Deep Neural Network based Distance Estimation for Geometry Calibration in Acoustic Sensor Networks

Tobias Gburrek[1], Joerg Schmalenstroeer[1], Andreas Brendel[2], Walter Kellermann[2], Reinhold Haeb-Umbach[1]
[1]*Department of Communications Engineering, Paderborn University, Germany*
[2]*Multimedia Communications and Signal Processing, FAU Erlangen-Nuernberg, Germany*
{gburrek, schmalen, haeb}@nt.uni-paderborn.de, {Andreas.Brendel, Walter.Kellermann}@FAU.de

***Abstract*—We present an approach to deep neural network based (DNN-based) distance estimation in reverberant rooms for supporting geometry calibration tasks in wireless acoustic sensor networks. Signal diffuseness information from acoustic signals is aggregated via the coherent-to-diffuse power ratio to obtain a distance-related feature, which is mapped to a source-to-microphone distance estimate by means of a DNN. This information is then combined with direction-of-arrival estimates from compact microphone arrays to infer the geometry of the sensor network. Unlike many other approaches to geometry calibration, the proposed scheme does only require that the sampling clocks of the sensor nodes are roughly synchronized. In simulations we show that the proposed DNN-based distance estimator generalizes to unseen acoustic environments and that precise estimates of the sensor node positions are obtained.**

***Index Terms*—DNN, CDR, acoustic distance estimation, geometry calibration**

I. Introduction

A wireless acoustic sensor network (WASN) consists of small devices called "nodes", which are connected via wireless links. Each node is equipped with memory, a wireless network interface, a processing unit and one or multiple microphones. WASNs are used in surveillance, human-machine interfaces and environmental monitoring tasks [1]. Distributing microphones in an environment comes with the promise that there is always a sensor close to each relevant sound source. Thus, WASNs offer the potential of improved signal enhancement and acoustic localization capabilities, compared to a single compact microphone array.

Acoustic source localization can, e.g., be used to steer a camera towards a moving speaker in a smart home scenario [2]. In such multi-modal setups the usage of a common coordinate system eases the process of data fusion. Hence, knowledge of the position and orientation of the sensor nodes within a chosen coordinate system is required in these scenarios to provide absolute positioning information. The process of determining the nodes' position and orientation is called geometry calibration.

However, manual geometry calibration is a tedious task, in particular if the network consists of a large number of nodes, and any change in the setup asks for recalibration. Therefore, automatic geometry calibration from the observed acoustic signals is desirable, and, indeed, has been studied extensively, see [3] for an overview.

It appears natural to consider geometry calibration and sampling clock synchronization jointly, because correlation-based measures, such as the time difference of arrival (TDoA) of signals at different nodes, typically used to infer geometric relations among the nodes, require a synchronous network. However, such methods (e.g., [4], [5]) often require additional information, such as the position of anchor sources [4], which may not be available in practice.

In this paper we take a different approach: rather than relying on a synchronous network or jointly estimating the geometry and the sampling clock offsets, we develop a technique which only needs a rough synchronization across sensor nodes. We do, however, assume that each node has a synchronous microphone array of known topology instead of only a single microphone. The rough synchronization between the nodes is needed to match the observations made by the different nodes.

Our proposed approach to geometry calibration is an extension of the method we presented in [6], which utilizes direction-of-arrival (DoA) estimates computed from the microphone array signals of each node. But, DoA information alone can only infer a "relative" geometry, lacking any absolute distances. Thus, the inferred geometry can be determined only up to an unknown scaling factor. In this contribution we propose to compute this scaling factor from source-to-sensor distance estimates gleaned from the acoustic properties of the microphone signals.

In [7] and [8] it was shown that the coherent-to-diffuse power ratio (CDR) can be utilized to estimate the distance between a microphone pair and an acoustic source. There, Gaussian processes (GPs) were used for CDR-based distance estimation. However, the GPs were learned for a certain acoustic environment and generalization to unseen environments is expected to yield poor results.

We overcome this restriction by using DNNs, which are trained on various acoustic environments such that both, the room characteristics and the distance, can be extracted from the CDR. To further support the generalization of the learned model to varying acoustic environments, we employ the recently proposed R-vectors as additional input to the network, which are meant to capture the room properties [9].

The remainder of the paper is organized as follows: In Sec. II the idea of CDR-based distance estimation is reviewed, followed by the new DNN approach in Sec. III. Subsequently, geometry calibration using DoAs and distance estimates is explained in Sec. IV. Finally, simulation results are presented in Sec. V and some conclusions are drawn in Sec. VI.

II. CDR-Based Distance Estimation

We consider a microphone pair, which records a single acoustic source in a reverberant environment. The recorded signal can be decomposed into a coherent component and a diffuse part. The CDR measures the power ratio of these components, which is related to the distance between the source and the microphones as shown in [7]. In [10, Eq. 12] a DoA-independent estimator, yielding the estimate $\widehat{\mathrm{CDR}}(l,k)$, was derived, where l indexes the time frame and k the frequency bin, respectively. From $\widehat{\mathrm{CDR}}(l,k)$, the so-called diffuseness

$$\hat{\mathrm{D}}(l,k) = \frac{1}{1+\widehat{\mathrm{CDR}}(l,k)} \tag{1}$$

can be computed, which we will use in the following, because, unlike the CDR, it is limited to the interval [0, 1].

To achieve robustness against temporal inactivity of the acoustic source, it was proposed in [7] to average the diffuseness $\hat{\mathrm{D}}(l,k)$ across a sufficiently large number of time frames and frequency bins, resulting in the averaged diffuseness ζ.

A. GP-Based Distance Estimation

In prior works on CDR-based distance estimation, e.g., [7] and [8], it was proposed to use GP regression trained on the averaged diffuseness ζ for distance estimation.

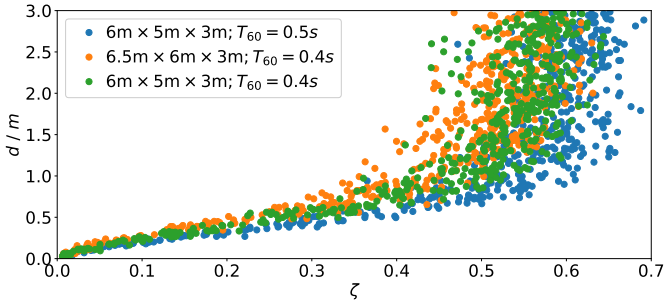

Fig. 1. Dependency of the relationship between ζ and the distance d on the acoustic environment: Each datapoint corresponds to a randomly drawn source-microphone constellation. The legend in the plot shows the dimensions of the considered rooms and the corresponding reverberation time T_{60}.

Fig. 1 shows the relationship between the averaged diffuseness ζ and the distance d for three exemplary rooms with different acoustic characteristics. Obviously, this relationship strongly depends on the acoustic properties of the room, since the energy decay of the coherent signal component is directly affected by the reverberation time T_{60} and many other factors [7]. Consequently, a GP, learned for a certain acoustic environment, will not generalize to other acoustic environments. Additionally, ζ tends for large values towards an asymptote, prohibiting the exact estimation of the regression function and, thus, degrading the performance also for smaller distances.

III. DNN-Based Distance Estimation

DNNs have many learnable parameters, which gives them an increased modeling power compared to GPs. Hence, they may be able to take advantage of the additional information present in the high-resolution estimate $\hat{\mathrm{D}}(l,k)$ compared to the scalar value ζ. Keeping all the information contained in the time-frequency pattern, the DNN has the freedom to decide by itself, how to best combine this information, rather than defining this combination beforehand. Actually, we used a window consisting of several frames of the diffuseness as input feature. Additionally, we enable the DNN to learn room characteristics by presenting data from various acoustic environments. The underlying hypothesis is that this will allow the DNN to map $\hat{\mathrm{D}}(l,k)$ to a distance estimate irrespective of the present room characteristics.

Fig. 1 shows that the variance of ζ grows as a function of the distance, which also holds for $\hat{\mathrm{D}}(l,k)$. Therefore, to avoid unreliable measurements, we concentrate on small distances and exclude distances that are larger than an upper bound $r_{\max}$. This restriction is not detrimental for scaling geometries, since only one reliable distance estimate is sufficient and there is mostly a node of a WASN close to each relevant sound source.

To handle the growing variance of $\hat{\mathrm{D}}(l,k)$ for larger distances we formulated distance estimation as a classification problem rather than a regression problem, by which small and large deviations from the ground truth distance (class) are penalized equally. Thus, a distinction between the large distances, which are more tricky to be estimated correctly, is enforced. When distance estimation is formulated as a regression problem this can be circumvented by the DNN because the loss can be minimized by estimating an average distance for the larger distances. The error due to the categorization of the distance into classes has a negligible effect on geometry scaling because this error is rather small (a few cm) compared to the inter-node distances (a few m). For distances larger than $r_{\max}$ an additional class called out-of-range (OoR) is introduced.

We investigate two types of DNNs, a simple multilayer perceptron (MLP) and a convolutional recurrent neural network (CRNN). The architectures of the DNNs are given in Tab. I and Tab. II, respectively. Hereby, B denotes the size of the mini-batches, F the number of frequency bins, T the number of time frames, C the number of classes, and R the dimension of the R-vector, which will be introduced later.

The major difference between the two types of DNNs lies in the usage of temporal information. When using the MLP, we average the diffuseness over all time frames in the considered observation interval to obtain the input feature vector of the DNN. Thus, all time information is discarded. In contrast, the CRNN is able to utilize temporal information contained in $\hat{\mathrm{D}}(l,k)$, e.g., information about the activity of the coherent source. This will also be reflected by the simulation results in Sec. V.

Distance estimation can be further improved by utilizing R-vectors as additional input feature. R-vectors have been introduced in [9] to capture information about the acoustic environment in automatic speech recognition.

This idea can be transferred to distance estimation, whereby the R-vector is used to capture information about the current environment, e.g., the reverberation time T_{60}. As shown in

TABLE I

ARCHITECTURE OF THE MLP USED FOR DISTANCE ESTIMATION: DROPOUT WITH A PROBABILITY OF 0.5 IS USED IN ALL HIDDEN LAYERS.

Block	Output shape
Diffuseness	$B \times F$
optional: Concat R-vector	$B \times (F + R)$
$3 \times \text{fc}_{\text{ReLU}}(1024)$	$B \times 1024$
$\text{fc}_{\text{Softmax}}(C)$	$B \times C$

TABLE II

ARCHITECTURE OF THE CRNN USED FOR DISTANCE ESTIMATION: EACH CONV{1,2}D LAYER INCLUDES RELU AS ACTIVATION AND BATCH NORMALIZATION. ONLY THE LAST OUTPUT VECTOR OF THE GRU IS FORWARDED TO THE CLASSIFICATION NETWORK.

Block	Output shape
Diffuseness	$B \times 1 \times F \times T$
$2 \times$ Conv2d$(7 \times 3; 16)$	$B \times 16 \times F \times T$
MaxPool2d(4×2)	$B \times 16 \times \lfloor F/4 \rfloor \times \lfloor T/2 \rfloor$
$2 \times$ Conv2d$(7 \times 3; 32)$	$B \times 32 \times \lfloor F/4 \rfloor \times T$
MaxPool2d(4×2)	$B \times 32 \times \lfloor F/16 \rfloor \times \lfloor T/4 \rfloor$
Reshape	$B \times 32 \cdot \lfloor F/16 \rfloor \times \lfloor T/4 \rfloor$
Conv1d$(3; 512)$	$B \times 512 \times \lfloor T/4 \rfloor$
Conv1d$(3; 256)$	$B \times 256 \times \lfloor T/4 \rfloor$
$2 \times$ GRU(256)	$B \times 256$
optional: Concat R-vector	$B \times (256 + R)$
$\text{fc}_{\text{ReLU}}(256)$	$B \times 256$
$\text{fc}_{\text{Softmax}}(C)$	$B \times C$

Tab. I and Tab. II the R-vector will be concatenated either with the input feature vector of the MLP or the output of the gated recurrent unit (GRU) layer of the CRNN.

R-vectors correspond to the output of an intermediate layer of a DNN trained to classify room impulse responses (RIRs) from reverberated signal recordings. In [9] it was suggested to use a time delay neural network (TDNN) for R-vector extraction. However, we decided to replace the TDNN by a convolutional neural network for simplicity, whereby this decision was inspired by the x-vector extractor presented in [11]. The architecture of our R-vector extractor can be found in Tab. III. We use the output of the first fully connected layer as R-vector.

TABLE III

ARCHITECTURE OF THE R-VECTOR EXTRACTOR: EACH CONV1D LAYER INCLUDES RELU AS ACTIVATION. BATCH NORMALIZATION IS USED.

Block	Output shape
MFCC	$B \times 23 \times T$
Conv1d$(3; 128)$	$B \times 128 \times T$
Conv1d$(5; 128)$	$B \times 128 \times T$
Conv1d$(1; 128)$	$B \times 128 \times T$
StatisticPool	$B \times 256$
R-Vector $= \text{fc}_{\text{ReLU}}(512)$	$B \times 512$
$\text{fc}_{\text{ReLU}}(512)$	$B \times 512$
$\text{fc}_{\text{Softmax}}(C)$	$B \times C$

IV. Geometry Calibration

We simulated two-dimensional scenarios with K nodes and N independent acoustic sources, with only one source being active at any given time (see Fig. 2 for an example).

In order to determine the nodes' positions and orientations, we use the geometry calibration method introduced in [6], which also provides the position of the acoustic sources. There, an objective function is defined which assesses the compatibility of the $K \cdot N$ DoA estimates with an assumed geometry.

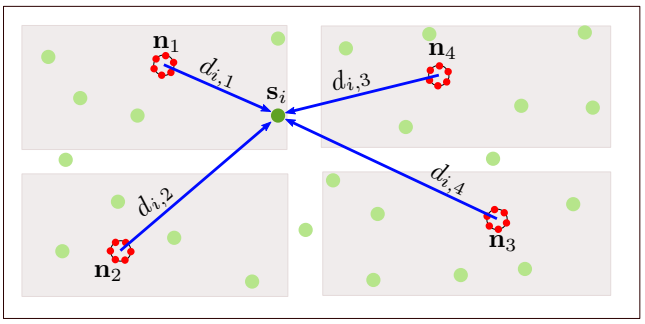

Fig. 2. Example of a random setup with four microphone arrays at positions $\mathbf{n}_j$, and highlighted i-th source position $\mathbf{s}_i$ (red dots: microphones; green dots: source positions; gray area: possible positions to randomly place nodes (microphone arrays); all nodes and sources have a minimum distance of 0.5 m to the closest wall; 1 m spacing between the gray areas; the dimensions of the room are drawn from $[6\,\text{m}, 7\,\text{m}] \times [5\,\text{m}, 6\,\text{m}]$; room height=3 m)

This nonlinear objective function is iteratively minimized, see [6] for details. Additionally, we embed the calibration method into a similar random sample consensus (RANSAC) method as the one described in [12] to be more robust against outliers in the DoA estimates.

Due to the fact that this method only utilizes DoA estimates, the optimization problem suffers from scale invariance [12]. To avoid the trivial solution (all unknowns equal to zero), the following equality constraint, which relates all inter-node distances, is added to the optimization problem:

$$\sum_{i=1}^{K} \sum_{j=i}^{K} ||\mathbf{n}_i - \mathbf{n}_j||_2 = 1, \tag{2}$$

with $\mathbf{n}_i$ and $\mathbf{n}_j$, $i, j \in \{1, 2, 3, 4\}$, denoting the node positions.

We use the estimated source-node distances to determine the unknown scaling factor v of the calibration results, which arises from the introduced constraint. As mentioned earlier, a single source-node distance estimate would ideally be sufficient. But better results are obtained if all available distance estimates are utilized.

The unknown $v \in \mathbb{R}^+$ is determined by scaling the source-node distances of the unscaled geometry to the distance estimates. This results in the following weighted least squares problem

$$\hat{v} = \underset{v}{\text{argmin}} \sum_{i=1}^{N} \sum_{j=1}^{K} w_{ij} \left(v||\hat{\mathbf{s}}_i - \hat{\mathbf{n}}_j||_2 - \hat{d}_{ij} \right)^2, \tag{3}$$

where $\hat{\mathbf{s}}_i$ denotes the unscaled estimate of the position of the i-th source, $\hat{\mathbf{n}}_j$ the unscaled estimate of the position of the j-th node, and $\hat{d}_{ij}$ the estimate of the distance between source i and node j. The weights w_{ij} are introduced to account for the distance dependence of the variance of the distance estimates, see Fig. 1. They are chosen to be: $w_{ij} = 1/||\hat{\mathbf{s}}_i - \hat{\mathbf{n}}_j||_2$.

The optimization in Eq. (3) leads to

$$\hat{v} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{K} w_{ij}\, \hat{d}_{ij}\, ||\hat{\mathbf{s}}_i - \hat{\mathbf{n}}_j||_2}{\sum_{i=1}^{N} \sum_{j=1}^{K} w_{ij}\, ||\hat{\mathbf{s}}_i - \hat{\mathbf{n}}_j||_2^2}. \tag{4}$$

The properties of the sensor nodes used in the simulations were inspired by the hardware described in [13], where each

node is equipped with a circular array that consists of six microphones. The two opposite microphones which are $5\,\mathrm{cm}$ apart form a pair used for distance estimation, giving three distance estimates per array, which are combined by checking the consistency of the three estimates. Using the microphones exhibiting the largest distance in an array is a reliable choice in practice (see, e.g., [7]). If at least two estimates coincide, we select the corresponding estimate for geometry scaling and exclude the source-node pair otherwise ($w_{ij} = 0$). Besides, we do not utilize the corresponding source-node pair for geometry scaling if at least one node provides the OoR class.

V. Simulation Results

Simulated data is used for the evaluation of our approach as well as for training the DNNs. We utilize the image source method [14] to simulate RIRs, using the implementation of [15]. The RIRs are used to reverberate the source signals, which can be either white Gaussian noise or speech, whereby the used speech samples are taken from the TIMIT database [16]. Due to additional physical effects that are not considered by the simulated data, e.g., directional sources, an adaptation of DNN-based distance estimation to real data is expected to be beneficial for real microphone recordings and will be considered in future work.

The distance estimators are trained and evaluated on data sets, consisting of single source-node pairs, which are uniformly drawn from the room layout at a height of $1.5\,\mathrm{m}$. Due to the fact that the accuracy of distance estimates degrades if the source or the node is located in the vicinity of walls (see [17]), a minimum distance of $0.5\,\mathrm{m}$ to the closest wall is ensured for all nodes and sources.

We use separate data sets for distance estimator training and R-vector extractor training, which we consider to be more realistic than using the same data sets for the training of both. In both data sets rooms with a height of $3\,\mathrm{m}$ are considered. The room dimensions are uniformly drawn from the set $[6\,\mathrm{m}, 7\,\mathrm{m}] \times [5\,\mathrm{m}, 6\,\mathrm{m}]$ for the distance estimator and from the set $[5\,\mathrm{m}, 8\,\mathrm{m}] \times [4\,\mathrm{m}, 7\,\mathrm{m}]$ for the R-vector extractor. Besides, the reverberation time T_{60} is uniformly drawn at random from $[0.2\,\mathrm{s},\ 0.5\,\mathrm{s}]$ and $[0.1\,\mathrm{s},\ 0.6\,\mathrm{s}]$, respectively. We placed the sources such that the distance to the nodes is uniformly drawn from $[0.03\,\mathrm{m},\ 3\,\mathrm{m}]$. In the R-vector data set all sources are uniformly distributed in the room. Both training data sets contain 10000 source-node pairs, whereby additional 1000 OoR examples are added for the simulations corresponding to Tab. V and Tab. VI.

In order to evaluate our approach to geometry calibration, we consider scenarios with a setup as depicted in Fig. 2. All scenarios consist of $K{=}4$ nodes and $N{=}30$ successively active sources, whereby each source corresponds to a $3\,\mathrm{s}$ long speech signal, which is generated as described before.

A. Model Configuration

All DNNs are trained using Adam [18] with a mini-batch size of $B{=}32$ and a learning rate of $3{\cdot}10^{-4}$, whereby the distance is linearly quantized into $C{=}31$ classes plus one additional class for OoR. The short-time Fourier transform, which is needed to estimate $\widehat{\mathrm{CDR}}(l,k)$, uses a Blackman window with a length of $25\,\mathrm{ms}$ and $10\,\mathrm{ms}$ shift. Additionally, we estimate the power spectral densities, used for CDR estimation, by recursive averaging with forgetting factor $\lambda{=}0.95$, as described in [7]. Furthermore, $\hat{\mathrm{D}}(l,k)$ is calculated for all frequencies between $125\,\mathrm{Hz}$ and $3.5\,\mathrm{kHz}$, which corresponds to the frequency range, where speech has significant power.

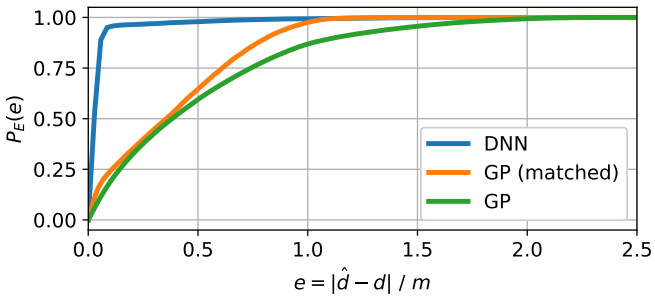

Fig. 3. Cumulative distribution function of the error $P_E(e)$ of CRNN-based and GP-based distance estimates: The GP (zero-mean prior and γ-exponential covariance function [8]) is trained on a single acoustic environment and tested either on the same environment (matched) or on the evaluation data set that contains multiple environments. Speech is used as source signal.

B. Distance Estimation

We first evaluate the proposed distance estimators, and use the mean-absolute error (MAE) as performance metric

$$e_{AE} = \frac{1}{M} \sum_{m=1}^{M} |\hat{d}_m - d_m|. \tag{5}$$

Here, d_m denotes the ground truth distance and $\hat{d}_m$ the distance estimate. The evaluation set contains 10000 source-node constellations, which results in $M{=}30000$ source-microphone-pair constellations.

Fig. 3 shows a comparison of GP-based and DNN-based distance estimation. It can be seen that the proposed approach outperforms the GP-based method, even so if the GP is applied to the acoustic room characteristics, on which it was trained.

Tab. IV provides results for distance estimation using different input signals and DNN architectures. It becomes obvious that the best results can be achieved, when a CRNN is used, which is able to utilize temporal information. Additionally, R-vectors, which contain distance information by itself (see first row of Tab. IV), are helpful to reduce the error in all cases. Nevertheless, R-vectors have a diminishing effect, when a CRNN is used and speech is the input signal. This means

TABLE IV
PERFORMANCE OF THE DISTANCE ESTIMATOR FOR DIFFERENT TYPES OF SOURCE SIGNALS AND DIFFERENT DNN ARCHITECTURES

Architecture	Diffuseness	R-vector	MAE / m Noise	MAE / m Speech
MLP		✓	0.176	0.151
MLP	✓		0.119	0.148
MLP	✓	✓	0.064	0.070
CRNN	✓		0.087	0.055
CRNN	✓	✓	0.062	0.052

that the diffuseness contains already enough information about the environment. Moreover, better results can be achieved when speech is used instead of white Gaussian noise. We hypothesize that the correlation properties of the diffuseness resulting from speech support the learning process of the convolutional layers and, thus, are beneficial for gathering information about distances and environments.

TABLE V
INFLUENCE OF SNR AND OOR DETECTION ON DISTANCE ESTIMATION: 2500 OOR EXAMPLES ARE ADDED TO THE EVALUATION SET. THE CRNN (DIFFUSENESS + R-VECTOR) IS APPLIED TO SPEECH.

SNR/dB	**Fusion**	**# Discards**	**MAE**/m	**F_1-score (OoR Detection)**
30		-	0.046	91.44%
30	✓	148	0.033	94.90%
20	✓	166	0.034	94.67%
10	✓	308	0.044	91.86%
5	✓	745	0.062	86.00%

The influence of sensor noise, which is simulated by adding white Gaussian noise to the reverberated signal, and the introduced OoR class can be seen in Tab. V. To generate the corresponding training data, integer values in the range from $5\,\mathrm{dB}$ to $30\,\mathrm{dB}$ are randomly chosen for the signal-to-noise ratio (SNR). Obviously, our approach is robust against a wide range of sensor noise levels. Furthermore, the fusion of the three distance estimates per node improves the performance.

C. Geometry Calibration

Next, we examine the geometry calibration performance. It is to be mentioned that all results, which are provided by our geometry calibration method, are given relative to the position and orientation of a reference node. Thus, the calibration results are matched to the ground truth geometry by a rigid body transformation for evaluation. For DoA estimation, the complex Watson kernel method [19] is used.

The mean position error (MPE) of the nodes' positions is used as metric

$$e_{PE} = \frac{1}{4M} \sum_{m=1}^{M} \sum_{j=1}^{4} ||\hat{\mathbf{n}}_{j,m} - \mathbf{n}_{j,m}||_2. \tag{6}$$

Tab. VI shows the MPE of the calibration results for different values of T_{60} for $M{=}100$ scenarios. Noticeably, the MPE increases for larger T_{60} values. This is caused by the degradation of the DoA estimates in more reverberant environments (see [19]). Additionally, the distance estimation errors influenced the calibration results only marginally.

TABLE VI
MPE OF THE SCALED GEOMETRY CALIBRATION RESULTS

T_{60}/ms / Distances	200	300	400	500
Ground truth	0.080m	0.134m	0.187m	0.213m
Estimates (CRNN)	0.084m	0.14m1	0.197m	0.228m

VI. CONCLUSIONS

In this paper we proposed a DNN-based distance estimator. It takes acoustic signal diffuseness information and, optionally, R-vectors to capture acoustic properties of the room, as input and predicts the distance between an acoustic source and a recording node. The distance estimates are combined with DoA estimates to infer the positions and orientations of the sensor nodes of a WASN. Simulations have shown that the distance estimator provides estimates, which exhibit an average error that is smaller than $6.5\,\mathrm{cm}$ under various acoustic conditions, and enables accurate geometry calibration results.

ACKNOWLEDGMENT

This work was supported by DFG under contract no <SCHM 3301/1-2> and <KE 890/10-2> within the framework of the Research Unit FOR2457 "Acoustic Sensor Networks".

REFERENCES

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. IEEE SCVT*, Nov 2011.

[2] J. Schmalenstroeer, V. Leutnant, and R. Haeb-Umbach, "Audio-visual data processing for ambient communication," in *Proc. Conference on Artificial Intelligence*, 2009.

[3] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink, "Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms," *IEEE Signal Processing Magazine*, vol. 33, no. 4, July 2016.

[4] F. Ma, Z. Liu, and F. Guo, "Direct position determination in asynchronous sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 8790–8803, Sep. 2019.

[5] C. Jia, J. Yin, D. Wang, and L. Zhang, "Lagrange programming neural network for TOA-based localization with clock asynchronization and sensor location uncertainties," *Sensors*, vol. 18, p. 2293, 07 2018.

[6] F. Jacob, J. Schmalenstroeer, and R. Haeb-Umbach, "Microphone array position self-calibration from reverberant speech input," in *Proc. IWAENC*, Sep. 2012, pp. 1–4.

[7] A. Brendel and W. Kellermann, "Distributed Source Localization in Acoustic Sensor Networks using the Coherent-to-Diffuse Power Ratio," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[8] A. Brendel, A. Regensky, and W. Kellermann, "Probabilistic modeling for learning-based distance estimation," in *Proceedings of the 23rd International Congress on Acoustics*, Aachen, Germany, Sep. 2019.

[9] Y. Khokhlov, A. Zatvornitskiy, I. Medennikov, I. Sorokin, T. Prisyach, A. Romanenko, A. Mitrofanov, V. Bataev, A. Andrusenko, M. Korenevskaya, and O. Petrov, "R-vectors: New technique for adaptation to room acoustics," in *Proc. Interspeech 2019*, 09 2019.

[10] A. Schwarz and W. Kellermann, "Unbiased coherent-to-diffuse ratio estimation for dereverberation," in *Proc. IWAENC*, Sep. 2014.

[11] H. Zeinali, L. Burget, and J. Cernocky, "Convolutional neural networks and x-vector embedding for DCASE2018 acoustic scene classification challenge," *arXiv preprint arXiv:1810.04273*, 2018.

[12] J. Schmalenstroeer, F. Jacob, R. Haeb-Umbach, M. Hennecke, and G. A. Fink, "Unsupervised geometry calibration of acoustic sensor networks using source correspondences," in *Interspeech 2011*, 2011.

[13] H. Afifi, J. Schmalenstroeer, J. Ullmann, R. Haeb-Umbach, and H. Karl, "MARVELO - a framework for signal processing in wireless acoustic sensor networks," in *ITG Fachtagung Sprachkommunikation*, Oct. 2018.

[14] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 04 1979.

[15] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," 1993.

[17] A. Brendel, I. Altmann, and W. Kellermann, "Acoustic Source Position Estimation based on Multi-Feature Gaussian Processes," in *Proc. EUSIPCO*, A Coruna, Spain, Sep. 2019.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[19] L. Drude, F. Jacob, and R. Haeb-Umbach, "DOA-estimation based on a complex Watson kernel method," in *Proc. EUSIPCO*, Aug 2015.