On Predicting the Perceived Annoyance of Wind Turbine Sound

Stephan Preihs[1], Jakob Bergner[1], Daphne Schössow[1], Jürgen Peissig[1]
[1] *Leibniz University Hannover (LUH), Institute of Communications Technology (IKT)*
Appelstr. 9A, 30167 Hannover, Email: {name.surname}@ikt.uni-hannover.de

Introduction

Human perception of environmental sound sources receives more and more attention nowadays due to an increased sensitivity to noise pollution. In the case of wind turbines, in recent years there have been additional efforts to include perceptual aspects in the assessment and prediction of noise immissions in supplement to standardized sound measurements.

For gaining further insights on different perceptual aspects of wind turbine noise (WTN) annoyance, laboratory studies are a popular and accepted methodology in current research [1]. In addition it would be desirable to be able to robustly predict the outcome of such studies on the basis of a psycho-acoustically motivated signal analysis of WTN stimuli. One of the goals of our project WEA-Akzeptanz [2] therefore was the development of a methodology for WTN annoyance prediction.

The basic idea of our approach for a combined signal analysis of WEA sound and a neural network training for psychoacoustic annoyance modeling presented in the following is inspired by an algorithm that has been established for over two decades in audio coding for assessing the quality of lossy audio codecs. In the case of the method *PEAQ* covered by the ITU in the standard ITU-R BS.1387-1. (Perceptual Evaluation of Audio Quality) [3] the approach is a combination of a hearing model and a neural network trained on listening test data.

Figure 1 shows the basic structure of the PEAQ-processing chain. Here, an original signal and a signal to be evaluated (the encoded audio signal) are to be passed to the method at the input. On the basis of a peripheral hearing model, a number of signal characteristics of the coded audio signal in comparison to the original signal are then determined. These are then combined in further processing stages into so-called *Model Output Variables* (MOVs) and finally mapped via a neural network to the so-called *Objective Difference Grade* (ODG). This gives a reproducible estimate of the subjective perception with respect to the annoyance of coding artifacts rated by a listening test participant on a scale from 0 to −4 (cf. box on the right in Figure 1).

Our goal within the scope of this contribution was to transfer this approach to our application of rating the annoyance caused by WTN stimuli. This paper therefore is structured as follows. First we describe how our reference annoyance values were gathered and which stimuli were used. Then we give an overview about the signal processing techniques we used for extracting features from the stimuli. The following section will describe the actual training of the annoyance model. In a final section an evaluation of the training results and additional investigation will be presented. We will then conclude this paper and give an outlook on possible future work.

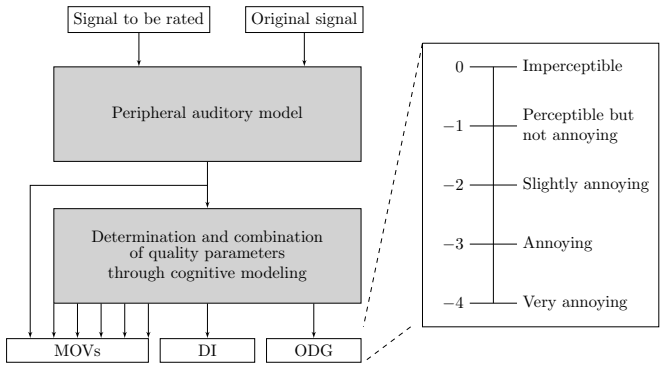

Figure 1: Source of inspiration: Flow graph of the PEAQ-algorithm (on the basis of [4]) and verbalization of the scale of the PEAQ-"Objective Difference Grade".

Methodology

In the following we will explain in which way we gathered our reference annoyance values, how we extracted meaningful signal characteristics from the stimuli and formed so called features out of them and how the actual training of the annoyance model was carried out.

Reference Annoyance Values

According to the methodology of using listening test data for the training of an annoyance prediction model, a set of stimuli accompanied by a subjective rating of the perceived annoyance is needed as input for the training procedure. This is one reason why we conducted a series of listening experiments which we also describe in [5]. For further details about the experimental setup as well as the statistical analysis please refer to the paper. At this point we will only explain those parts of the experiment that are relevant for the creation of the annoyance model.

The study consisted of 13 binaural stimuli that were to be rated by the subjects on an 11-point Likert Scale. The set of stimuli can be seen in Table 1 and is composed of soundscape recordings of wind tubines in different distances in their respective acoustic environment. The test set also contains two anchors which are a helicopter flyover and a rural soundscape that does not include any WT sound. In total 60 participants aged from 18 to 74 finished our online survey and in the end we identified 48 valid data sets.

Looking on the results (see Figure 3) of the annoyance ratings done by the subjects we can see that the mean of the ratings falls between 2 and 8. The error bars show the 95% confidence intervals according to a normal distribution that was fitted to the data. The most annoying

stimuli are as expected the wind turbine at a distance of 150 m as well as the helicopter flyover. The least annoying ones are the wind turbine at a distance of 700 m and the rural soundscape. In general the trend of a distance related annoyance rating is quite well reproduced.

Table 1: Details of the WTN scenes used in the online survey.

Scene	Distance	Special features
01	150 m	
02	390 m	Tractor + car passing
03	390 m	
04	390 m	More vegetation noise
05	700 m	Agriculture sounds
06	700 m	Car passing
07	700 m	View towards street (no WT), car passing
08	700 m	View towards street (no WT)
09	700 m	Audio: scene 8, video: scene 5
10	N/A	Anchor: Helicopter flyover
11	N/A	Anchor: Rural soundscape
12	400 m	Training example
13	700 m	Training example

Extracted Signal Parameters

Another important step in the creation of an annoyance model is the extraction of meaningful signal parameters from the stimuli used in the listening experiments.

As a basic physical parameter describing one major signal property, the LAeq of the stimuli was calculated. Therefore it was necessary to decide which microphone signals to use as a basis, either those of a sound level meter microphone on a rigid plate or those of the calibrated dummy head at ear level (approx. 165 cm). The comparison of both recordings with respect to the total, as well as third-octave sound levels showed varying differences, which could not be explained with a simple offset. Therefore, for our modeling it was purposeful to use the LAeq of the signal that was also presented in the subject studies, i.e. that of the dummy head. This decision ensures that results of the studies can be reconciled with those of the physical analysis. This is true even if the measurements with the measurement microphone on a rigid plate potentially contain less interference from wind as well as the high quality of the measurement microphone used results in less spectral artifacts.

The traditional perception-based parameters that were calculated include loudness, sharpness, and roughness (see Table 2). These parameters were each calculated separately for the left and right channels of the binaural recording in short time frames, resulting in a time series for each channel. The loudness and sharpness parameters were implemented directly using the Matlab Audio Toolbox, while roughness was calculated using an adapted version of the fluctuation strength implemented there.

As can seen in Table 3 together with a short description, we computed the mean and standard deviation for those parameters to create a feature value. As pre-processing, the values of each feature are normalized to the respective maximum over all items.

In addition we developed two features that were specifically designed to quantify the occurrence of two characteristic effects of wind turbine sound since we suppose that they have an influence on the perceived annoyance. The first one we called binaural periodicity index and is based on an analysis of periodical fluctuations in binaural features like ILD and ITD [6]. The second one is the so called blade whistling index and is a measure for the occurrence of periodically changing tonal components caused by defects in the WT blades [7].

Table 2: Overview of the calculation of conventional perceptual parameters.

Parameter	**Reference**	**Configuration**
Loudness	ISO 532-1 [8] (Zwicker)	Time variant (2 ms frame, result interpolated to 0,1 s), free field
Sharpness	DIN 45692 [9]	Calculated for frames of 0,1 s
Roughness	[10]	Calculated for frames of 0,1 s

Table 3: Overview of the features used for training the annoyance model.

ID	**Name of feature**	**Description**
1	LAEq	Equivalent A-weighted sound pressure level
2	Loudness mean	Loudness according to ISO 532-1 averaged over the item
3	Loudness std	Standard deviation of loudness
4	Sharpness mean	Sharpness according to DIN 45692 averaged over the item
5	Sharpness std	Standard deviation of sharpness
6	Roughness mean	Roughness according to Fastl averaged over the item
7	Roughness std	Standard deviation of roughness
8	Binaural period. index	Measure for the occurrence of binaural fluctuations
9	Blade whistling index	Measure for the occurrence of blade tip whistling

Training of the Annoyance Model

The actual training of the mapping from feature values to a global annoyance parameter is done using a so-called *Shallow Neural Network* resp. *Function Fitting Neural Network* with the Matlab Neural Network Toolbox. Figure 2 shows the network topology. A nine-dimensional vector is fed to the network as input. The feature data is processed in 10 so-called *hidden layers*. Another layer, the so-called *Output Layer*, provides the mapping to the scalar output value. The optimization of the network parameters is done with the *Bayesian Regularization Backpropagation*, which is known for achieving a high degree of generalization even for smaller data sets at the expense of a higher computational cost. The splitting (in the following also called *split*) to training, validation and test set is done in the order of established ratios with 68, 16 and 16 %, so that two items each are stored in the validation and test set.

Our neural network training framework allows to perform various investigations on the dependence of the training results on the number and composition of the input features used. Also, the division of the dataset

into training, validation and test sets can be traced and arbitrarily adjusted according to different metrics. Thus, within the constraints of the dataset used, it is possible to trace the ability of generalization of the trained network.

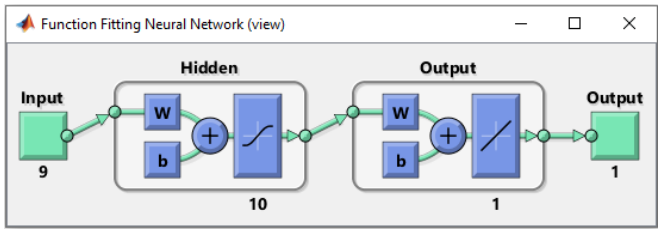

Figure 2: Layout of the neural network used (Screenshot from the Matlab Neural Network Training toolbox [11]).

Evaluation of Training Performance

The results of the training are shown in the following as an example, and the most important studies on the dependence of the training result on the split and the number of features used are presented. Figure 3 shows the result of a training where items 2 and 8 were in the test set and all others were in the training and validation sets, respectively. The graph plots the mean from the listening test results including the respective 95 % confidence interval over the 13 items in blue. In addition, the mean predicted by the neural network is shown in red. It can be seen that this falls within the confidence interval of the true means for all items. However, the result in Figure 3 is a *best case* estimate, since the training result depends more or less strongly on the split used due to the small size of the data set used.

Figure 4 shows the corresponding *worst case* estimate. Here, items 10 and 11 were in the test set, placing the prediction well outside the confidence intervals. The results of both graphs are from an investigation in which several hundred trainings of the neural network were performed with random splits and the RMSE between true means and predicted means was computed as a quality criterion. Figure 5 also shows a statistical analysis of this study as a boxplot. Here it can be seen from the very small interquartile ranges and the whiskers close to the median that for a large number of the items and trainings the prediction falls within the confidence interval of the listening test results regardless of the split. As might be expected, however, items 1 and 10 are particularly susceptible to assignment in the split due to their significantly higher perceived annoyance, and thus there may be outliers in the prediction results. However, this fact does not call into question the feasibility of training a discriminative model. Accordingly, it must only be ensured that the model in the training gets to see as much as possible of the entire range of annoyance levels and thus feature expressions.

In order to investigate the influence of the number of features as well as their contribution to the overall result and thus to ensure that they are meaningful, trainings were performed with an increasing number of features. Here, for each number of features (starting with the first

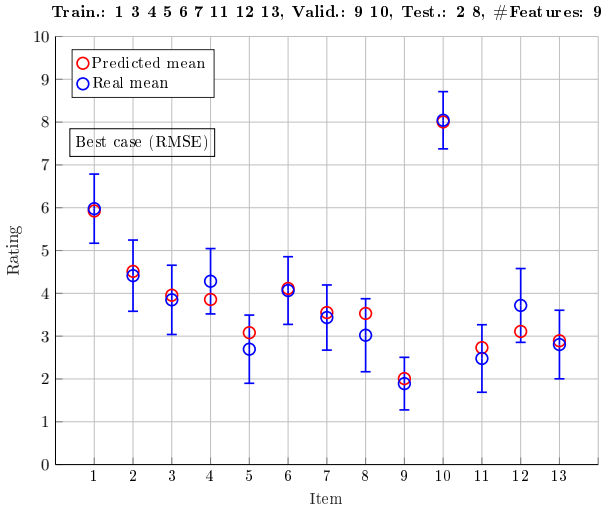

Figure 3: Best-case result of listening test result prediction using signal features and a neural network.

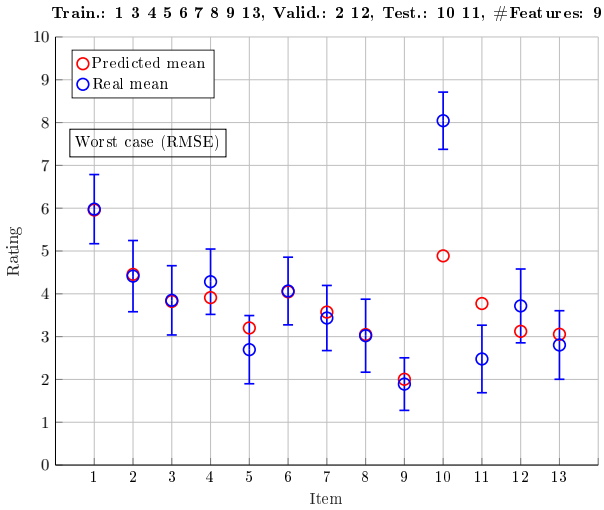

Figure 4: Worst-case result of listening test result prediction using signal features and a neural network.

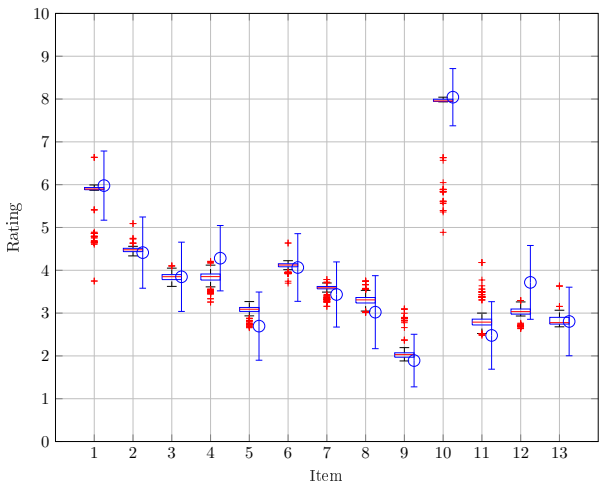

Figure 5: Statistical analysis of the dependence of the training result on the split used.

one from Table 3 and then always adding the next one in the list as another one), a network was trained 100 times and the training error was evaluated according to different metrics. Figure 6 shows the mean of the averaged absolute errors between prediction and listening test results as a function of the number of features.
Figure 7 shows the mean correlation between the vector of prediction and listening trial results as a function of the number of features. For both metrics, it seems to be confirmed that, in tendency, adding more features to the training is useful. The intermediate degradation of the results is not unusual for a neural network, since the information available with an additional feature can possibly only be exploited with another additional feature. Of course, this kind of evaluation also depends on the order or combination of features. However, since in our implementation the number of parameters of the neural network also increases with increasing number of features as well as a multitude of permutations of the feature vector is possible, this aspect was not further examined in the context of the investigations described here, especially since the small number of items might dominate the result under certain circumstances.

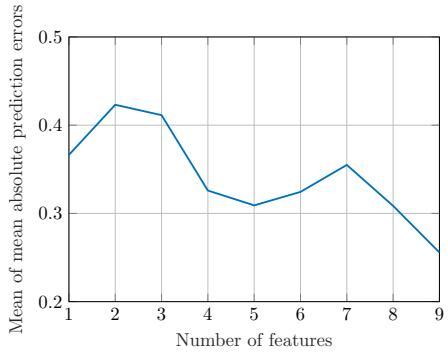

Figure 6: Analysis of prediction error as a function of number of features used: mean value of the mean absolute error.

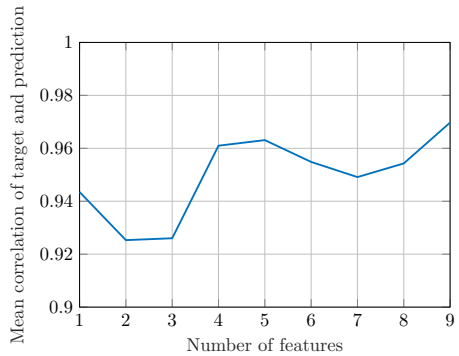

Figure 7: Analysis of prediction error as a function of number of features used: mean correlation of prediction and listening test result.

Conclusions and Outlook

In this contribution, we described the training of a WTN annoyance model using a neural network. We could show that a prediction of mean annoyance ratings based on characteristic signal properties is possible in principle. However, due to the relatively small data set we used and despite the additional investigations we conducted, we can only speak of a *proof of concept*, since *overfitting* to the data is to be assumed. Nevertheless, the developed framework for signal analysis and training of the annoyance model allows a flexible training even on larger datasets.
Looking ahead, the developed methodology can be used to conduct a variety of further investigations. For example, one of our next steps is to incorporate further parameters such as empirical data of the test subjects and the attitude towards wind turbines into the training of the model. With such a trained network, it would then be possible to make even more differentiated predictions of the expected annoyance rating by a test subject. In addition, further investigations into the robustness of the model and its integration into the WT acceptance prediction model are conceivable.

Acknowledgement

The project WEA-Akzeptanz was funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) under the funding code 0324134A.

References

[1] Alamir, M., Hansen, K., Zajamsek, B., and Catcheside, P., Subjective responses to wind farm noise: A review of laboratory listening test methods, Renewable and Sustainable Energy Reviews, 2019.

[2] Homepage of the Project WEA-Akzeptanz, URL: `https://www.wea-akzeptanz.uni-hannover.de`, Online; accessed 01-07-2021.

[3] ITU-R BS.1387-1, *Method for Objective Measurements of Perceived Audio Quality*, 2001.

[4] Thiede, T. et al., PEAQ - der künftige ITU-Standard zur objektiven Messung der wahrgenommenen Audioqualität, 20. Tonmeistertagung, 1998.

[5] Preihs, S., Bergner, J., Schössow, D., and Peissig, J., Assessing Wind Turbine Noise Perception by means of Contextual Laboratory and Online Studies, Wind Turbine Noise Conference, 2021.

[6] Preihs, S., Bergner, J., and Peissig, J., Ansätze zur binauralen Erweiterung einer Algorithmik zur lästigkeitsbezogenen Analyse und Synthese der Schallemissionen von Windenergieanlagen, Fortschritte der Akustik - DAGA, 2019.

[7] Leibniz Universität Hannover, Abschlussbericht zum Projekt WEA-Akzeptanz, Technical report, 2021.

[8] DIN, DIN ISO 532-1: Akustik - Verfahren zur Berechnung der Lautheit - Teil 1: Verfahren nach E. Zwicker, Technical report, 2020.

[9] DIN, DIN 45692 - Messtechnische Simulation der Hörempfindung Schärfe, 2009.

[10] Fastl, H. and Zwicker, E., *Psychoacoustics - Facts and models*, Springer Berlin Heidelberg, 3rd edition, 2007.

[11] The Mathworks, Inc., Natick, Massachusetts, *MATLAB version 9.9.0.1495850 (R2020b) Update 1.*