59th ILMENAU SCIENTIFIC COLLOQUIUM
Technische Universität Ilmenau, 11 – 15 September 2017
URN: urn:nbn:de:gbv:ilm1-2017iwk-083:1

APPLICATION OF ARTIFICIAL NEURAL NETWORKS FOR EDITING MEASURED ACOUSTICAL DATA FOR SIMULATIONS IN VIRTUAL ENVIRONMENTS

Antje Siegel[1], Christian Weber[1], Albert Albers[2], David Landes[2], Matthias Behrendt[2]

[1] Engineering Design Group, Technische Universität Ilmenau
[2] IPEK – Institute of Product Engineering at Karlsruhe Institute of Technology (KIT)

ABSTRACT

Acoustic simulation tools are used and demanded by various groups of people. Architects and urban planners as well as product designers and engineers are interested in simulating the acoustical properties of buildings, machines or other products. Acoustic simulation techniques are continually evolving. The current trend is towards integrating forward-looking technologies like virtual reality (VR) into the simulation process. Common acoustical simulation tools, such as numerical methods, are computationally expensive and cannot be applied in real time. This, however, is a mandatory requirement for VR-applications. For that reason, techniques based on measured acoustical data are often used for acoustic simulations in VR. However, various disturbance variables, such as interfering noise, can distort measurement results immensely. In this paper an Artificial Neural Network (ANN) is described which can be used for the post-processing of measured data. A concept specifically for the noise cancellation in acoustic measurement data is outlined.

Index Terms – artificial neural networks, acoustic VR-simulation, noise cancellation

1. INTRODUCTION

Acoustic simulation tools are used and demanded by various groups of people. Architects and urban planners as well as product designers and engineers are interested in simulating the acoustical properties of buildings, machines or other products. Acoustic simulation techniques are continually evolving. The current trend is towards integrating forward-looking technologies like virtual reality (VR) into the simulation process. Common acoustical simulation tools, such as numerical methods, are computationally expensive and cannot be applied in real time. This, however, is a mandatory requirement for VR-applications. For that reason, techniques based on measured acoustical data are often used for acoustic simulations in VR. However, various disturbance variables, such as interfering noise, can distort measurement results immensely. In most cases, artefacts in recorded data cannot be completely avoided. Due to this, methods have to be developed which allow a successful reduction or cancelation of undesired interferences with measured signals. In this paper an Artificial Neural Network (ANN) is described which can be used for the post-processing of measured data. A special kind of ANN, a so called autoencoder, is introduced and a concept specifically for acoustic measurements is outlined.
The investigations of this paper are based on an autoencoder that is described in [1]. As an example for the implementation of the concept a virtual traffic simulation is used. As described in [2], the goal of this traffic simulation is the realistic and plausible representation of vehicle noise for different traffic scenarios. For this purpose a virtual acoustic vehicle model was developed. This model, outlined in [3] and in [4], has a modular structure. In this way, the main sound sources of a vehicle can be analysed separately. The noise of each of the

vehicles sound sources has to be recorded. Afterwards the recorded audio signals are processed with an internally developed software tool that simulates the sound transmission from each sound source to several listener positions. Transfer functions that describe the acoustic transmission properties of the car as well as a simplified model for sound reflections at plane walls are included in the software tool. For the traffic simulation an acoustical database was created. This database includes the audio recordings of vehicles with different drive topologies: Cars with conventional internal combustion engines as well as vehicles with electric and hybrid drive topologies. The audio and vehicle data originates from recordings which were conducted on a vehicle driven on the acoustic roller test bench at IPEK – Institute of Product Engineering at Karlsruhe Institute of Technology (KIT). An aspect of roller test benches is that the rollers of the test bench generate a significant noise which is not negligible during acoustic measurements. The reasons for this are the material, shape and resonance body of the rollers. The noise interferes with the audio signal that actually is to be recorded. The autoencoder mentioned at the beginning of this section can be used to successfully remove the unwanted interferences. This paper shows how the autoencoder can be trained with a limited amount of training data.

2. CONCEPT

One challenge with regard to applying an ANN is to create a useful dataset for the training. In order to get a universal solution an ANN has to be trained with comprehensive datasets. In case of the underlying application example this is unavoidable because the noise of the test bench's rollers does differ between each measurement. Thus, a solution has to be found which can remove the rollers' noise out of different measurement data despite the fact that the exact noise signal cannot be forecasted exactly. However, it is not easy to generate a suitable training dataset. Often the effort for collecting meaningful training data is too high and disproportionate in relation to the benefit. Therefore, the investigations outlined in the following sections are aimed to design well-working solutions based on a manageable amount of training data.
For the investigations an autoencoder, which is a special kind of ANN, is used. The general structure of an autoencoder is shown in the following figure.

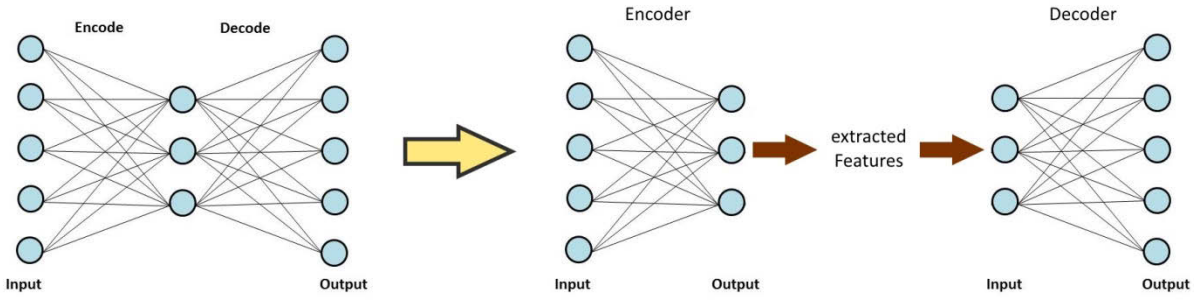

Figure 1: Structure of an autoencoder

An autoencoder is composed of an encoder and a decoder part [5]. The encoder's task is to reduce and compress the input data. As a result, so called features are extracted. These features are characteristic and representative for the content of the input data. The extracted features are used in the decoder to reconstruct data of the same format as the input data. For the investigations outlined in this paper an autoencoder that was introduced in [1] was used. This autoencoder can process input data of a length of 784 audio samples. In order to limit the complexity of the autoencoder this input size was maintained and audio files provided at the input are divided into individual parts of 784 audio samples. The audio data is processed

block by block. The autoencoder is a convolutional network since it consists, among other things, of two convolutional layers in the encoder and respectively two deconvolution layers in the decoder. The goal is to apply a disturbed signal on the input of the autoencoder and to get an undisturbed signal at the output of the autoencoder. This method is called “denoising”. In [5] use is made of this method in order to reduce artefacts in images. Also noise cancellation in audio signals can be realised by applying an autoencoder. This is described in [6]. In order to get a universal solution an autoencoder has to be trained with comprehensive datasets. Datasets containing audio recordings of several hours duration are not unusual. The TIMIT dataset [7], for example, is used for acoustic phonetic studies and contains recordings of 360 different speakers. Each of the 360 speakers performs 10 different sentences. Therefore, the TIMIT dataset contains 3600 spoken sentences in total. In case of the paper’s underlying application example, comparatively few measurement data could be collected. The following sections it describe how the training data was measured. Moreover, it is outlined how to use the measured data for an effective training.

3. MEASUREMENTS FOR THE TRAINING DATA

For the training data the sound of a vehicle was recorded. No special driving manoeuver was done; just the stationary noise with the car idling was measured. Fifteen microphones were placed at defined different positions around the car (figure 2). Seven microphones were placed in the nearfield of the vehicle. Furthermore, eight microphones were installed in the far field at a distance of 7.5 meters of the vehicle’s longitudinal axes. The car was fixated on a roller test bench, which was switched-off, since only the stationary noise of the vehicle was subject of the measurements. The goal was to create audio data without the noise of the rollers. In that way, 15 audio files with a length of 4.927 seconds for each file were recorded. These files served as reference files in order to validate the autoencoder’s performance during the training.

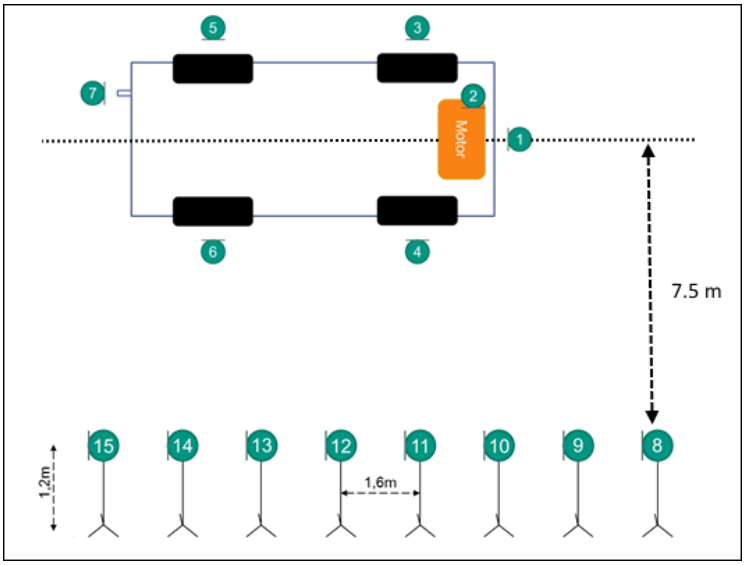

Figure 2: Measurement setup for stationary sound of vehicle

Further measurements were made in order to record the noise of the rollers. For this purpose, the rollers were driven in the same way as for the simulation of a steady speed driving at a speed of 50 km/h. This driving manoeuvre was chosen because steady speed driving is the focus of the traffic simulation which is used for the exemplary implantation of the concept that is introduced in this paper. The rollers’ noise was recorded with one microphone at a

distance of 7.5 m of the car's longitudinal axes (figure 3). The car was on the rollers during the recordings since the sound of the rollers is dependent on the mass which is placed on the rollers.

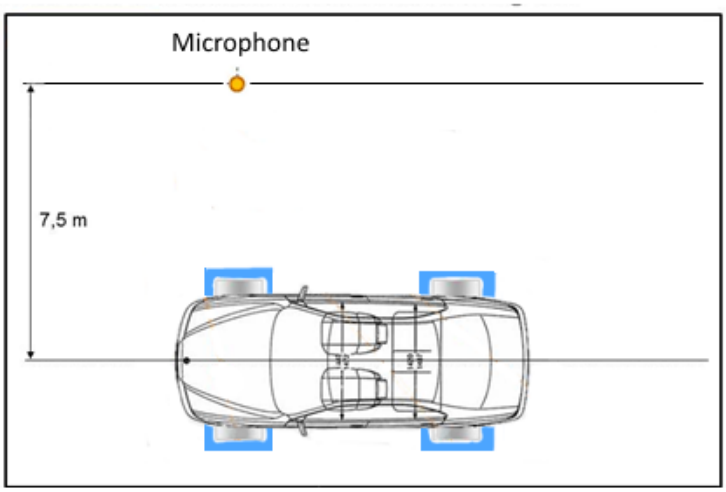

Figure 3: Measurement setup for the recording of the noise of the rollers

The 15 audio files from the microphones recording the car idling without the rollers were mixed with the noise signal of the rollers. For the mixed signals a signal to noise ratio (SNR) between 13.1 dB and 18.8 dB was chosen.

4. TRAINING OF THE AUTOENCODER

4.1 Direct training

During direct training the 15 mixed audio signals, which contain the noise of the rollers, are provided to the input of the autoencoder. The task of the autoencoder is to remove the rollers' noise out of the mixed signals. That means that at the output of the autoencoder a noise-free signal is generated. The output signal is compared to the audio files that were described at the beginning of the previous section. Each of the output signals is compared with the respective noise-free reference file and an error is computed. The autoencoder is adjusted in order to minimize this error. In that way the performance of the autoencoder is optimised.

4.2 Pre-training and fine tuning

Another training method is to pre-train the autoencoder. During this pre-training the autoencoder is taught to reconstruct and generate output signals with sufficient precision. Moreover, the feature extraction at the encoder's output is improved by the pre-training. The extracted features should be meaningful and characteristic for the essential information in the reference signals. For this purpose the reference signals without the rollers' noise are put at the input of the autoencoder. The task of the autoencoder is to generate an output signal that is equal to the input signal. The generated output is compared with the input of the autoencoder and an error is computed. On basis of this error the behaviour of the autoencoder is optimized. After the pre-training the autoencoder is fine-tuned. For that reason, the training of the pre-trained autoencoder is continued in the same way as described for the direct training. Now, the mixed audio files which also contain the roller's noise are used as input for the autoencoder and the error to be minimized is determined from the difference of the generated signal at the autoencoder's output and the noise-free reference signals.

4.3 Datasets for the training

For the training three different datasets were prepared. For the first dataset the audio records described in section 4.2 were divided into individual audio blocks with a length of 748 audio samples. The audio files without noise as well as the recorded noise of the rollers were edited as shown in the following figure (figure 4). The two kinds of signals were added to generate the mixed files. In that way 4515 audio blocks were created from the reference files and respectively of the mixed signals.

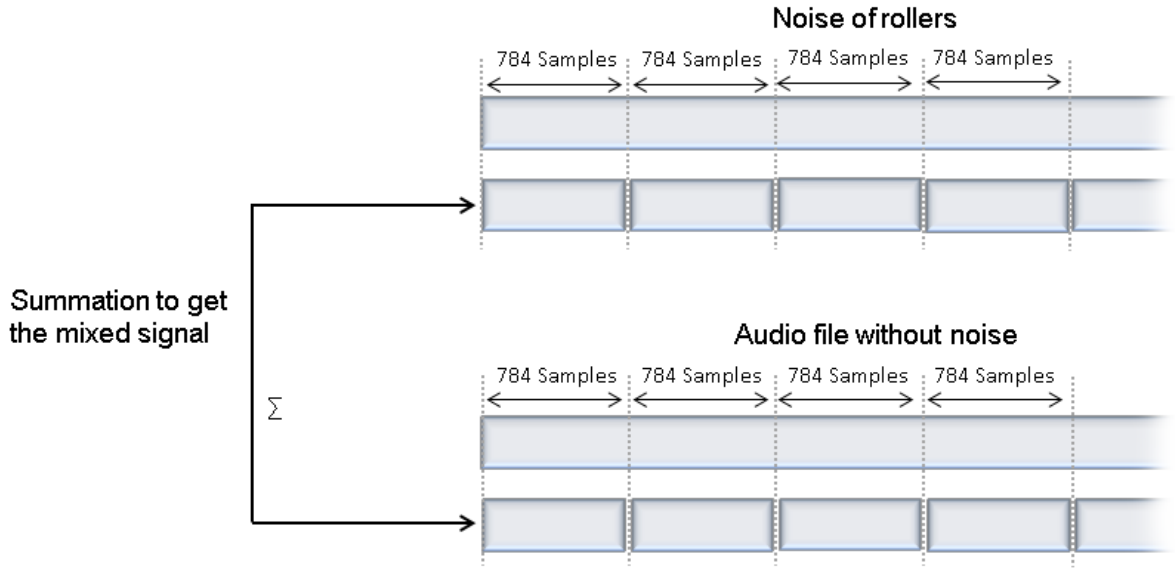

Figure 4: Dataset 1

For the second and third dataset the mixed signals were generated by varying the extracted audio blocks by a few samples. The separation into audio blocks was not started at the beginning of the audio files. Instead, the starting point for the first block was shifted by a few samples. The shift was 98 samples for the rollers' noise and 196 samples for the reference files in dataset 2. In dataset 3 a shift was chosen of 2 samples for the rollers' noise and of 4 samples for the reference files. The procedure for dataset 3 is outlined in figure 5.

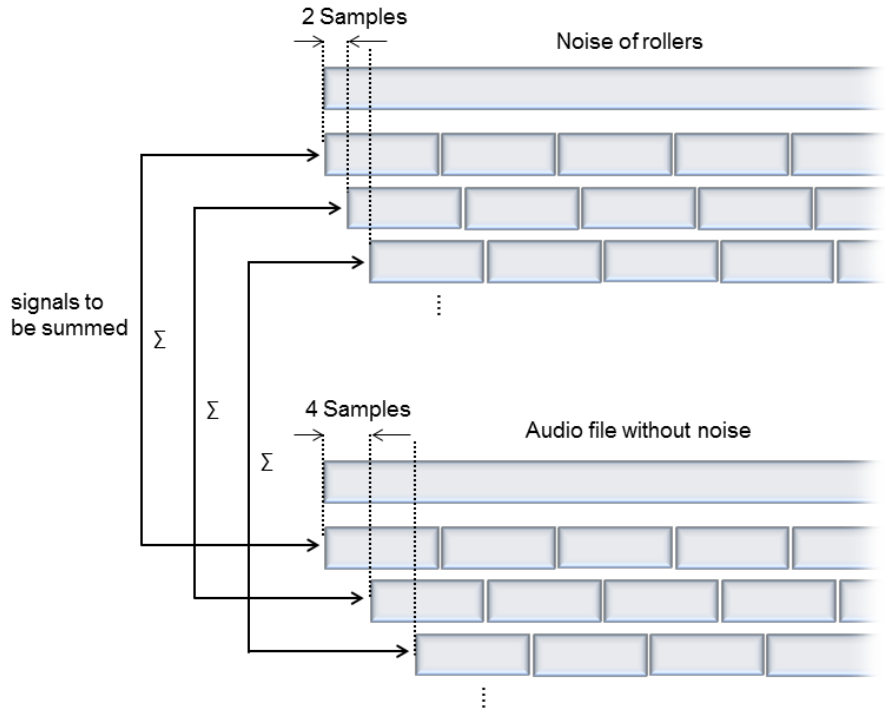

Figure 5: Dataset 3

The signals with shifted starting points were added in order to generate mixed signals. In comparison to dataset 1 the amount of the training data could be increased immensely. Dataset 2 consisted of 13543 audio blocks and dataset 3 contained 884011 audio blocks, (table 1).

Table 1: number of audio blocks for the three different training datasets

	Dataset 1	Dataset 2	Dataset 3
Number of audio blocks	4515	13543	884011

The autoencoder was trained with the three different datasets. The results of the different trainings are compared and discussed in the following section.

5. RESULTS

For validating the training results a new audio file was generated. This validation signal was created by also adding the recorded noise of the rollers to a stationary sound of a vehicle. However, the rollers' noise as well as the sound of the vehicle differ in each recording and are irreproducible signals. For that reason, the noise of the rollers was recorded several times during the measurements described in section 3. For validation, a different recording of the rollers' noise was used than for the training. Also the stationary sound, which is used for the validation, is different from the vehicle sound that was used for the training. Indeed, the stationary sound for the validation comes from a webpage that actually supports producers of audio dramas [8]. The independence of the training and validation data is important because only in this way reliable statements can be made about the universal validity and transferability of the results. In that way, it is ensured that the autoencoder does not only work for the data that was used for the training but also for noise signals and vehicle sounds that differ from the training signals. The disturbed validation data was also divided into audio blocks with a length of 784 samples. These audio blocks were put into the trained autoencoder

and the generated audio blocks at the autoencoder's output were rearranged one after another. In that way, a continuous audio signal could be reconstructed. The reconstructed audio data was compared to the original audio data that contained the noise-free validation signal. For the reconstructed and the original data the mean squared error (MSE) and the correlation coefficient were computed (table 2). With direct training an MSE of 0.005 could be reached for dataset 1. This result could be improved with dataset 2 by 24% and with dataset 3 by 44%. Two identical signals would result in a correlation coefficient of 1. Two different and absolutely uncorrelated signals would lead to a correlation coefficient of 0. After the training with dataset 1 a correlation coefficient of 0.7961 could be reached. This value could be improved by the training with dataset 2 to 0.8408 and with dataset 3 to 0.8856. The second training method (pre-training and fine tuning) was, for reasons of time, only done with dataset 3. As a result an MSE of 0.0027 and a correlation coefficient of 0.8916 were computed.

Table 2: Mean squared errors and correlation coefficients for the different trainings

	MSE	**Correlation coefficient**
Direct training		
Dataset 1	0.0050	0.7961
Dataset 2	0.0038	0.8408
Dataset 3	0.0028	0.8856
Pre-training and fine tuning		
Dataset 3	0.0027	0.8916

Figure 6 shows the logarithmic magnitude spectrum of the noise-free validation signal, here referred to as original signal (red). Also shown is the spectrum of the noise signal that was used for validation (blue). It can be seen that the two signals interfere over the whole frequency range. In such a case the whole frequency range has to be processed and reconstructed by the autoencoder. It is not possible to "copy" certain frequency ranges from the input signal. This can be regarded as a difficult task for the autoencoder.

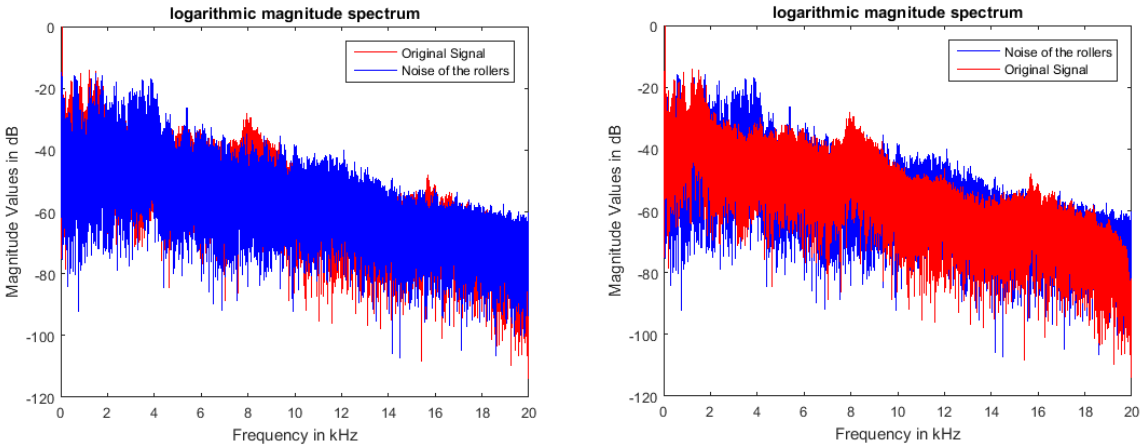

Figure 6: Logarithmic magnitude spectrum of original signal and the noise signal

In figures 7-10 the frequency spectrums of the noise-free validation signal (red) and the reconstructed signals (green) for the different datasets and training methods are shown. With all training datasets a good reconstruction up to frequencies of about 2500 Hz could be realised. In all cases the biggest differences between reference and reconstructed signals occurred in the frequency range of 4 to 10 kHz. The results shown in table 2 are also reflected by the magnitude spectrums. The trainings with dataset 3 caused the smallest deviations between reference and reconstructed signal. Compared to the direct training the pre-training

and fine tuning could additionally improve the performance of the autoencoder for a small frequency range of 4800 to 5800 Hz. The authors also listened to the reconstructed audio files. In all reconstructed files artefacts were audible. These artefacts sound similar to the crackling of old vinyl records. The reconstructed signals from the autoencoder trained with dataset 3 contained much less artefacts than the results of the autoencoder that was trained with datasets 1 and 2.

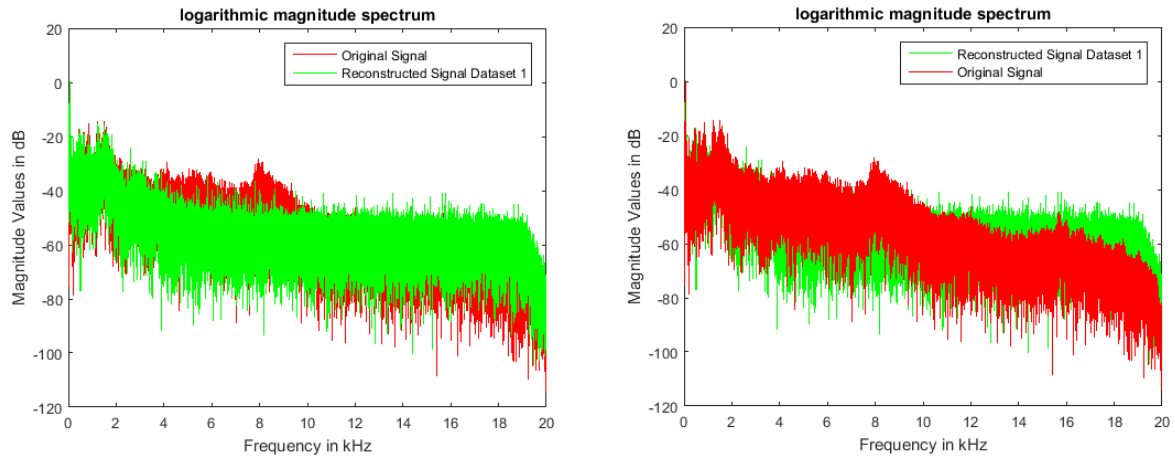

Figure 7: Comparison of the spectrums of original and reconstructed signal for dataset 1 (direct training)

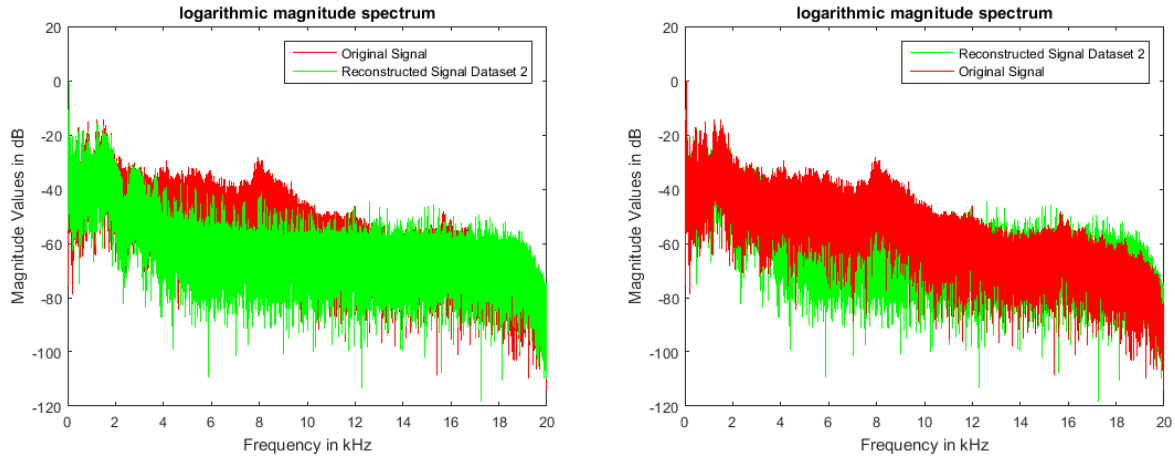

Figure 8: Comparison of the spectrums of original and reconstructed signal for dataset 2 (direct training)

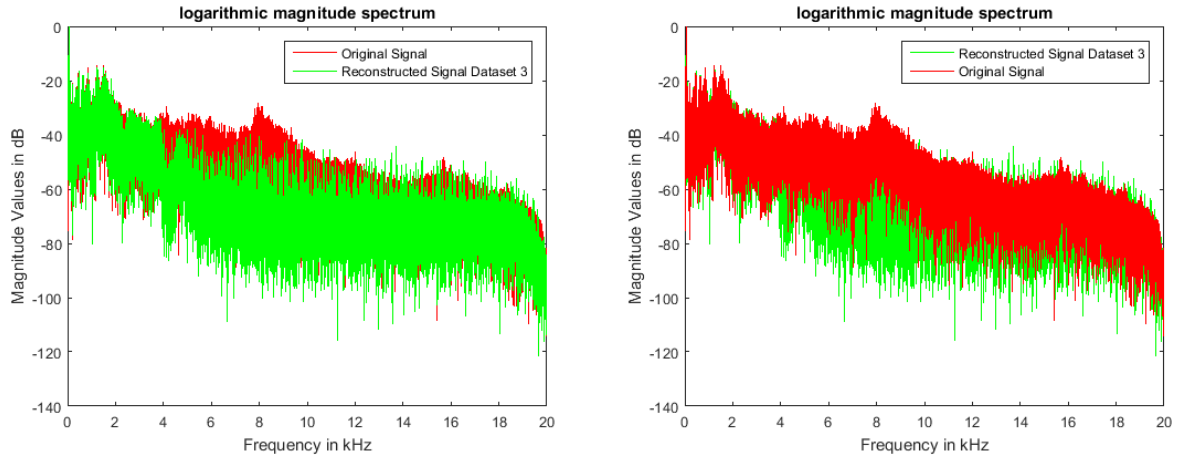

Figure 9: Comparison of the spectrums of original and reconstructed signal for dataset 3 (direct training)

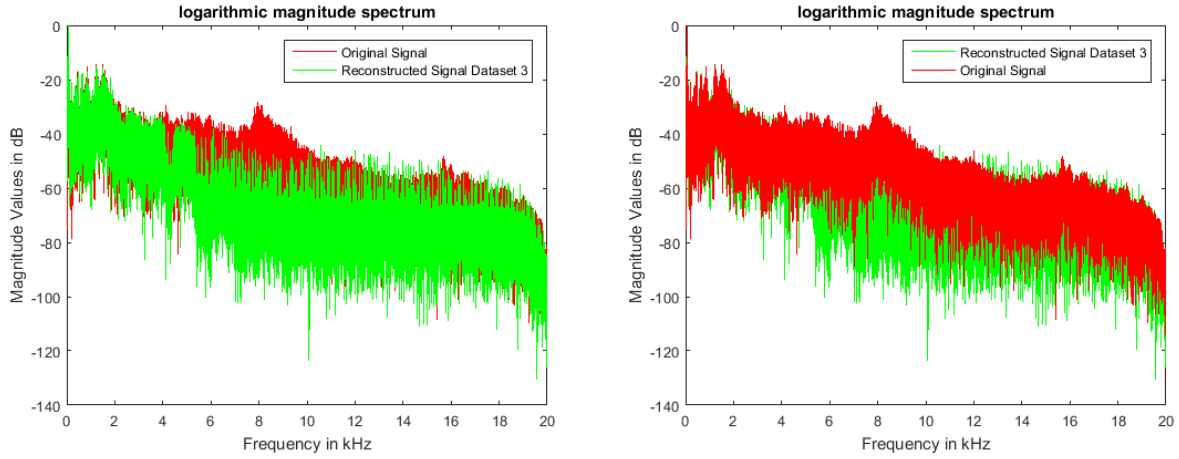

Figure 10: Comparison of the spectrums of original and reconstructed signal for dataset 3 (pre-training and fine tuning)

6. CONCLUSION

In this paper an Artificial Neural Network (ANN) is described which can be used for the post-processing of measured data. A concept specifically for the noise cancellation in acoustic measurement data is outlined. The concept was implemented with acoustic data that was measured and recorded for a database that is used for a virtual traffic simulation. The measurements were done on a roller test bench. As an unavoidable effect the rollers produce some noise which interferes with the acoustic data that is actually to be measured. An ANN was trained in order to remove the rollers noise out of the recorded audio data. The focus of this paper is on investigating different training techniques in order to find well-working solutions based on a manageable amount of training data. For this purpose three different training datasets were generated. Since the audio data is processed by the ANN in blocks audio files for the training are divided in individual parts. This division in audio blocks is varied for each dataset. A division of the training data in overlapping blocks (dataset 3) could improve the results of the training significantly. After the training with dataset 3 the original signal could be well reconstructed for most of the frequency spectrum. Differences between reconstructed and original signal could be detected mainly for mid-range frequencies. In comparison to the direct training the pre-training and fine tuning of the autoencoder led only to small improvements. The investigations outlined in this paper confirmed that the performance of an ANN, especially the performance of autoencoders, is highly correlated with the training. Meaningful and comprehensive datasets are important to generate good and

precise results. This contribution showed that well working training datasets can also be generated out of a small amount of data. The investigated datasets are based on audio recordings of a total length of solely less than 80 seconds. Thus, the results should also be interpreted in regard to the effort-to-benefit ratio. With very small effort a good approximation to a noise-free signal could be reached. This is very promising for further research.

ACKNOWLEDGEMENTS
The authors would like to thank the members of the Zeidler-Forschungs-Stiftung for their support.

REFERENCES

[1] A. Siegel, C. Weber, A. Albers, D. Landes and M. Behrendt, *Akustische Simulation von Fahrzeuggeräuschen innerhalb virtueller Umgebungen basierend auf künstlichen neuronalen Netzen (KNN),* Wissenschafts- und Industrieforum Intelligente Technische Systeme 2017, Paderborn, 2017.

[2] A. Siegel, C. Weber, A. Albers, D. Landes and M. Behrendt, *SIMULATION OF ACOUSTIC PRODUCT PROPERTIES IN VIRTUAL ENVIRONMENTS BASED ON ARTIFICIAL NEURAL NETWORKS (ANN),* ICED17: 21st International Conference on Engineering Design, University of British Columbia, Vancouver, Canada, to be published

[3] S. Husung, A. Siegel and C. Weber, *Acoustical Investigations in Virtual Environments for a Car Passing Application,* ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Buffalo, USA, 2014.

[4] A. Siegel, C. Weber, A. Mahboob, A. Albers, D. Landes and M. Behrendt, *Virtual Acoustic Model for the Simulation of Passing Vehicle Noise,* ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference IDETC2016, Charlotte, USA, 2016.

[5] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio and P. Manzagol, *Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion,* Journal of Machine Learning Research 11 (2010) 3371-3408, 2010.

[6] A. Maas, Q. Le, T. O'Neil, O. Vinyals, P. Ngyen and A. Ng, *Recurrent Neural Networks for Noise Reduction in Robust ASR,* INTERSPEECH, pp. 22-25, 2012.

[7] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1,* Philadelphia: Linguistic Data Consortium, 1993.

[8] *2-4-10038.mp3, 1999,* copyright - www.hoerspielbox.de; http://www.hoerspielbox.de, last access: 2015-11-15.

CONTACTS

Univ.-Prof. Dr.-Ing. C. Weber christian.weber@tu-ilmenau.de
Dipl.-Ing. A.Siegel antje.siegel@tu-ilmenau.de
Prof. Dr.-Ing. Dr. h.c. A. Albers albert.albers@kit.edu
Dipl.-Ing. D. Landes david.landes@kit.edu
Dr.-Ing. M. Behrendt matthias.behrendt@kit.edu