Machine listening in spatial acoustic scenes with deep networks in different microphone geometries

Jörn Anemüller* and Hendrik Schoof

Computational Audition Group, Dept. Medical Physics and Acoustics, and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany

Abstract

Multi-channel acoustic source localization evaluates direction-dependent inter-microphone differences in order to estimate the position of an acoustic source. We here investigate a deep neural network (DNN) approach to source localization that improves on previous work with learned, linear localizers. DNNs with depths between 4 and 15 layers were trained to predict the direction of target speech in an isotropic, multi-speech-source noise field. Several system parameters were varied, in particular number of microphones in the bilateral hearing aid scenario was set to 2, 4, and 6, respectively. Results show that DNNs provide a clear improvement over the linear classifier reference system. Increasing the number of microphones from 2 to 4 results in a larger increase of performance for the DNNs than for the linear system. 6 microphones provide only a small additional gain. The DNN architectures perform better with 4 microphones than the linear approach does with 6 microphones, thus indicating that location-specific information in source-interference scenarios is encoded non-linearly in the sound field.

1 Introduction

The human auditory systems routinely performs acoustic source localization, a task is important also in technical systems since it permits detection of relevant event such as speech, facilitates reconfiguration of (auditory) spatial signal processing, and may trigger subsequent actions such as obstacle avoidance in robots.

Location-specific information as measured with multi-channel microphone arrays is encoded in relative transfer functions (RTFs, [8]), dominated by, but not limited to, time-differences of arrival (TDOA) of the direct-path component of the acoustic signal. Classic approaches for source localization are based on TDOA analysis which commonly uses the generalized cross-correlation (GCC) method to yield robust TDOA estimates [2, 9].

Data adaptive systems that form an implicit RTF representation through learning on training data, have been proposed as systems that do not rely on direct TDOA estimation. In some real-world scenarios, e.g., when amplitude modulation is characteristically present in target and interference source, they have shown robust localization performance [7, 1, 11, 4].

The present work evaluates a non-linear extension of an earlier linear approach [5] by employing deep feed-forward networks that learn the transformation from multi-channel audio signals to a probabilistic location map. Specific emphasis is put on a systematic comparison across several deep network architectures and with a linear reference network that serves as baseline. We investigate the question as to what extent the density of spatial sound field sampling, i.e., number of microphone sensor channels, influences localization accuracy and whether there might be a trade-off between number of sensors and complexity of the classifiers' architecture. In conclusion, the results presented here for speech sources embedded in isotropic noise are indicative

*Corresponding author: joern.anemueller@uol.de. The authors acknowledge support by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grants FOR 1732 "Individualized Hearing Acoustics" and SFB 1330 "Hearing Acoustics", Project B3 "Hierarchical Models of Acoustic Information Processing", Projektnummer 352015383.

https://doi.org/10.7557/18.5151

of a qualitative difference between non-linear (deep network) and linear localizers that cannot be overcome by the inclusion of additional sensor channels.

2 Methods

2.1 Probabilistic Acoustic Source Localization with Deep Nets

The discriminative approach to source localization builds on a standard classification framework that is employed to build decision models for directional sound source presence. Relevant acoustic parameters are learned implicitly, thus no direct impulse response measurements and no additional assumption on the acoustics are required.

Source presence is indicated by cross-correlation function features $\rho_{ij}(\tau)$, cf. section 2.2, containing a main peak centered around the TDOA $\tau_{ij}(\zeta)$ corresponding to location ζ. The cross-correlation functions should therefore permit a classifier to adaptively learn to discriminate patterns that imply source presence from those that occur when no source is active in the direction of interest. They are denoted by

$$\boldsymbol{\phi}_{ij} = \left[\rho_{ij}\left(-D_{ij}\right), \ldots, \rho_{ij}\left(0\right), \ldots \rho_{ij}\left(D_{ij}\right)\right]^{\top}, \quad (1)$$

$$\boldsymbol{\phi} = \left\{\boldsymbol{\phi}_{ij}\right\}_{i=1,\ldots,M;\, j=i+1,\ldots,M}, \quad (2)$$

where D_{ij} is the maximum absolute delay between two sensors that is included and $\boldsymbol{\phi}$ denotes the feature vector concatenating all cross-correlation vectors from all pairs of M sensors i, j.

During classifier training, example feature vectors $\boldsymbol{\phi}$ are labeled as positive examples for their respective source direction ζ, whenever a source is present at the corresponding location during the time-frame across which the feature vector has been computed. We here employ deep feed-forward neural network classifiers in order to build implicit direction-dependent models during training. Their output layer contains a set of N output units, one for each direction ζ.

When trained with the categorical cross-entropy cost-function, network outputs converge to a-posteriori probability estimates for the respective classes. Hence, the output of a trained deep network localization algorithm provides us with a spatio-temporal probabilistic localization map

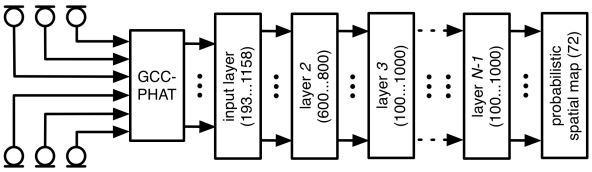

Figure 1: Processing diagram of the proposed algorithm. Multi-channel audio data from between two and six microphones is transformed into GCC-PHAT feature vectors that are used to train and evaluate three different deep network architectures and a linear reference net on the task of predicting the correct source position in 72 azimuth angle directions. The output of each network is a probabilistic spatio-temporal localization map, estimating the probability of source activity for each time-point and direction.

$\hat{P}_{source}(\zeta, t)$ that indicates the probability of a source being active for each time frame t and each direction ζ. Maximum a-posteriori estimates are computed from the probabilistic location map according to

$$\hat{\zeta}(t) = \mathrm{argmax}_{\zeta}\left[\hat{P}_{source}(\zeta, t)\right]. \quad (3)$$

Multi-source DOA estimation is achieved by evaluation of the J most probable occurrences of sound source positions. Note that estimation of the number of sound sources is not attempted here although the probabilistic information about the directional source distribution may lend itself to such an approach.

2.2 Feature Extraction

Deep-network localization is based on input features that capture the spatial covariance structure of the sound field as observed at the microphones, using generalized cross-correlation phase transform (GCC-PHAT) [10] functions

$$\rho_{ij}(\tau) = \frac{1}{\Omega}\sum_{\omega=0}^{\Omega} \Psi(\omega) \cdot X_i(\omega) X_j^*(\omega) \cdot e^{2\pi i \omega \tau / \Omega}, \quad (4)$$

$X_i(\omega)$ denotes the Fourier-transform of the i-th microphone signal $x_i(t)$, ω frequency and $\Psi(\omega) = H_i(\omega)H_j^*(\omega)$ a spectral weighting. The phase transform (PHAT) weighting has been shown to be robust against noise and reverberation, and is often used in direction-of-arrival (DOA) estimation

	#Mic	Chan. left	Chan. right	#GCC
G1	2	front-left	front-right	193
G2	4	front-left rear-left	front-right rear-right	579
G3	6	front-left center-left rear-left	front-right center-right rear-right	1158

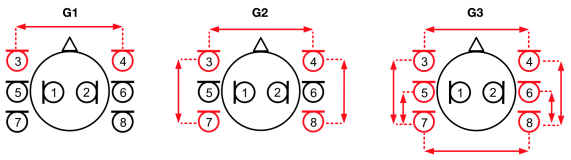

Table 1: Top: Geometries of the bilateral hearing aid setup with two to six microphones and the resulting number of GCC-PHAT coefficients that capture the pairwise spatial cross-correlation functions. Bottom: Depiction of microphones' geometric arrangement relative to the head (not to scale) and the microphones pairs for computation of cross-correlation functions (red). In-the-ear microphones No. 1 and No. 2 were not used in this study.

[7, 1, 3], resulting in the choice

$$\Psi(\omega) = \frac{1}{|X_i(\omega)X_j^*(\omega)|}. \quad (5)$$

Thus, $\Psi(\omega)$ equalizes the amplitude of the signals with a uniform spectral weighting.

		layer & size			
	layer	in	2	3...N-1	out
Net 1	4	#GCC	800	1 x 1000	72
Net 2	5	#GCC	600	2 x 600	72
Net 3	7	#GCC	600	4 x 300	72
Net 4	15	#GCC	600	12 x 100	72
Net R	2	#GCC	n/a	n/a	72

Table 2: Deep network architectures compared in the experiments. Size of the input layer (#GCC) depends on the size of the GCC-PHAT feature vector associated with the respective microphone geometry, cf. Table 1. The output layer size of 72 corresponds to number azimuth directions, number of hidden layers varied between 2 and 13. *Net R* denotes the linear reference network.

input layer:	min. 193 ... max. 1158 units
hidden layers:	min. 2 ... max. 13 layers
hidden layer size:	min. 100 ... max. 1000 units
output layer:	72 units
dropout:	50% of units per layer
activation function:	rectifying linear (ReLU)
output non-lin.:	softmax
optimizer:	adam
cost-function:	categorical-cross-entropy

Table 3: Parameters of DNN architectures used in training and evaluation, also cf. Table 2.

3 Experimental evaluation

3.1 Training and Evaluation Data

Data for training and evaluation of the proposed algorithm was generated from a database of multi-channel head-geometry room impulse response function [6] and the TIMIT speech corpus. Target speech sources were placed in 5 degree intervals at one of 72 azimuth angles at distance 80cm for which impulse responses were available. Single-channel speech signals from the TIMIT corpus were convolved with 6-channel behind-the-ear hearing aid impulse responses set to obtain multi-channel sound field data for the corresponding speaker location. Depending on the experiment, between two and six channels were used during training and testing, resulting in three different array geometries $G1$, $G2$, $G3$, as described in Table 1. Isotropic, speech-simulating noise field data was generated by placing 72 randomly selected speech sources simultaneously at all 72 azimuth positions, ensuring spectral and temporal properties of each interfering source to be (on average) identical to those of the target source. 6-channel speech- and noise-fields were superimposed at signal-to-noise-ratios (SNR) of clean (∞ dB), 20 dB, 10 dB, 0 dB, and -10 dB. In total, 10 hours of multi-channel training data were generated from the training portion of the TIMIT dataset for each SNR condition , comprising 144 unique speaker-utterance combinations (72 male, 72 female) per direction. Evaluation data amounted to 5 hours from 72 unique speaker-utterance combinations (36 male, 36 female) from the testing portion of TIMIT. Thus, the total amount of data for training and evaluation was of sufficient size to train large deep-network architectures.

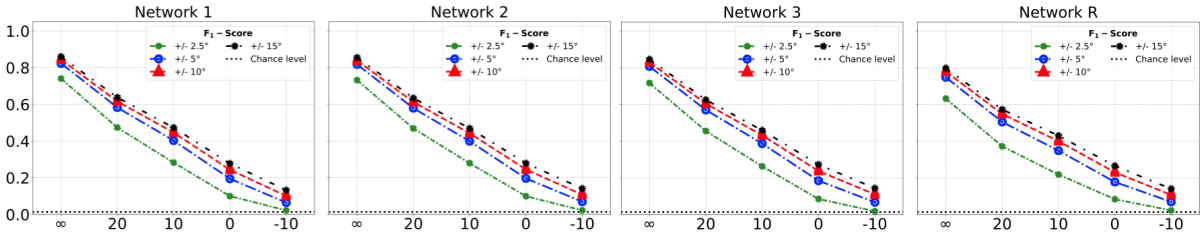

Figure 2: Performance with 2 behind-the-ear microphones (geometry *G1*) and 10 ms temporal resolution of localization. F_1-scores given for different network architectures, different azimuth resolutions and different SNR conditions.

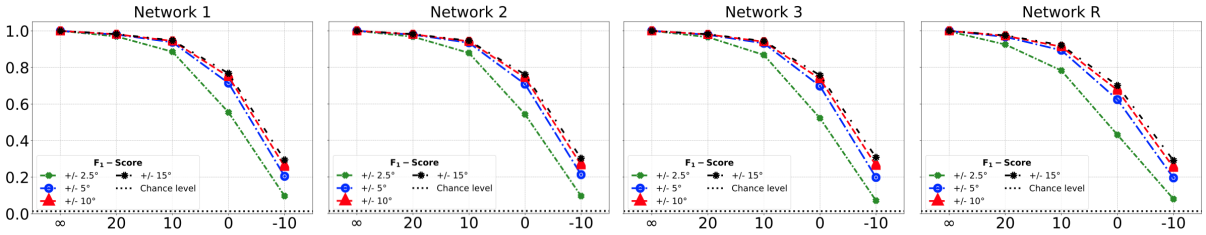

Figure 3: Performance with 6 behind-the-ear microphones (geometry *G3*) and 100 ms temporal resolution of localization. F_1-scores given for different network architectures, different azimuth resolutions and different SNR conditions.

3.2 Algorithm setup

GCC-PHAT coefficients were computed from 10 ms windows and band-limited with an upper cut-off frequency of 4 kHz. A moderate window-shift of 5 ms was chosen to generate training and test data for the evaluation setting. After reducing the length of the GCC-PHAT vectors to 4 ms around the center, limiting their maximum delay to ±2 ms, feature vectors with dimensionality between 193 and 1158 were obtained as input vectors for DNN processing. Depending on the number of microphones, we used 1, 3, or 6 pairwise cross-correlations that were subsequently arranged in a single feature vector, cf. Table 1 for a summary.

A number of deep feed-forward network architectures were chosen with different depths and number of units per layer, while holding the total number of neuron units approximately constant. In total, four networks (*Net 1*, ..., *Net 4*) as indicated in Table 2, with parameters listed in Table 3, were evaluated for each scenario. A linear reference network *Net R* served as a baseline for comparison with linear discriminative source localization [5].

3.3 Performance evaluation

Performance of the trained localizers was evaluated as its F_1 score, the harmonic mean of precision and recall, the latter being averaged across all azimuths,

$$F_1 = \left(\frac{1}{2 \cdot precision} + \frac{1}{2 \cdot recall} \right)^{-1}. \quad (6)$$

To compute relative effects across architectures and geometries, van Rijsbergen's effectiveness E, defined as

$$E = 1 - F_1, \quad (7)$$

was used, which attains a value of zero for perfect classification.

3.4 Experiments

Experiments were carried out with the goal to systematically investigate the effect that different sensor geometries and deep network architectures as outlined above have on localization performance. Several additional parameters were varied in the experiments: Signal-to-noise ratio (SNR) ranged from clean to −10 dB. The maximum a-posteriori direction estimate has been computed on (unaveraged)

localization probability outputs of the networks on a 10 ms time-scale, as well as after temporal pooling of probabilities across 100 ms frames. Spatial precision with which a correct vs. false localization decision of the systems was evaluated ranged from $\pm 2.5^{\circ}$, i.e., within a single azimuth bin, to $\pm 15^{\circ}$, permitting an azimuth range around the true location as a faithful estimate. Results from a subset of experiments are reported below, which highlight the observed effects in a number of typical acoustic scenarios.

3.5 Results

Results obtained with deep networks architectures *Net 1*, ..., *Net 3* and the linear reference network *Net R* are shown in Fig. 2 for the 2-microphone behind-the-ear geometry ($G1$) without temporal pooling, and in Fig. 3 for the 6-microphone behind-the-ear geometry ($G3$) with temporal pooling of 100 ms. These two scenarios also represent the hardest and least-hard settings in which the algorithms have been evaluated, with all other scenarios (data not presented here due to space limitations) achieving performance measures in between. Network *Net 4* with the largest number of layers, but the smallest number of units per layer, resulted in poor performance on the localization task (data not shown here), likely indicating that wider processing layers are required, given the parameters in Table 3 which include 50% dropout units. Thus, *Net 4* was excluded from subsequent analysis.

The dominant effect of increased performance shown in Fig. 3 is due to the combination of temporal pooling and an increased number of sensors. Note that chance level is at 1/72, thus most of the datapoints shown fall well above chance. Scenario $G3$ shows localization performance near or above 90% for SNRs at 10 dB or better, which may be the most relevant range for real-world applications. For subsequent analysis, we have chosen to closer investigate results at 10 dB with an azimuth accuracy of $\pm 5^{\circ}$ due to its relevance in practice.

Table 4 shows van Rijsbergen's effectiveness E 3.3, indicating that DNN architectures perform significantly better than the linear reference net, albeit the differences between DNN architectures being minor. The improvement with 6 microphones ($G3$) instead of 4 microphones ($G2$) appears small, with the linear network in situation $G3$ still performing poorer than the DNN localizers in situation $G2$. Thus, information about source location in an interfering noise field may require non-linear processing for decoding, an effect that linear methods cannot compensate for by denser spatial sampling, cf. situation $G3$ with *Net R*. Table 5 investigates the effect of increasing the number of recording channels, showing relative improvement of geometries $G2$ and $G3$ over the 2-microphone geometry $G1$ (with the respective network architecture and pooling time-constant being held equal). The results show that DNN=processing obtains a larger benefit from an additional microphones compared to the linear network *Net R*.

$\tau = 10$ms

	Net 1	*Net 2*	*Net 3*	*Net R*	rel. imp.
G1	**0.60**	0.60	0.61	0.65	8.3%
G2	**0.31**	0.32	0.32	0.42	26.7%
G3	**0.30**	**0.30**	0.31	0.39	23.8%

$\tau = 100$ms

	Net 1	*Net 2*	*Net 3*	*Net R*	rel. imp.
G1	**0.28**	0.29	0.33	0.36	22.0%
G2	**0.06**	0.07	0.07	0.12	46.7%
G3	**0.06**	0.07	0.07	0.11	41.5%

Table 4: Relative improvement of best DNN architecture compared to linear reference network *Net R* (right column). Van Rijsbergen's effectiveness $E = 1 - F_1$ in acoustic scenario with 10 dB SNR and $\pm 5^{\circ}$ azimuth resolution, computed for all combinations of temporal resolution τ (10 ms, 100 ms), network architecture (*Net 1, 2, 3, R*), and microphone geometry (*G1, G2, G3*).

4 Summary and Discussion

In the present contribution, we have proposed a deep network approach to acoustic source localization in a hearing aid scenario with multiple behind-the-ear microphones mounted bilaterally on a head. While our previous work has shown that source localization in this setup can be carried out with high accuracy using learned linear filters, results presented here show that performance can be further increased through the use of non-linear learning algorithms such as deep feedforward networks. While the specific network architecture appeared to be of lesser significance, it may be of interest that

τ = 10ms

	Net 1	*Net 2*	*Net 3*	*Net R*
G2	**48.1%**	47.6%	47.3%	35.0%
G3	**50.3%**	49.8%	49.4%	40.1%

τ = 100ms

	Net 1	*Net 2*	*Net 3*	*Net R*
G2	76.5%	76.4%	**78.4%**	65.6%
G3	77.6%	77.1%	**79.3%**	70.1%

Table 5: Effect of increasing number of recording channels from 2 microphones (geometry *G1*) to 4 (*G2*) and 6 (*G3*), respectively. Relative improvement in effectiveness E compared to baseline geometry *G1*. Non-linear processing with DNNs more effectively extracts information conveyed in the additional channels than linear reference network *Net R*.

the improved performance of non-linear localization cannot be achieved with linear methods even if the sensor number is increased further: Linear models on 6-channel data were incapable of reaching the performance that non-linear networks achieved on 4-channel data. A saturation effect at 4 microphone setups that had been observed in previous studies with linear classifiers, i.e., use of 6 microphones would lead to only a comparably small increase in localization performance, has been confirmed in the present study for non-linear networks. Thus, practical applications should consider a reasonable number of microphones and devote additional resources to non-linear signal processing approaches in order to achieve optimum performance. Doing so has been shown here to result in relative improvements of effectiveness E of up to 46.7% over linear approaches. Depending on the specific situation, this is approximately equivalent to a 5 dB increase in signal-to-noise-ratio (SNR) of the recording condition.

References

[1] J. Anemüller and H. Kayser. Multi-channel signal enhancement with speech and noise covariance estimates computed by a probabilistic localization model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 156–160, 2017.

[2] C. Blandin, A. Ozerov, and E. Vincent. Multi-Source TDOA Estimation in Reverberant Audio Using Angular Spectra and Clustering. *Signal Processing*, 92:1950–1960, 2012.

[3] A. Brutti, M. Omologo, and P. Svaizer. Comparison between different sound source localization techniques based on a real data collection. In *IEEE Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pages 69–72, 2008.

[4] S. Chakrabarty and E. A. P. Habets. Multi-speaker doa estimation using deep convolutional networks trained with noise signals. *IEEE Journal of Selected Topics in Signal Processing*, 13:8–21, 2019.

[5] H. Kayser and J. Anemüller. A discriminative learning approach to probabilistic acoustic source localization. In *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 100–104, 2014.

[6] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier. Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses. *EURASIP Journal on Advances in Signal Processing*, 2009:ID 298605, 2009.

[7] H. Kayser, N. Moritz, and J. Anemüller. Probabilistic Spatial Filter Estimation for Signal Enhancement in Multi-Channel Automatic Speech Recognition. In *Interspeech*, pages 2562–2566, 2016.

[8] B. Laufer, R. Talmon, and S. Gannot. Relative transfer function modeling for supervised source localization. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[9] B. Loesch and B. Yang. Adaptive Segmentation and Separation of Determined Convolutive Mixtures Under Dynamic Conditions. In *Latent Variable Analysis and Signal Separation*, pages 41–48, 2010.

[10] M. Omologo and P. Svaizer. Acoustic Event Localization Using a Crosspower-Spectrum Phase Based Technique. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages II/273–II/276, 1994.

[11] R. Takeda and K. Komatami. Sound source localization based on deep neural networks with directional activate function exploiting phase information. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 405–409, 2016.