inter.noise
HAMBURG 2016

# Modeling sound quality from psychoacoustic measures

Lena SCHELL-MAJOOR[1]; Jan RENNIES[2]; Stephan D. EWERT[3]; Birger KOLLMEIER[4]

[1,2,4] Fraunhofer IDMT, Hör-, Sprach- und Audiotechnologie & Cluster of Excellence Hearing4All, Oldenburg

[3,4] Medizinische Physik & Cluster of Excellence Hearing4All, Universität Oldenburg, 26111Oldenburg

## ABSTRACT

Sound quality is a complex perceptual measure depending on several factors, which can often be associated with more basic psychoacoustic measures such as roughness, sharpness, tonality or loudness. One frequently used method for modeling the overall quality of sounds is therefore to fit a linear combination of psychoacoustic measures to subjectively measured data of quality. The psychoacoustic measures are usually calculated using standards or descriptions from literature references. However, these standards or references may not always be accurate for the specific class of sounds for which sound quality shall be determined. In this study we therefore applied this method to subjectively measured psychoacoustic measures and compared the results to data from instrumental measures and to data from literature. By means of regression analyses model approaches were derived from parts of the data and used to predict another part of the data for validation. The results indicate that subjectively measured basic psychoacoustic measures are not always in line with standardized calculation procedures, and that subjectively measured annoyance can be predicted quite well by subjectively measured sharpness, loudness, roughness and tonality. These data serve as benchmarks for further model approaches and underline the general applicability of this method.

Keywords: Psychoacoustics, Sound quality, Sharpness I-INCE Classification of Subjects Number(s): 63.2, 63.7

## 1. INTRODUCTION

Quality assessment of sounds is relevant for the development process and utilization of many technical devices and products, e.g., household and automotive appliances. One key aspect in the applied research is developing models for predicting sound quality. They provide a cost-efficient way to estimate sound quality. One successfully applied method for predicting the overall quality of sounds is to fit a linear combination of psychoacoustic measures to subjectively measured data of quality (1, 2, 3). The psychoacoustic measures, e.g., roughness, sharpness, tonality or loudness, are usually calculated using standards or descriptions from literature references and their calculation is also included in many commercially available sound analyzing software tools (referred to as 'instrumental measures' in the following). These instrumental measures are usually derived from experimental data obtained by asking subjects but they are not necessarily validated or applicable for all different kinds of sounds. Although the method of linearly combining instrumental psychoacoustic measures for predicting the sound quality may work well for a certain class of sounds, it is not clear how well this method generalizes over a large set of sounds. Furthermore the prediction accuracy of predicted sound quality based on instrumental measures has not been benchmarked in reference to experimental data. This would illustrate how well instrumental measures exploit the potential of psychoacoustic measures to predict sound quality, because experimental data can be seen as the gold standard and illustrate the limits of this method.

In this study we investigated and the general applicability of the combination of weighted acoustic parameters for the prediction of an overall quality judgement by applying this method to a set of data where the psychoacoustic parameters roughness, sharpness, tonality and loudness had been assessed

[1] Lena.Schell-Majoor@idmt.fraunhofer.de
[2] Jan.Rennies@idmt.fraunhofer.de
[3] Stephan.Ewert@uni-oldenburg.de
[4] Birger.Kollmeier@idmt.fraunhofer.de

by subjects together with annoyance as an overall measure for sound quality. Furthermore, we evaluated to what extent instrumental measures exploited the potential of this method by comparing the accuracy of the predictions derived from experimental measures to predictions derived from instrumental measures.

## 2. METHOD

Two experiments were conducted, which used the same method but different sets of stimuli. The apparatus and experimental procedure was the same. As loudness was found to be one of the most important factors for sound quality in most studies on sound quality, we decided to put the focus on the other psychoacoustic measures in this study, and equalized the instrumental loudness for most of the sounds in experiment 1. These sounds were mainly simple artificial sounds (tones, bandpass-filtered noises) that were also used in other studies before. In experiment 2 mainly more complex, real sounds were used. A subset of the stimuli was the same in both experiments to assess the question to which degree the subjective ratings of the same stimuli were influenced by the entire set of stimuli in each experiment.

### 2.1 Subjects

Thirty subjects participated in each experiment. Their age ranged from 20 to 32 years (experiment 1, median 25 years) and 20 to 32 years (experiment 2, median 25 years), respectively. Two subjects participated in both experiments. All subjects reported normal hearing abilities and had pure-tone thresholds of less than or equal to 20 dB HL at audiometric frequencies in the range 125 Hz–8 kHz. Most of the subjects were unexperienced as they had not participated in listening studies before. The subjects were paid for their participation.

### 2.2 Stimuli

The stimuli included simple artificial sounds, which have been used in other studies before, and more complex, real stimuli, which are of interest for industrial applications. Seventy-four different sounds were presented in experiment 1. Four of them were recorded sounds, the remaining sounds were synthetic bandpass-filtered (BP) noises, amplitude-modulated (AM) tones, pure tones, and white noise. Based on ISO 532 B / DIN 45631All sounds had an instrumental loudness of 4 sone, except for the white noise (10 sone) and the pure tone with f = 2 kHz (5 sone). So white noise and all pure tones had a level of 60 dB SPL. Three AM-sounds were presented twice to evaluate the reliability of data within a session.

Experiment 2 included 78 sounds. 53 sounds were recorded signals. Most of them were presented at levels corresponding to an instrumental loudness of 12 sone, including those four sounds that were presented in experiment 1 where they had a loudness of 4 sone. The other sounds were white noise (10 sone), a 1-kHz pure tone (4 sone) and AM-tones (4 sone). Except for six additional AM-tones, all of these sounds were presented identically to experiment 1. Most sounds had length of about 2 seconds. Due to their characteristics, some sounds were longer, e.g., the sound of church bells. The longest sound had a length of 4 seconds. The sampling frequency of all sounds was $f_s$ = 44100 Hz. To avoid clicking at the beginning or the end of sounds, a hanning window with flanks of 1000 samples was applied and 4096 zeros were added at the end of all sounds.

### 2.3 Apparatus and experimental procedure

The experiments were conducted individually for each subject in a sound-attenuating booth. All sounds were presented diotically via Sennheiser HD650 headphones. At the beginning of the experiment the subjects got written instructions which mentioned the attributes to be rated (see below), but did not explain them. Each session started with a pre-experimental phase, where the subjects had to listen to nine sounds from the current set of stimuli to familiarize them with the range of stimuli to be expected in the main experiment. This familiarization was the same in test and retest sessions, and the same sounds were used for all subjects. After the familiarization was completed, the sound scaling started. A scale from 0 to 50 with numerical marks every five units was used. Above this numerical scale five verbal marks from “not” to “very” were displayed. The words (original German words: nicht, etwas, mittelmäßig, ziemlich, sehr) were taken from Rohrmann (4). In this study subjects had to rate their perception of the five attributes roughness, sharpness, tonality, loudness and annoyance (original

German words: Rauigkeit, Schärfe, Tonhaltigkeit, Lautheit, Lästigkeit) for each sound using sliders. The sliders were arranged one below the other each labeled with one of the five attributes in the order as given above (see Figure 1). The initial position of all sliders was at 0 at the beginning of each rating. Above the upper slider the numerical and verbal marks were displayed. All sliders were always active, so subjects could switch between rating the different attributes for one sound and change them until they confirmed the given ratings. Each slider had to be moved at least once before the ratings could be confirmed. Once the ratings were confirmed, the subjects could neither see nor change the past ratings. The order of the sounds was randomized for each subject. The duration of each session including the familiarization and the assessment of a little less than 80 sounds was approximately one hour. Each subject participated in two sessions (test and retest) with a minimal distance of one week in between.

### 2.4 Instrumental measures

The instrumental measures used for the regression analyses in this study were obtained using the software Artemis Classic 11.0. Before the measures listed in Table 2 were calculated, the signals were filtered using an FIR filter with 1024 coefficients to equalize the transfer function of the headphones. This filter was matched to the inverse transfer function measured with the headphone on an artificial ear (Brüel & Kjær Type 4153).

Table 1 – List of instrumental measures calculated in Artemis

| Measure from Artemis |
|---|
| 'Roughness vs. Time' |
| 'Sharpness vs. Time [FFT ! ISO 532 B, Aures]' |
| 'Tonality DIN 45681 vs. Time' |
| 'Loudness vs. Time [FFT ! ISO 532 B]' |

## 3. RESULTS

Linear regression analyses were performed with varying input data and predictors. The resulting coefficients were used to predict different sets of validation data. For input and validation data either data from either experiment 1 (Exp 1) or experiment 2 (Exp 2) or both experiments (All data) were used. The predictors were all included measures (sharpness, tonality, roughness, and loudness) or just sharpness or just tonality. Sharpness was chosen to be tested as a single predictor because it showed the highest correlation with the annoyance data in the experiment (R = 0.85). Among the instrumental measures tonality was found to yield the highest correlation with the annoyance from the experiment (0.52), therefore it was also tested as a single predictor.

### 3.1 Experimental Measures

The results showed that the annoyance ratings could be predicted by a weighted combination of the psychoacoustic measures rated in the experiments. Including all measures as predictors led to explained variances (i.e., coefficients of determination, $R^2$) between 70 % (data from experiment 2 predicted with data from experiment1) and 91% (data from experiment 1 predicted with data from experiment 1). Using only sharpness as predictor resulted in explained variances of 72% or 73% for the different conditions. Tonality did not predict the annoyance well with explained variances between 1% and 20%.

Table 2 – Explained variance of predictions of validation data with coefficients from linear regression with different input data and different experimental predictor variables

| Input data | Predictor | Validation data | Explained variance |
|---|---|---|---|
| Exp 1 | All measures | Exp 1 | 91% |
| Exp 1 | All measures | Exp 2 | 70% |

| | | | |
|---|---|---|---|
| Exp 2 | All measures | Exp 1 | 79% |
| Exp 2 | All measures | Exp 2 | 79% |
| All data | All measures | All data | 84% |
| Exp 1 | Sharpness | Exp 1 | 72% |
| Exp 1 | Sharpness | Exp 2 | 73% |
| Exp 2 | Sharpness | Exp 1 | 72% |
| Exp 2 | Sharpness | Exp 2 | 73% |
| All data | Sharpness | All data | 72% |
| Exp 1 | Tonality | Exp 1 | 20% |
| Exp 1 | Tonality | Exp 2 | 1% |
| Exp 2 | Tonality | Exp 1 | 20% |
| Exp 2 | Tonality | Exp 2 | 1% |
| All data | Tonality | All data | 8% |

### 3.2 Instrumental Measures

To transfer the time-dependent measures extracted from Artemis into single numbers, the maximum and the mean value were calculated. Regression analyses were performed with both sets of data. The first value in the last column of Table 3 refers to the regression with maximum values and the second value to regression with mean values. In most cases using the maximum values from the instrumental measures leads to better results in terms of more explained variance. The highest amounts of explained variances were reached with all measures as predictors and the same set of data as input and validation data (Exp 1: 53%, Exp 2: 72%, all data: 47%). The results also showed that sharpness was not applicable as a potential predictor yielding only up to 17% explained variance. Tonality on the other hand was not as good as using all measures as predictors, but with explained variances up to 43% much better than sharpness alone.

Table 3 - Explained variance of predictions of validation data with coefficients from linear regression with different input data and different instrumental predictor variables

| Input data | Predictor | Validation data | Explained variance (max / mean) |
|---|---|---|---|
| Exp 1 | All measures | Exp 1 | 53% / 3% |
| Exp 1 | All measures | Exp 2 | 6% / 19% |

| | | | |
|---|---|---|---|
| Exp 2 | All measures | Exp 1 | 16% / 2% |
| Exp 2 | All measures | Exp 2 | 72% / 50% |
| All data | All measures | All data | 47% / 13% |
| Exp 1 | Sharpness | Exp 1 | 1% / 1% |
| Exp 1 | Sharpness | Exp 2 | 11% / 17% |
| Exp 2 | Sharpness | Exp 1 | 1% / 1% |
| Exp 2 | Sharpness | Exp 2 | 11% / 17% |
| All data | Sharpness | All data | 5% / 7% |
| Exp 1 | Tonality | Exp 1 | 43% / 1% |
| Exp 1 | Tonality | Exp 2 | 15% / 10% |
| Exp 2 | Tonality | Exp 1 | 43% / 1% |
| Exp 2 | Tonality | Exp 2 | 15% / 10% |
| All data | Tonality | All data | 27% / 6% |

## 4. DISCUSSION

The results of the regression analyses with experimental data showed that the approach to predict the annoyance of sounds by the psychoacoustic measures roughness, sharpness, tonality and loudness is generally applicable. The annoyance of the sounds in this study can be predicted quite well by the experimental psychoacoustic measures yielding explained variances up to 91%. Usually a large influence of (instrumental) loudness on annoyance or sound quality in general is reported as, e.g., in (1, 5, 6). In this study the instrumental loudness of most of the sounds was equalized to avoid the dominance of this measure. Given this reduced influence of loudness, sharpness had the largest influence in this experimental data in this study and was a good predictor without any additional measures. This is in line with findings from several other studies, e.g., (3,7,8), although these studies used instrumental data to predict sound quality. However, in the present study the instrumental measures did not show such a good prediction performance and, among the instrumental measures, tonality was found to have the largest correlation with annoyance. One possible cause for these differences could be that there may be different descriptions and calculation rules for the psychoacoustic measures. Even if there are at least standards for loudness and sharpness, also other ways to calculate these measures can be used. Furthermore the validation of the instrumental measures is often done with artificial sounds in order to control for certain parameters of the sounds and, especially in studies with real sounds, differences between experimental and instrumental data is often found as, e.g., in (9) for loudness. Another reason could be that most of the studies investigated rather limited classes of sounds and therefore one measure might be of particular importance for the respective sounds. The results showed that the potential to predict annoyance from roughness, sharpness, tonality and loudness is not fully exploited when current instrumental measures are used, and that better predictions can be achieved when the basic psychoacoustic measures are determined experimentally by the same subjects. Even if some instrumental measures are able to predict the sound quality of a certain class of sounds, these results do not generalize to a large set of different sounds. The employed method allows for a comparatively fast and efficient assessment of both overall quality measures and basic psychoacoustic quantities and can thus be valuable tool for sound quality studies.

## 5. SUMMARY AND CONCLUSIONS

In this study regression analyses were performed with experimental and instrumental psychoacoustic measures as predictors for the annoyance of a large set of different sounds. Results from using experimentally assessed psychoacoustic measures were compared to results obtained with instrumental ones. The results showed a larger amount of explained variance for predictions with experimental measures. The following conclusions were drawn

- Psychoacoustic measures can serve as predictors for the annoyance of sounds (explained

variances between 70% and 91% for experimental data).
- Experimentally measured sharpness alone is a good predictor for the annoyance of the investigated sounds with explained variances between 72% and 73%.
- The instrumental measures used in this study did not exploit the potential of this method to predict the annoyance of the sounds in this study.

## ACKNOWLEDGEMENTS

The authors would like to thank Marcel Schulze for his help in data collection.

## REFERENCES

1. Ellermeier W, Kattner F, Kurtze L, Bös, J. Psychoacoustic characterization of the noise produced by photovoltaic inverters. Acta Acustica united with Acustica 2014; 100:1120-1128.
2. Kim EY, Shin TJ, Lee SK. Sound quality index for assessment of sound quality of laser printers based on a combination of sound metrics. Noise Control Engr. J. 2013;6 (6): 534-546.
3. Kuwano S, Seiichiro N. Subjective impression of copy machine noises: An examination of physical metrics for the evaluation of sound quality. Proc INTER-NOISE 09; 23-26 August 2009; Ottawa, Canada 2009. p. 746-52.
4. Rohrmann B. Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. Zeitschrift für Sozialpsychologie 1978; 9: 222-245.
5. Schell-Majoor L, Mores R. Comfort parameter for aircraft acoustics. Proc. of the EAA Forum Acusticum 2011, 26 June - 1 July 2011; Aalborg, Denmark.
6. Fastl H. Sound quality of electric razors – effects of loudness. Proc. of INTER-NOISE 00; 27-30 August 2000; Nice, France 2000. p. 4425-31.
7. Fastl H, Zwicker E. Psychoacoustics - Facts and Models. 3rd ed. Heidelberg, Germany: Springer; 2007.
8. Zwicker E. A proposal for defining and calculating the unbiased annoyance. In: Schick, A. Contributions to Psychological Acoustics. Oldenburg, Germany: BIS Oldenburg; 1991. p. 187-202.
9. Rennies J, Wächtler M, Hots J, Verhey J. Spectrotemporal characteristics affecting the loudness of technical sounds: Data and model predictions. Acta Acustica united with Acustica 2015;101:1145-1156.