On sound source localization of speech signals using deep neural networks

Reinhild Roden[1,2], Niko Moritz[1], Stephan Gerlach[1], Stefan Weinzierl[2], Stefan Goetze[1]
[1] *Project Group Hearing, Speech and Audio Technology, Fraunhofer IDMT*
[2] *Audio Communication Group, TU Berlin*
E-mail: reinhild.roden@idmt.fraunhofer.de

Abstract

In recent years artificial neural networks are successfully applied especially in the context of automatic speech recognition. As information processing systems, neural networks are trained by, e.g., backpropagation or restricted Boltzmann machines to classify patterns at the input of the system. The current work presents the implementation of a deep neural network (DNN) architecture for acoustic source localization.

Introduction

Beside the detection of the 2-dimensional angle of sound incidence, a reliable 3D-sound-source localization can improve the performance of hearing aids and spatial filtering algorithms (beamformers), for instance. The distance is usually more difficult to estimate than the direction of arrival using auditory motivated localization cues, i.e., interaural level difference (ILD), interaural phase difference (IPD), and interaural time delay (ITD), cf. e.g. [1] [2] [3]. Auditory depth perception is rather imprecise, cf. e.g. [4], pp 116-137. In addition, most studies are limited to detect the direction of sound incidence, often just for one dimension - the azimuth. We propose to use a deep neural network (DNN) using multiple hidden layers to estimate the position of a sound source in all three spatial dimensions, azimuth ϑ , elevation φ and distance r. In this contribution, different features and feature combinations are evaluated as an input of the DNN. Tested feature types are ILD, ITD, binaural spectral magnitudes and phases, as well as real and imaginary parts of the signal spectrum.

Signal Processing

Deep Neural Network

A DNN consists of many neurons that are organized in multiple layers. A feedforward network connects each neuron in one layer with each neuron of the next layer in a single direction. Forward passing the network as depicted in Fig. 1, for each layer the single set of input values (out of a data set) $\boldsymbol{\theta^{in}}$ is weighted and summed up according to Eq. (1). Following this, the result θ^{net} is forwarded to the activation function, here chosen to be the sigmoid logistic function in Eq. (2). The parameters I, H, O in Fig. 1 denote the number of neurons in the respective layer, vector $\boldsymbol{w}$ defines the weights of the connections between neurons of neighboring layers and b is the bias weight of a layer. All output values θ^{out} of previous layer are either the set of input values for the next one or the output of the network.

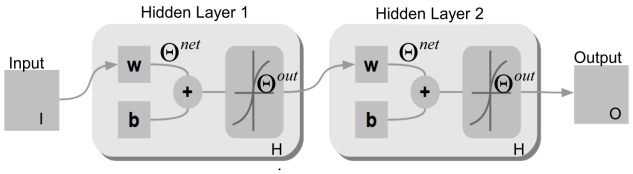


Fig. 1: Path of neural network with two hidden layers.

$$\theta^{net} = \sum_{a=1}^{A} (w_a \cdot \theta_a^{in}) + b \tag{1}$$

$A = I$ computing the output values of the first hidden layer, otherwise $A = H$.

$$\theta^{out} = \frac{1}{1 + exp(-\theta^{net})} \tag{2}$$

Eq. (1-2) calculated H respectively O times for each neuron in the layer.

The training of the DNN is realized as follows: Firstly, the weights are randomly initialized with values between $w = \pm 0.1$. Secondly, for one input vector the forward passing is done as described above for Fig. 1. Next, the network output is compared with a target vector by computing the error as absolute difference which is propagated back to the DNN input. The used method is known as error backpropagation by stochastic gradient descent minimizing the error as a function of the weights. Due to the minimization the last steps are the calculation of weight adjustments (with momentum) and the update of weights. The procedure is repeated for the epoch length[1] multiplied by a given number of epochs[2].

Feature extraction from binaural signal

Feature vectors containing spatial information and feeding the DNN are calculated using a binaural time signal $y_{\{l,r\}}[n]$ Eq. (3). The left and right channel is de-

[1] The epoch length is the number of input vectors in the data set feeding the DNN.

[2] The number of epochs is a rational number. Without a replacement of an input vector during the training and according to a number of epochs of one, each vector is given to the DNN for one time, for instance.

noted by the subscripted letters l and r; n defines the time index.

$$y_{\{l,r\}}[n] = [y_{\{l,r\}}[n], y_{\{l,r\}}[n-1], ...y_{\{l,r\}}[n-N+1]]^T \quad (3)$$

ILD and ITD are extracted as follows (f_s - sampling frequency):

$$\mathrm{ILD} = 20\log_{10}\left(\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} y_l^2[i]}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} y_r^2[i]}}\right) \ [\mathrm{dB}] \quad (4)$$

$$\mathrm{ITD} = \max(\underbrace{\sum_{m=-\infty}^{\infty} y_l^*[m]y_r[n+m]}_{\text{cross correlation}})/f_s \cdot 10^6 \ [\mu\mathrm{s}] \quad (5)$$

In detail ILD and ITD are applied to 30 binaural time signals coming from a gammatone filter bank in a frequency range of 0.1 to 8 kHz [5]. Therefore the length of the input vector amounts to 30 for separate used features and 60 for combined features. Further audio representations are the binaural magnitude $|\mathcal{F}(\mathrm{y}_{\{l,r\}})|$, the phase $\arg(\mathcal{F}(\mathrm{y}_{\{l,r\}}))$, the real $\mathrm{Re}(\mathcal{F}(\mathrm{y}_{\{l,r\}}))$ and the imaginary part $\mathrm{Im}(\mathcal{F}(\mathrm{y}_{\{l,r\}}))$ computed using the short-term Fourier transformation (STFT) of the time signal as well as combinations of these features. Assuming a STFT-length of 512 samples at a sampling rate of $f_s = 16$ kHz, the half of the spectrum includes 257 bins for the left or right channel of the signal. For example, combining magnitude and phase, the number of input neurons is long as four times the number of bins representing the half spectrum. Tab. 1 shows possible input vectors to the DNN and their specific lengths.

Tab. 1: Input vectors to the DNN and its length.

feature	length
ILD	30
ITD	30
combined ILD, ITD	30+30
$\lvert\mathcal{F}(\mathrm{y}_{\{l,r\}})\rvert$	257+257
combined $\mathrm{Re}(\mathcal{F}(\mathrm{y}_{\{l,r\}}))$, $\mathrm{Im}(\mathcal{F}(\mathrm{y}_{\{l,r\}}))$	2·(257+257)
combined $\lvert\mathcal{F}(\mathrm{y}_{\{l,r\}})\rvert$, $\arg(\mathcal{F}(\mathrm{y}_{\{l,r\}}))$	2·(257+257)

Preliminary experiments

Spatial speech signals

The binaural speech signals are obtained by convolution of head related impulse responses (HRIRs) [6] with logatomes of the Oldenburg logatome corpus [7]. Used logatomes were spoken by a male speaker without dialect. The sampling rate is 16 kHz. Silent parts of the signals are removed. HRIRs were measured using a KEMAR mannequin by Thiemann et al. [6]. Seven directions were chosen for the azimuth ($\vartheta \in \{-30°, -20°, \dots 30°\}$) and the elevation ($\varphi \in \{-10°, 0°, \dots\dots 50°\}$). The chosen angles are visualized in Fig. 2.

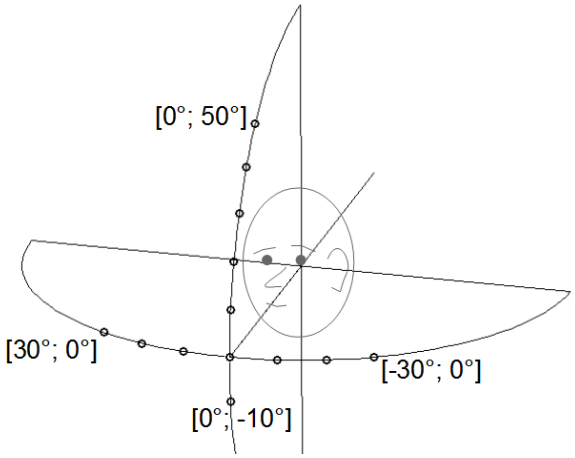


Fig. 2: Grid of used HRIRs, $[\vartheta; \varphi]$

Due to the lack of measurements for different distances just one feature for the distance is examined. For five distances ($r \in \{10, 20 \dots 50\}$m) the atmospheric absorption is calculated based on ISO 9613-1 including formulae for the pure tone attenuation coefficient [dB/m] (cf. [8] Eq. (3-5)). This is not considered for speech because the frequency range of the logatomes is limited to 8 kHz. Instead the attenuation of sound is calculated for monaural pink noise at a sampling rate of f_s =44.1 kHz in order to examine the sensitivity of the network for this feature also for higher frequency ranges.

Comparison of input vectors for independent spatial dimensions

For the evaluation of different input vectors parameters of the network are chosen to be constant for all cases: Two hidden layers are trained, whereby the learning rate is set to a value of 0.1 and the momentum to 0.3. In each case a set of input vectors calculated from 50 logatomes are prepared for training and for testing purposes. Features were calculated block-wise with a frame size of 25 ms, a hop size of 10 ms and a STFT-length of 512 samples at a sampling rate of f_s =16 kHz. The runs of training and testing for certain kinds of input vectors are conducted independently for each spatial dimension.
As mentioned logatomes are used just for the horizontal and median plane. Examining the atmospheric absorption for different distances the network is trained and tested with the magnitude extracted from monaural pink noise. Due to the changed sampling rate (f_s =44.1 kHz) the STFT-length is 2048 samples. Used input neurons are all spectral values representing the frequency range between 18 and 22.05 kHz.
Tab. 2 lists considered conditions whose results are shown in this work. Seven cases of different kinds of input vectors are examined for the horizontal and for the median plane. One case is related to the distance. All cases are numbered by Roman numerals. Furthermore, Tab. 2 includes the epoch length and the number of epochs as well as the number of neurons I, H, O in the layers. The epoch length differs occasionally for the reason that sometimes

Tab. 2: First columns list tested cases and respective parameters. The last column contains the average hit rates of overall performances (guessing rate: 14.3 for horizontal/median plane, 20% for distance).

	Dim	**Kind of input**	**Epoch length**	**Trained epochs**	**I**	**H**	**O**	**average hit rate** [%]
I	$[\vartheta;0°]$	ILD	19915	30.1	30	100	7	99.8
II	$[\vartheta;0°]$	ITD	20118	29.8	30	100	7	99.1
III	$[\vartheta;0°]$	ILD, ITD	18959	31.6	60	100	7	98.8
IV	$[\vartheta;0°]$	$\lvert\mathcal{F}(\mathrm{y}_{\{l,r\}})\rvert$	19400	30.9	514	600	7	100
V	$[\vartheta;0°]$	$\mathrm{Re}(\mathcal{F}(\mathrm{y}_{\{l,r\}}))$, $\mathrm{Im}(\mathcal{F}(\mathrm{y}_{\{l,r\}}))$	17633	34.0	1028	1100	7	100
VI	$[\vartheta;0°]$	$\lvert\mathcal{F}(\mathrm{y}_{\{l,r\}})\rvert$, $\arg(\mathcal{F}(\mathrm{y}_{\{l,r\}}))$	20209	29.7	1028	1100	7	14.3
VII	$[0°;\varphi]$	ILD	19330	31.0	30	100	7	100
VIII	$[0°;\varphi]$	ITD	19351	31.0	30	100	7	37.1
IX	$[0°;\varphi]$	ILD, ITD	19848	30.2	60	100	7	97.2
X	$[0°;\varphi]$	$\lvert\mathcal{F}(\mathrm{y}_{\{l,r\}})\rvert$	18788	31.9	514	600	7	94.2
XI	$[0°;\varphi]$	$\mathrm{Re}(\mathcal{F}(\mathrm{y}_{\{l,r\}}))$, $\mathrm{Im}(\mathcal{F}(\mathrm{y}_{\{l,r\}}))$	18315	32.8	1028	1100	7	23.5
XII	$[0°;\varphi]$	$\lvert\mathcal{F}(\mathrm{y}_{\{l,r\}})\rvert$, $\arg(\mathcal{F}(\mathrm{y}_{\{l,r\}}))$	17790	33.7	1028	1100	7	14.3
XIII	r	$\lvert\mathcal{F}(\mathrm{noise})\rvert$	12000	50	190	200	5	71

more, sometimes less input vectors are generated for 50 randomly chosen logatomes of different signal lengths. The varying (rounded) numbers of epochs arise from a constant number of training iterations which were set to $(6 \cdot 10^5)$.

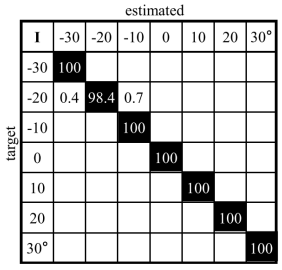
Results and Discussion

The last column of Tab. 2 presents achieved average hit rates and a summary of results. Moreover, each single case is shown as a confusion matrix in Fig. 3. The captions include the Roman numerals as seen in Tab. 2 assigning a certain panel with the related feature and spatial dimension.
In the horizontal plane the hit rates are high as expected for ILD and ITD and their combination (cf. Fig. 3, panels a, b). A perfect confusion matrix could be reached with longer time frames for ILD and ITD calculation or a time recursive smoothing of these values. The combination of magnitude and phase could not reveal the same result (cf. Fig. 3, panel c). Perhaps another set of network parameters or a different representation of the magnitude and phase information could change this performance. In contrast, the binaural magnitude and the combination of the real and imaginary part show maximal hit rate (cf. Fig. 3, panel b). Hence, the DNN probably extracts the spatial information by itself. This fact is promising for the median plane.
In the median plane the DNN is expected to extract the monaural spectral coloration that is induced by changing the elevation. Considering rather high precision for the binaural magnitude it could be assumed that the DNN learned to extract the mentioned coloration (cf. Fig. 3, panel e). Since the ILD and the combination of ILD and ITD show high accuracies in median plane (cf. Fig. 3, panels d, e), it is possible that the DNN has detected a cue in the narrowband ILD features, which is not characterized by the actual ILD itself. The reason is that the ILD should remain constant for all elevation angels, i.e., was expected to be insensitive to the median plane. Also the performance of the ITD only is above guessing rate (cf. Fig. 3, panel f). Conceivably, a small but systematic deviation of the azimuth ϑ at each elevation is the reason for changing characteristics of the ILD and ITD (approx. 40 μs distributed over the whole range of all used elevation angles). This could be caused by a minimal and nearly imperceptible unbalance of the measurement setup for the HRIR data set, respectively a slight inclination od the median plane. A different explanation for the precise localization using the ILD in the median plane focus the directivity of recording microphones while the HRIR measurement. Without an omni-directional characteristic deviations in the ILD are plausible if the microphone membrane is changed in its orientation while measuring the reference signal[3] (cf. [9]). In this case the deconvolution to extract the HRIR cannot eliminate the influence of directivity which supports the localization performance of the DNN, here.
For localizing the distance, the atmospheric absorption cannot be considered for speech but potentially for broadband acoustic events. The DNN is able to detect distances by the attenuation of sound within the frequency range of 18 – 22.05 kHz (cf. Fig. 3, panel i). Lower frequency ranges were tested but could just reach results near the guessing rate. With decreasing frequency the training becomes more difficult because attenuation coefficients become smaller which leads to a blurred pattern for the DNN.

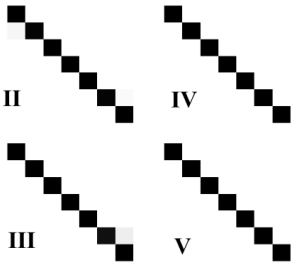
Conclusion, Outlook

To conclude, the localization of speech signals by a DNN is possible for the horizontal and median plane. In the horizontal plane the use of ILD and ITD features should be preferred for the smaller size of the input vectors supporting an efficient training. For the median plane, spectral features that contain additional monaural information are theoretically expected to perform better. A directivity of recording microphones could support the localization performance using the ILD.
In future, experiments with a set of measured HRIRs for

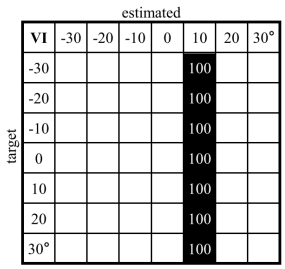
[3] A reference signal without KEMAR mannequin is measured for a deconvolution of the recorded signal in order to extract the pure HRIR. Hereby, the excitation signal and fault effects, like reflections or the influence of the speaker, are eliminated.



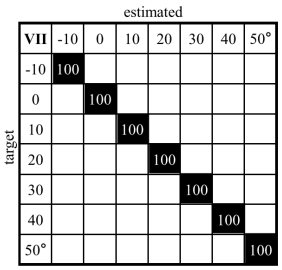
(a) case I



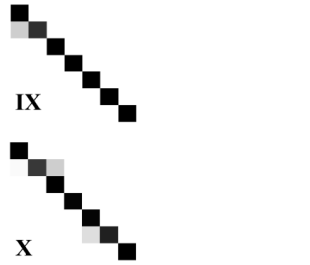
(b) almost identical looking plots for cases I - V



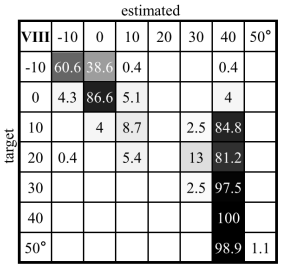
(c) case VI



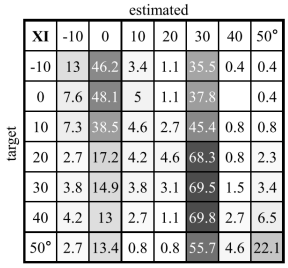
(d) case VII



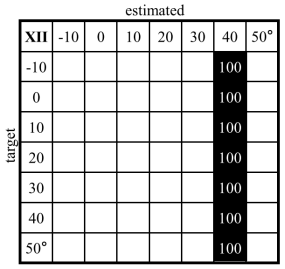
(e) almost identical looking plots for cases VII, IX, X



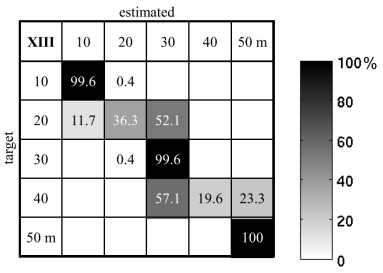
(f) case VIII



(g) case XI



(h) case XII



(i) case XIII

Fig. 3: DNN performance of the testing. The average of 100 ms (10 frames) is rated as one decision of the network.

different distances and smaller ranges are planned. Long-term, it is intended to implement a DNN architecture for a spherical localization and a parallel estimation of the source distance. Attention should be paid to the fact that cues for azimuth and elevation are not independent [10]. The system will be evaluated for robustness against changes in room acoustic conditions and additional diffuse noise.

Acknowledgments

This work was partially funded by the European Commission. Grant no. 318381 EAR-IT – Experimenting Acoustics in Real environments using Innovative Testbeds, S4ECoB - Sounds for Energy Control of Buildings under grant no. 284628 and EcoShopping - Energy efficient & Cost competitive retrofitting solutions for shopping buildings grant no. 609180.

References

[1] Michael S. Datum, Francesco Palmieri, and Andrew Moise. An artificial neural network for sound localization using binaural cues. The Journal of the Acoustical Society of America, 100(1):372-383, 1996

[2] Andrzej Czyzewski. Automatic identification of sound source position employing neural networks and rough sets. Pattern Recognition Letters, 24(6):921-933, 2003.

[3] Tobias May, Steven van de Par, and Armin Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. Audio, Speech, and Language Processing, IEEE Transactions on, 19(1):1-13, 2011.

[4] Jens Blauert. Spatial hearing. Cambridge, MIT Press, 1997.

[5] Malcolm Slaney. An efficient implementation of the patterson-holdsworth auditory filter bank. Apple Computer, Perception Group, Technical Report, 1993.

[6] Joachim Thiemann, Andreas Escher, and Steven van de Par. Multiple model high-spatial resolution HRTF measurements. In DAGA 2015, Nürnberg, page 264, 2015.

[7] medi.uni-oldenburg.de/ollo/ 15.03.2015, 10:00

[8] ISO 9613-1 (1993). Acoustics – Attenuation of sound during propagation outdoors – Part 1: Calculation of the absorption of sound by the atmosphere. Geneva, Switzerland: International Organisation for Standardization (ISO).

[9] Møller, H. (1992). Fundamentals of binaural technology. Applied acoustics, 36(3), 171-218.

[10] Chalapathy Neti, Eric D. Young, and Michael H. Schneider. Neural network models of sound localization based on directional filtering by the pinna. The Journal of the Acoustical Society of America, 92(6):3140-3156, 1992.