic determination of the calculation rule are possible. In semi-automatic mode, the Metric Project will support you and map the jury test results to measurement analysis results in the best possible way by calculating a linear regression model. You can select any number of previously calculated analysis single values, and ArtemiS SUITE will automatically determine a corresponding metric.



Example of a regression analysis result

Correlation

- Linear dependency of two data series

Correlation coefficient R

- Measure for the strength of the correlation, range of values {–1; 1}
  - R = –1: strong, negative correlation
  - R = 0: stochastic correlation
  - R = 1: strong, positive correlation

Regression

- Numerical modeling of the relationship between two data series

Coefficient of determination R^2

- Measure for the quality of the modeling, range of values {0; 1}
- In linear regression, the coefficient of determination corresponds to the square of the correlation coefficient.
  - R^2 = 0: no correlation
  - R^2 = 1: high modeling quality

The correlation coefficient R is one of the values displayed for each calculated analysis. If there is a strong linear correlation between the analysis and the data from the jury test, this will result in a high correlation coefficient (maximum value: R = 1).

In addition, the quality of the current metric can be checked in the Metric Project. For this purpose, the coefficient of determination R^2 may be used which indicates the proportion of the variance in the jury test data explained by the regression model. A high coefficient of determination thus indicates that the jury test results can be reproduced very well by implementing the mathematical formula found and the results from the technical measurement analysis (maximum value: R^2 = 1).

Developing a robust metric

In principle, metric development should aim for a high coefficient of determination of the resulting calculation rule. This is because a high coefficient of determination indicates that the metric maps the jury test results well with the single values of the technical measurement analyses. However, a high coefficient of determination must not be the sole optimization criterion.

Instead, the goal of defining a robust metric is to be able to predict not only the results of the current jury test (training data) but also the sound quality of additional sound samples.

Concrete procedures

The following procedures are to be considered for the development of a robust metric:

- The impact of each single value needs to be systematically examined and selected with regard to the causal relationship regarding the noise aspect under investigation.
  Example: If you exclusively wish to evaluate broadband sounds, a high correlation to the single values of the tonality can only occur by chance. Despite an apparently high correlation, this single value is not to be included in the metric.
- The jury test results are not to be mapped with too large a number of single values from many different technical analyses (predictors). This usually only leads to an overfitting of the metric to the noise samples that were used as trainings data for the metric creation.
  That is, a high number of predictors may increase the correlation to the jury test results used. However, it usually does not increase the predictive quality for unknown sounds, i.e., sounds that were not used to determine the metric. In order to create a robust metric, it is often more appropriate to use only a small number of analyses as predictors.
- A best practice for creating a robust metric is to randomly split your data into two groups. Use the first data set (training data set) to create your metric and the second data set to validate it (validation data set). With perceptual evaluations also being available for the validation data set, you can compare the calculated results with the results of the jury test and check your metric. If you did not focus exclusively on a high coefficient of determination when creating the metric but instead limited yourself to using a few plausible single values, your metric is supposed also to predict the results of the second data set well.

Further information

Much useful guidance on creating robust metrics is provided in the following publication: Fiebig, Kamp; “Development of metrics for characterizing product sound quality”, Proceedings Aachen Acoustics Colloquium 2015, 123–133.

Example

The following example[1] is intended to illustrate that the use of additional analyses is not always expedient, and that the selection of technical analyses also needs to be made with care. For the example, the sounds of twelve hair dryers were evaluated in a jury test by several participants on a categorical, ten-point scale.
In order to determine a metric, the single values of the following analyses were calculated first:

- Loudness
- Sharpness
- Tonality
- Speech Intelligibility Index (SII)

[1] The example is purely fictitious and serves only to illustrate the procedure. The numerical values given do not indicate when a sufficiently high correlation or coefficient of determination is achieved. There are no generally valid limits for these values. Instead, it must be decided on a case-by-case basis when the agreement between jury test results and the metric is sufficiently high.

Interpretation of the indicated values

After calculating the analysis single values, the correlation values R between the analysis values and the jury test results for each analysis are displayed in a table. In this table, the desired single values for the metric can be activated. A corresponding metric based on the activated single values is automatically calculated by ArtemiS SUITE. The formulæ for this metric as well as its correlation coefficient R and coefficient of determination R^2 are directly displayed.

#		Name	Math. Function	Scale	R	R^2	P [%]	R resid.
1	☑	Loudness.Max	None	Lin.	0,91	0,82	0	
2	☑	Sharpness.Max	None	Lin.	0,87	0,76	0	
3	☐	Tonality.Default	None	Lin.	0,45	0,20	14	0,50
4	☐	SII.P5	None	Lin.	-0,83	0,68	0	-0,36

Correlation values between analysis values and jury test results

R_{resid}

In the present example, the single values of loudness and sharpness were activated first, as the values of these analyses show a significantly high correlation with the jury test results (*R = 0,91* and *R = 0,87* respectively). The coefficient of determination of the modeling based on these two analyses is *R^2 = 0,88*. In order to increase this further, another single value is to be integrated into the metric. When selecting the additional parameter, it is not only the correlation coefficient R of the respective analysis that is to be considered but also, for example, the correlation coefficient to the residuals R_{resid}. In this case, the residuals represent the deviations between the values from the jury test and the values calculated with the current metric. If the residuals are small, the current metric will reflect the values from the jury test well. The indicated value R_{resid} describes the linear dependency between the respective analysis single value and the residuals. A high R_{resid} value means that the single values of this (deactivated) analysis are highly correlated to the residuals. That is, this analysis can probably reduce any existing prediction errors of the present metric and improve the coefficient of determination of the metric.
This is even possible if the correlation coefficient R of this analysis is lower than that of others. A small R_{resid} value suggests that activating this analysis will barely improve the metric.
In the present example, the speech intelligibility index SII actually has a higher correlation[2] to the jury test results than has tonality. However, the correlation to the residual is higher for the tonality:

- Tonality: *R = 0,45*, *R_{resid} = 0,50*
- SII: *R = –0,83*, *R_{resid} = –0,36*

Improvement by an additional predictor

This means that activating tonality will increase the coefficient of determination of the resulting metric to a greater extent than that of the speech intelligibility index. This is because the speech intelligibility index responds to high levels, as does loudness. Thus, activating the speech intelligibility index does not provide any additional information if a single value such as loudness is already activated. In contrast to that, tonality does provide additional information (a measure for the tonal components contained in the sound) and improves the coefficient of determination of the resulting metric to *R^2 = 0,9*.

[2] The SII is defined in such a way that high values indicate good speech intelligibility and low values indicate poor speech intelligibility. The jury test results were coded such that a good evaluation corresponds to a low numerical value. In contrast to loudness, sharpness, and tonality, speech intelligibility thus shows a negative correlation to the jury test results.

Nonetheless, the increase from *0,88* to *0,9* is only a minor improvement. In order to determine whether the metric can actually be improved by using the additional predictor, it should be checked with the help of a validation data set. This will help to rule out that the additional predictor does not cause an overfitting to the training data, but that the metric improves the prediction quality for other noise samples as well.

Conclusion

Thus, the example shows that when creating metrics, analyses with high correlation coefficients are not to be the only ones to be taken into account. Instead, users need to contemplate which analysis contains additional information on the sounds and covers another relevant aspect of noise.

4. Using the created metric

Applying the sound quality metric

Provided that a robust metric was created, the following evaluations of further sounds can be predicted numerically by using the mathematical formula and the results of the technical measurement analysis. It is very important to consider that the sound quality metric is only applied to sounds with similar sound characteristics. Only in this way can the metric be used to make meaningful predictions. Using the metric for other types of sounds does not provide convincing results in many cases.

Example: Sounds of sports cars during acceleration were used to perform the jury tests and create the metric. The resulting metric will reproduce the sound quality of comparable recordings very well. However, the metric will fail if it is used to evaluate idle measurements of luxury cars. Even though in both cases the sounds were generated by internal-combustion engines and measured at comparable positions (e.g., at the passenger position inside the car) and with comparable equipment, these sounds are barely comparable and cannot be analyzed with the same metric in a meaningful way.