Neural Networks for Localizing Sound Sources

Magnus Schäfer, Lukas Stich
HEAD acoustics GmbH, 52134 Herzogenrath, Deutschland, Email: telecom@head-acoustics.de

Abstract
There are several approaches for acoustically localizing sound sources. Typically, they can be classified depending on the choice of microphone setup: Arrays with multiple microphones and the corresponding beamforming algorithms are often utilized in communication devices (e.g. conference phones or smart speakers) or for technical investigations. Artificial heads in conjunction with binaural hearing models usually aim at modeling the capabilities of human listeners.

This contribution utilizes an eight-channel microphone array that is mounted on an artificial head and spatially samples the direct vicinity of both ears. This arrangement allows for a localization approach that closely resembles the experience of a human listener while simultaneously avoiding the limitations of an artificial head (e.g., with respect to front-back confusions).

The microphone signals are converted to the frequency domain by a short-time Fourier transform and the phase information is used by a convolutional neural network to perform the localization task. The structure of the neural network was adapted to the geometrical setup of the microphone array. A performance assessment of the localization system is presented that is based on real audio recordings and a comparison with two conventional beamforming algorithms is shown.

Introduction
There are different approaches to sound source localization depending on the task at hand. In a more technical application, one might consider the task of localizing the origin of an unwanted mechanical noise in, e.g., a household appliance. In a communication environment, the task could be the localization of a target speaker to enhance the signal quality by, e.g., using a beamforming algorithm to separate the target signal from the background noise. For both use cases, there are numerous approaches available that differ, e.g., depending on the circumstances, the sensors or sensor arrays used and the algorithmic solutions. Two methods that are also commonly used for performance comparisons are MUSIC [1] and SRP-PHAT [2].

There is a limited amount of work in the area of source localization utilizing machine learning approaches. A method for localizing speakers with a convolutional neural network is described in [3]. The method is designed under the assumption of W-disjoint orthogonality which was shown to be approximately valid for speech signals in [4] but can not be guaranteed for all types of signals. In [5], an efficient method for generating large simulated datasets for training, validation and test of machine learning localization systems is presented.

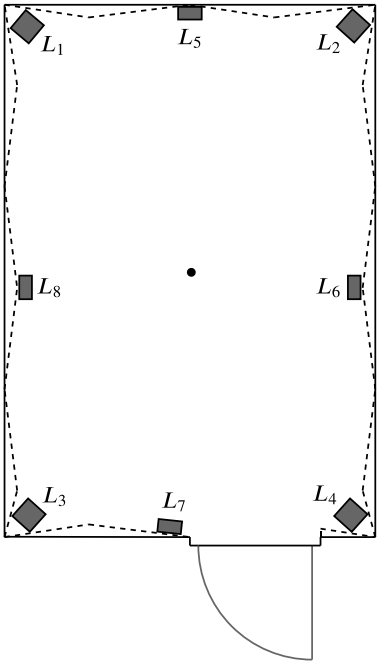

Figure 1: Overview of the measurement chamber with positions of the eight loudspeakers

Neural Networks
In recent years, advances in processing hardware along with algorithmic refinements have lead to an extensive deployment of machine learning models in different areas of signal processing. Most of these models are neural networks [6] with artificial neurons as their elementary building blocks.

The success of any machine learning approach relies on the data that is available for training, validating and testing the model. The training data set should be large enough so that overfitting of the model can be avoided. The data sets have to be disjoint to ensure a fair assessment of the capabilities of the model while at the same time each of them should be diverse enough that all steps of the model building can be performed on meaningful data.

Data Collection
The target of the present investigation is to lay the foundation for a localization approach that is capable of modeling human spatial perception by technical means. This comprises not only considering the influence of head and torso on the sound field but also analyzing the same spatial areas of the sound field that could be utilized by a human listener. To achieve this goal, a head-mounted microphone array is used to record the signals to be analyzed. The microphone array follows the design described in [7].

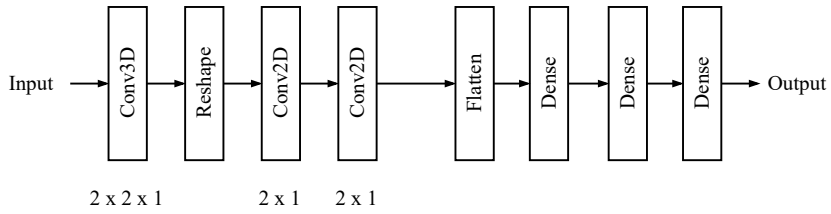

Figure 2: Layer structure of the neural network.

The microphone array has eight microphones arranged in two groups of four on each side of the head. The distribution of the microphones covers the typical range of small head movements that are used by human listeners to resolve ambiguities (cf. [8]). These ambiguities are also present in binaural recordings which essentially represent listening to an acoustic scene while keeping the head perfectly static.

Training, validating and testing a deep neural network for any task requires a fairly large amount of data. Recordings were made in a measurement chamber, its dimensions is given in Table 1 and its reverberation time is very low at 65.9 ms.

Length:	3.40 m
Width:	2.40 m
Height:	2.02 m

Table 1: Dimensions of the measurement chamber

There are eight loudspeakers mounted to the walls of the chamber, four are in the corners of the room, the remaining four are on the walls. Due to the construction of the chamber, which has differently slotted elements at different angles inside, the loudspeakers can not be mounted exactly in the center of the walls.

The recording setup consisted of the microphone array on an artificial head which was positioned on a turntable in the center of the room. The turntable was rotated in steps of 5° leading to 576 recording positions (8 loudspeakers and 72 different rotations) in total.

Two types of signals were recorded: White noise and speech. The noise signals are used to train the neural network. The speech signals are the English and German sentences from [9]. The German sentences are used for validation, the English sentences for testing and for comparison with the conventional approaches from [1, 2].

The recordings are not directly fed into the neural network. The signals are segmented and transferred to the frequency domain by a short-term Fourier transform. The feature that is analyzed by the neural network is the phase information as it showed the best performance in preliminary tests. The data is collected in a four-dimensional tensor. The dimensions are frame index of the transformation, frequency bins of the Fourier transform and two dimensions for the microphones (2x4) – roughly approximating the geometrical structure of the microphone array.

Neural Network Design

The structure of the neural network with its layers is shown in Figure 2. The different layers serve different purposes which are briefly explained and motivated in the following.

The localization is carried out individually for each time frame. Accordingly, the input data to the network is a three-dimensional tensor of dimensions 2x4x1025 (the transform length of the Fourier transform is set to 2048 samples). The first stage of the network performs convolutive operations on the data with filter kernels of different dimensions (corresponding to the dimensions of the data): After the first three-dimensional convolutional layer, the data is reshaped into a two-dimensional tensor which is subject to two sequential two-dimensional convolutional layers. These layers are typically intended to extract information and present it in a meaningful and concise manner for the following layers.

The defining parameter of convolutional layer besides the size of the filter kernel is the number of filter kernels. For the three layers in this network, the layers apply 256, 256 and 64 filter kernels going from left to right.

The following flatten layer reduces the dimensionality of the data to one. The dropout layer is a method to reduce the risk of overfitting the parameters of the network to the training data by randomly removing connections in the training phase of the network. The dropout rate is 75 %. The final three layers are dense (or fully-connected) layers which combine the information that was procured by the convolutive layers. The first two of these layers have 128 neurons, the last layer has 72.

The output of the network is the localization result. The localization task is interpreted as a classification in this system. The azimuth angles are split up into 72 equiangular classes leading to a resolution of 5°.

The activation function for all layers but the last is the exponential linear unit (ELU) [10], the last layer utilizes the Softmax function [11] to provide output data that can be interpreted as probabilities for the 72 possible source directions.

The network parameters are initialized by the He-Uniform approach [11, 12] and the categorical cross entropy [6] is the loss function. The adaptive moment estimation (Adam) [13] optimizer determines the parameters of the network, its stepsize and decay rates are set to $\alpha = 0.001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively.

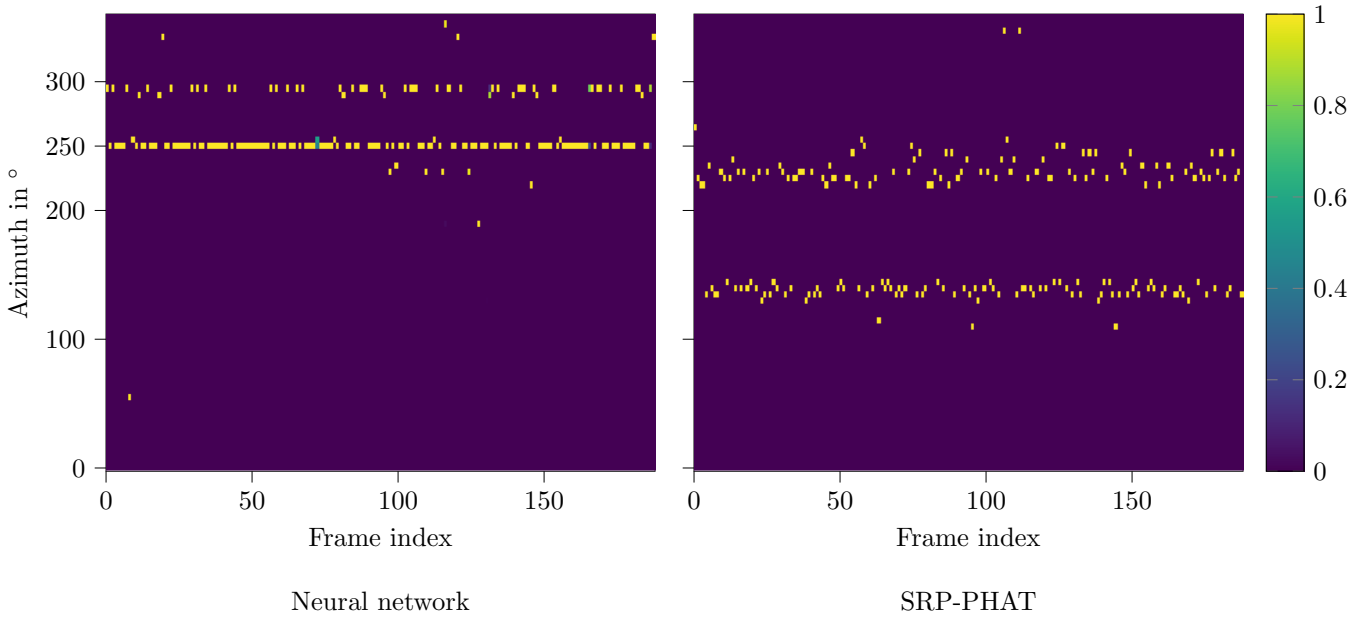

Figure 3: Probability map for the localization of a speech signal emitted by a loudspeaker at 250° in the additional acoustic environment

Experimental Results

The experimental results for the test signals are given in Table 2. The two result values are:

- Mean absolute error (MAE) between the actual source position and the estimate
- Accuracy of the localization (readily defined for the neural network; for the conventional approaches, the estimates are quantized to the nearest class center of an output class of the neural network)

Method	MAE	Accuracy
SRP-PHAT	9.30°	54.30 %
MUSIC	16.88°	43.39 %
Neural network	18.60°	66.50 %

Table 2: Experimental results for the test signals (English speech from [9])

It can be observed that the conventional approaches reach lower MAEs than the neural network. SRP-PHAT in particular has only half the estimation error of the neural network. The picture changes when looking at the accuracies. The neural network outperforms the conventional approaches by 12 % and 23 %, respectively.

A fairly low accuracy coupled with a low MAE (as can be seen for SRP-PHAT) essentially means that the localization is quite consistently close to the correct value but rarely hits it exactly. In contrast, the neural network often localizes the source perfectly but when it does not, the estimate is sometimes very far from the correct result.

This behaviour is very consistent with the loss function that was used when training the network. The categorical cross entropy only discriminates between two cases: The estimate is correct or it is wrong. There is no information in the loss function about how far from the correct value the current estimate is – the result is identical when the estimate is off by 5° or by 180°.

Additional Acoustic Environment

It is interesting to see that the neural network is capable of localizing signals which were not part of the training data set. Another interesting parameter is the acoustic environment: As described before, the data for training the network was gathered in a controlled environment in a measurement chamber with very little reverberation and no strongly reflective surfaces.

Some additional speech signals were also recorded inside a car cabin with the same microphone array on an artificial head. There are 15 loudspeakers in the car that are mounted at different azimuth angles (in the dashboard, the doors, the pillars and the boot). The data set is obviously fairly small compared to the set from the measurement chamber that had 576 different source directions so a full statistical investigation is not that meaningful. There is a basic trend that all methods have higher mean average errors compared to the measurement chamber but that the losses are larger for the neural network.

Some interesting observations can already be made by looking at examples from the localization results. One example is shown in Figure 3, the result for the neural network is on the left, the result for SRP-PHAT, the better of the two model-based approaches, is on the right. The loudspeaker is at an angle of 250° in relation to the head. Accordingly, a perfect localization result would be represented by a horizontal yellow line at that angle.

It can be observed that neither localization system is perfect in this situation. The neural network mostly localizes the loudspeaker correctly but also identifies an-

other source at 290° in some time frames. The result for SRP-PHAT also shows two sources, one is almost in the correct direction at 225° while the other source is at approximately 140°. An interesting aspect of the result is the different distribution of estimates: The neural network has two distinct horizontal lines at the two aforementioned angles with very few instances of high probability in the direct vicinity of the lines. The result for SRP-PHAT, on the other hand, has two groups of points that are spread in two angular regions. This is consistent with the observation that is also made for the data from the measurement chamber: Large deviations from the mean estimate are rare for the model-based approaches but small deviations are very common.

Conclusions and Outlook

A machine learning system for localizing sound sources in the azimuth plane was presented in this contribution. The system comprises several convolutional and fully-connected layers. The localization task is interpreted as a classification. A performance comparison between the neural network and two conventional model-based approaches revealed that the neural network achieves a high accuracy, i.e., it perfectly localizes the source. The mean absolute error of the neural network is higher than the error for the model-based approaches, though.

Another acoustical environment, a car cabin, was used to experimentally test if a neural network that was trained in one specific acoustic environment can be used in another environment. While the neural network outperformed the model-based approaches for some situations, it lost more performance than the other approaches on average. This again underlines the need for large training data sets that are representative of the later use cases for the neural network.

One aspect that was observed for the neural network in its current implementation is the impact of the chosen loss function. While the neural network often estimates the source position correctly, the false estimates are usually not near to the correct position leading to a high mean absolute error for the estimation. This could be improved by modifying the loss function to incorporate some measure of closeness to the target or by modeling the localization task as a regression altogether.

Acknowledgment

The study presented here was carried out within the framework of the research project *ALFASY – Altersgerechte Fahrerassistenzsysteme* (*Age-based Driver Assistance Systems*). This project was funded by the European Regional Development Fund (ERDF).

References

[1] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, March 1986.

[2] José Velasco, Daniel Pizarro, and Javier Macias-Guarasa. Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints. *Sensors (Basel, Switzerland)*, 12:13781–812, 12 2012.

[3] Soumitro Chakrabarty and Emanuël AP Habets. Multi-speaker localization using convolutional neural network trained with noise. *arXiv preprint arXiv:1712.04276*, 2017.

[4] Scott Rickard and Özgür Yilmaz. On the approximate W-disjoint orthogonality of speech. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–529–I–532. IEEE, 2002.

[5] Hadrien Pujol, Eric Bavu, and Alexandre Garcia. Source localization in reverberant rooms using deep learning and microphone arrays. In *International Congress on Acoustics (ICA)*, Aachen, Germany, 2019.

[6] Michael A Nielsen. *Neural networks and deep learning*. Determination press San Francisco, CA, USA:, 2015.

[7] Magnus Schäfer, Benedikt Koppers, Jan Reimes, and Hans-Wilhelm Gierlich. Joint reproduction of background noise and reverberation for development and testing of binaural devices. In *Audio Engineering Society Conference: 2019 AES INTERNATIONAL CONFERENCE ON HEADPHONE TECHNOLOGY*, Aug 2019.

[8] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[9] ETSI TS 103 281 V1.2.1. *Speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals*, January 2018.

[10] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[11] Aurlien Gron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 1st edition, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1026–1034, USA, 2015. IEEE Computer Society.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.