PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Deep network source localization and the influence of sensor geometry

Jörn ANEMÜLLER[(1)], Hendrik SCHOOF[(1)]

[(1)]Computational Audition Group, Medical Physics Section and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany, joern.anemueller@uni-oldenburg.de

Abstract

Learning-based localization approaches cast the acoustic speaker localization problem as a machine learning task where a classifier is trained on example data of acoustic feature vectors in order to predict likelihood of speech presence as a spatio-temporal distribution. We investigate the impact that fundamental acoustic parameters of the auditory scene (e.g. SNR, acoustic scene complexity, sensor geometry) exert on the ability to faithfully extract spatio-temporal activity maps for concurrent speakers. Our results indicate that to some degree shortcomings in the acoustic conditions can be compensated by increased complexity in the applied classification techniques. To this end, we systematically investigate localization performance for a set of deep neural network localizers of varying complexity, and for six different sensor configurations in a bilateral hearing aid setup. Deep networks result in improved performance compared to linear localizers, and their performance benefits more from an increase in the number of sensor channels. In specific configurations, deep networks with a smaller number of microphones perform better than a linear baseline network with a larger number of microphones. Thus, location-specific information in source-interference scenarios appears to be encoded non-linearly in the soundfield, requiring non-linear approaches for optimal decoding.

Keywords: Sound, Insulation, Transmission

1 INTRODUCTION

Acoustic source localization is a task routinely performed by the human auditory system. Several approaches have been proposed recently that formulate the acoustic source localization problem as a machine learning task where a mapping from acoustic features to the corresponding source location has to be learned from training data [4, 6, 8, 3, 1, 2].

The present work proposes a non-linear extension of our earlier linear approach [4] by employing deep feedforward 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 networks 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. The results presented here for speech sources embedded in isotropic noise are indicative 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 Discriminative deep network architecture for probabilistic acoustic source localization

The discriminative approach to source localization [4] builds on a standard classification framework that is employed to build decision models for directional sound source presence. Multi-layer feedforward neural network architectures are trained to learn an implicit representation of relevant acoustic parameters, thus no direct im-

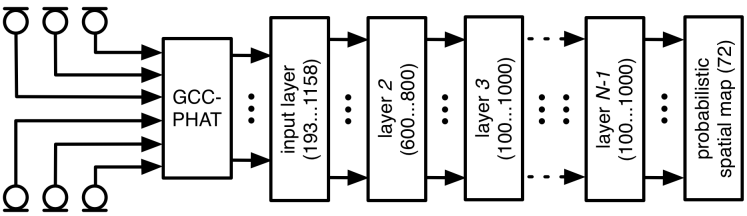

Figure 1. Processing diagram of the proposed algorithm.

Table 1. Geometries of the hearing aid setup and the resulting number of GCC-PHAT coefficients that form the feature vector input for discriminative localization.

Geom.	#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

pulse response measurements and no additional assumption on the acoustics are required.

Source presence is indicated by cross-correlation function features $\rho_{ij}(\tau)$, containing a main peak centered around the TDOA $\tau_{ij}(\zeta)$ corresponding to location ζ. Due to their invariance to spectral changes, generalized cross-correlation phase transform (GCC-PHAT) [7] are employed here. 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.

During classifier training, example feature vectors ϕ 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 $\hat{P}_{source}(\zeta,t)$ that indicates the probability of a source being active for each time frame t and each direction ζ. When the single most probable source direction is to be identified, maximum a-posteriori estimates are computed from the probabilistic location map. Multi-source DOA estimation is achieved by evaluation of the J most probable occurrences of sound source positions.

Fig. 1 provides a summary of the network structures examined in the present study, with varying number of network layers and different layer sizes. To investigate the influence also of acoustic parameters, the discriminative localization method has been applied to a number of different sensor geometries, with between two and six channels of the bilateral Hearing aid setup being used. Table 1 summarizes microphone geometries and resulting GCC-PHAT feature vector dimensionality.

Table 2. Performance of DNN architecture compared to linear reference network *Net R* in terms of van Rijsbergen's effectiveness E in acoustic scenario with 10 dB SNR.

	τ/ms	*Net 1*	*Net 2*	*Net 3*	*Net R*	rel. imp.
G1	10	**0.60**	0.60	0.61	0.65	8.3%
G2	10	**0.31**	0.32	0.32	0.42	26.7%
G3	10	**0.30**	**0.30**	0.31	0.39	23.8%
G1	100	**0.28**	0.29	0.33	0.36	22.0%
G2	100	**0.06**	0.07	0.07	0.12	46.7%
G3	100	**0.06**	0.07	0.07	0.11	41.5%

3 EXPERIMENTAL EVALUATION

3.1 Training and evaluation data

Data for training and evaluation comprising, in total, 15 hours of multi-channel data were generated from a database of multi-channel head-geometry room impulse response functions [5] and the TIMIT speech corpus. These included 144 unique speaker-utterance combinations for each SNR condition per direction, with space subdivided into 72 direction-of-arrival locations, spaced 5 degrees apart.

3.2 Experiments

Experiments were carried out in order to systematically investigate the effect that different sensor geometries and deep network architectures as outlined above have on localization performance. 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. Results from a subset of experiments are reported below, which highlight the observed effects in a number of typical acoustic scenarios. Experimental conditions *not* reported here include a variation in required spatial localization accuracy, additional temporal pooling time-constants, and presence of a localized interfering speaker in addition to isotropic noise.

3.3 Results

Table 2 shows van Rijsbergen's effectiveness $E = 1 - F_1$, 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 3 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*.

4 SUMMARY

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

Table 3. Effect of increasing number of recording channels from 2 microphones (geometry *G1*) to 4 (*G2*) and 6 (*G3*).

	τ/ms	*Net 1*	*Net 2*	*Net 3*	*Net R*
G2	10	**48.1%**	47.6%	47.3%	35.0%
G3	10	**50.3%**	49.8%	49.4%	40.1%
G2	100	76.5%	76.4%	**78.4%**	65.6%
G3	100	77.6%	77.1%	**79.3%**	70.1%

learning algorithms such as deep feedforward networks. While the specific network architecture appeared to be of lesser significance, it may be of interest that 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.

ACKNOWLEDGEMENTS

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.

References

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

[2] S. Chakrabarty and E. A. P. Habets. Broadband doa estimation using convolutional neural networks trained with noise signals. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017.

[3] S. Kataria, C. Gaultier, and A. Deleforge. Hearing in a shoe-box: binaural source position and wall absorption estimation using virtually supervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 226–230. IEEE, 2017.

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

[5] 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(1):ID 298605, 2009.

[6] H. Kayser, N. Moritz, and J. Anemüller. Probabilistic Spatial Filter Estimation for Signal Enhancement in Multi-Channel Automatic Speech Recognition. In *Proc. INTERSPEECH 2016*, 2016.

[7] M. Omologo and P. Svaizer. Acoustic Event Localization Using a Crosspower-Spectrum Phase Based Technique. *Proc. ICASSP 1994. IEEE International Conference on Acoustics, Speech and Signal Processing*, ii(2):II/273–II/276, 1994.

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