Gburrek *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:25
https://doi.org/10.1186/s13636-021-00210-x

EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY Open Access

Geometry calibration in wireless acoustic sensor networks utilizing DoA and distance information

Tobias Gburrek*, Joerg Schmalenstroeer and Reinhold Haeb-Umbach

Abstract

Due to the ad hoc nature of wireless acoustic sensor networks, the position of the sensor nodes is typically unknown. This contribution proposes a technique to estimate the position and orientation of the sensor nodes from the recorded speech signals. The method assumes that a node comprises a microphone array with synchronously sampled microphones rather than a single microphone, but does not require the sampling clocks of the nodes to be synchronized. From the observed audio signals, the distances between the acoustic sources and arrays, as well as the directions of arrival, are estimated. They serve as input to a non-linear least squares problem, from which both the sensor nodes' positions and orientations, as well as the source positions, are alternatingly estimated in an iterative process. Given one set of unknowns, i.e., either the source positions or the sensor nodes' geometry, the other set of unknowns can be computed in closed-form. The proposed approach is computationally efficient and the first one, which employs both distance and directional information for geometry calibration in a common cost function. Since both distance and direction of arrival measurements suffer from outliers, e.g., caused by strong reflections of the sound waves on the surfaces of the room, we introduce measures to deemphasize or remove unreliable measurements. Additionally, we discuss modifications of our previously proposed deep neural network-based acoustic distance estimator, to account not only for omnidirectional sources but also for directional sources. Simulation results show good positioning accuracy and compare very favorably with alternative approaches from the literature.

Keywords: Geometry calibration, Acoustic distance estimation, Deep neural network, Coherent-to-diffuse power ratio, Direction of arrival

1 Introduction

A wireless acoustic sensor network (WASN) consists of sensor nodes, which are connected via a wireless link and where each node is equipped with one or more microphones, a computing and a networking module [1, 2]. A network of distributed microphones offers the advantage of superior signal capture, because it increases the probability that a sensor is close to every relevant sound source, be it a desired signal or an interfering source. Information about the position of an acoustic source may be used for acoustic beamforming and for realizing location-based functionality, such as switching on lights depending on a speaker's position or steering a camera to a speaker who is outside its field of view. Source position information is also beneficial for the estimation of the phase offset between the sampling oscillators of the distributed sensor nodes [3, 4].

However, source location information can only be obtained from the audio signals without using additional prior knowledge, e.g., about source position candidates, like it is used in fingerprinting-based methods [5, 6], if the position of the sensors, i.e., the microphones, is known. This, however, is an unrealistic assumption, because one of the key advantages of WASNs is that they are typically an ad hoc network formed by non-stationary devices,

*Correspondence: gburrek@nt.upb.de
Department of Communications Engineering, Paderborn University, Paderborn, Germany

e.g., the smartphones of users, and, possibly, stationary devices, such as a TV or a smart speaker. For such a setup, the spatial configuration and even the number of sensor nodes is unknown a priori and may even be changing over time, e.g., with people, and thus smartphones, entering and leaving the setup.

Geometry calibration refers to the task of determining the spatial position of the distributed microphones [7]. In case of sensor nodes equipped with an array of microphones [8], the orientation of the array is also of interest. An ideal calibration algorithm should infer the geometry of the network while the network is being used, i.e., solely from the recorded audio signals, neither requiring the playback of special calibration signals nor human assistance through manually measured distances. The calibration should be fast, not only during initial setup but also when detecting a change in the network configuration [9] which triggers a re-calibration.

There is a further desirable feature, which is the independence from synchronized sampling clocks across the network (see [10–12]). Clearly, the tasks of geometry calibration and synchronization of the sensor nodes' sampling clocks are often closely linked [7]. Geometry calibration approaches relying on time difference of arrival (TDoA) [13, 14], time of arrival (ToA) [15], or time of flight (ToF) [16] information investigate time points of sound emission and/or intersignal delays, requiring that the clocks of the sensor nodes are synchronized.

Only the direction of arrival (DoA)-based approach does not require clock synchronization at (sub-)sample precision. Here, the assumption is that sensor nodes are equipped with microphone arrays to be able to estimate the angle under which an acoustic source is observed. This requires that the microphones comprising the array share the same clock signal, while the clocks at different nodes only need to be coarsely synchronized, e.g., via [17–20]. That coarse synchronization, i.e., a synchronization with an accuracy of a few tens of milliseconds, is necessary to identify same signal segments across devices. DoA-based calibration obviously suffers from scale indeterminacy: only a relative geometry can be estimated, as no information is available to infer an absolute distance.

Once measurements are given, be it ToA, TDoA, DoA or even combinations thereof [21, 22], the actual estimation of the spatial arrangement of the network amounts to the optimization of a cost function, which measures the agreement of an assumed geometry with the given measurements [13, 23–27]. This typically is a non-linear least squares (LS) problem [28, 29], for which no closed-form solution is known. Due to the non-convexity of the problem, iterative solutions depend on the initialization. What complicates matters further is the fact that the acoustic measurements, such as DoAs, suffer from reverberation, which results in outliers that can spoil the geometry calibration process. To combat those, the iterative optimization is often embedded in a random sample consensus (RANSAC) method [30], which, however, significantly increases the computational load.

The approach presented here offers two innovations. First, we employ acoustic distance estimates, in addition to DoA measurements, which will solve the scale ambiguity of purely DoA-based geometry calibration and still renders clock synchronization at sample precision unnecessary. Compared to our previous approach presented in [31] which already utilized DoA and distance estimates in a two-stage manner, the approach proposed in the paper at hand combines both types of estimates directly in a common cost function.

In [32, 33], it has been shown how the distance between an acoustic source and a microphone array can be estimated from the coherent-to-diffuse power ratio (CDR), the ratio between the power of the coherent, and the diffuse part of the received audio signal. The authors employed Gaussian processes (GPs) to estimate the distance between a close pair of microphones and the acoustic source. This technique performed well if the GP was trained in the target environment but generalized poorly to new acoustic environments. Better generalization capabilities were achieved by deep neural network (DNN)-based acoustic distance estimation, where the network was exposed to many different acoustic environments during training [31]. However, this approach to distance estimation needs signal segments where a coherent source is active for a time around 1 s to work well. This requirement excludes impulsive source signals but is generally fulfilled by speech. Therefore, we consider speech as source signal but do not exclude other acoustic sources. In the contribution at hand, we build upon the DNN approach and further generalize it to perform better in the presence of directional sources.

The second contribution of this paper is the formulation of geometry calibration as a data set matching problem, similarly to [13], however, employing both distance and DoA estimates. Since data set matching can be efficiently realized, it greatly reduces the computational complexity of the task and thus the time it takes to estimate the geometry compared to a gradient-based optimization of a cost function. Moreover, we integrate the data set matching into an error-model-based re-weighting scheme and present a formal proof of convergence for it. The re-weighting scheme robustifies the geometry calibration process w.r.t. observations with large errors without the need of using a RANSAC. Additionally, a detailed experimental investigation of the proposed approach to geometry calibration is presented beside the mathematical analysis. Furthermore, the formulation as a data set matching problem allows the inference of the network's geometry even if it only consists of two sensor nodes, each equipped with at least three microphones which do not lie on a line.

The paper is organized as follows: In Section 2, the geometry calibration problem and the notation is summarized, followed by the description of the cost function we investigate for geometry estimation in Section 3. Subsequently, the distance estimation via DNNs is briefly described in Section 4. In Section 5, the experimental results are summarized before we end the paper by drawing some conclusions in Section 6.

2 Geometry calibration setup

We consider a WASN, where a set of sensor nodes is randomly placed in a reverberant environment (see Fig. 1). Note that we investigate geometry calibration in a 2-dimensional space; however, the extension to 3-dimensional space is in principle straight-forward.

We assume that the internal geometric arrangement of each node's microphone array is known and that all microphones making up an array are synchronously sampled, which we consider a realistic assumption. To be able to identify which DoA and distance estimates made by the different sensor nodes correspond to the same source signal, we further assume that a coarse time synchronization, i.e., a synchronization with an accuracy of a few tens of milliseconds, exists between the clocks of the different sensor nodes. This can be established, e.g., by NTP [17] or PTP [18]. We do, however, not require time synchronization at the precision of a few parts per million (ppm).

The WASN consists of L sensor nodes (red dots in Fig. 1), each equipped with a microphone array centered at positions $\boldsymbol{n}_l = \left[\, n_{l,x} \;\; n_{l,y} \,\right]^{\mathrm{T}}$ with an orientation θ_l, $l \in \{1, 2, \ldots, L\}$ relative to the global coordinate system, which is spanned by the depicted coordinate axes x and y. Here, θ_l corresponds to the rotation angle between the local coordinate system of the l-th node and the global coordinate system, i.e., the angle between the positive x-axes of the global and the local coordinate system (measured counterclockwise from the positive x-axis to the positive y-axis). The K acoustic sources (blue dots in Fig. 1) are at positions $\boldsymbol{s}_k = \left[\, s_{k,x} \;\; s_{k,y} \,\right]^{\mathrm{T}}$, $k \in \{1, 2, \ldots, K\}$. We assume that only one source is active at any given time. Note that the positions of the sensor nodes $\boldsymbol{n}_l$, their orientations θ_l, and the positions of the acoustic sources $\boldsymbol{s}_k$ are all unknown and will be estimated through a geometry calibration procedure from the observed acoustic source signals.

The geometry calibration task amounts to determining the set $\Omega_{\text{geo}} = \{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_L, \theta_1, \ldots, \theta_L\}$. Furthermore, all source positions are gathered in the set $\Omega_{\boldsymbol{s}} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K\}$, which will be estimated alongside geometry calibration. This results in the set of all unknowns $\Omega = \Omega_{\text{geo}} \cup \Omega_{\boldsymbol{s}}$.

Since a sensor node does not know its own position or orientation within the global coordinate system, all observations are given in the node's local coordinate system (see Fig. 2 for an illustration). In the following, the superscript (l) denotes that a quantity is measured in the local coordinate system of the l-th sensor node. Thus, the position of the k-th acoustic source, if expressed in the local coordinate system of the l-th sensor node, is denoted as $\boldsymbol{s}_k^{(l)} = \left[s_{k,x}^{(l)}, s_{k,y}^{(l)} \right]^{\mathrm{T}}$. Quantities without a superscript are measured in the global coordinate system. For example, $\boldsymbol{s}_k$ corresponds to the position of the k-th acoustic source described in the global coordinate system.

Each sensor node l, $l \in \{1, \ldots, L\}$, computes DoA estimates $\widehat{\varphi}_k^{(l)}$ and distance estimates $\widehat{d}_k^{(l)}$ to the acoustic source k, $k \in \{1, \ldots, K\}$, all w.r.t. the node's local coordinate system. Altogether, this results in $K \cdot L$ DoA estimates and $K \cdot L$ distance estimates available for geometry calibration.

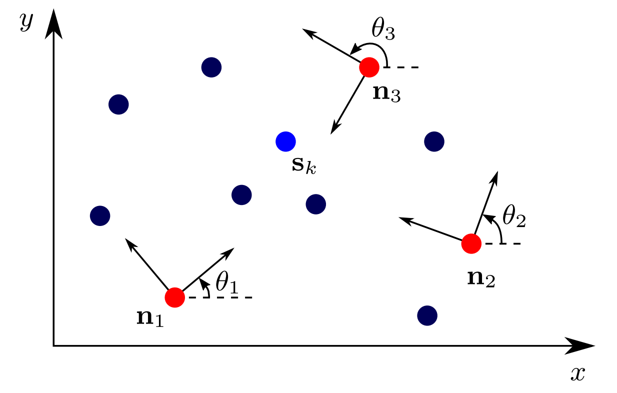

Fig. 1 Geometry calibration problem (red: sensor nodes; dark blue: acoustic sources; blue: source k; global coordinate system (x, y); local coordinate systems (x_l, y_l))

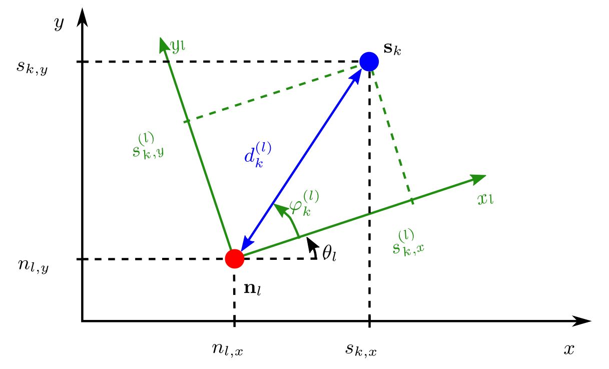

Fig. 2 Position of an acoustic source $\mathbf{s}_k$ within the global coordinate system (x, y) and local coordinate system (x_l, y_l) of node $\mathbf{n}_l$

3 Geometry calibration using DoAs and source node distances

To carry out geometry calibration, the given observations in the sensors' local coordinate systems have to be transferred to a common global coordinate system. Then, a cost function is defined that measures the fit of the transferred observations to an assumed geometry. The minimization of this cost function provides the positions and orientations of the sensor nodes, as well as the positions of the acoustic sources.

3.1 Development of a cost function

The position $\boldsymbol{s}_k^{(l)}$ of source k w.r.t. the local coordinate system of sensor node l is given by

$$\boldsymbol{s}_k^{(l)} = d_k^{(l)} \left[\cos\left(\varphi_k^{(l)}\right) \; \sin\left(\varphi_k^{(l)}\right) \right]^{\mathrm{T}}. \tag{1}$$

To project $\boldsymbol{s}_k^{(l)}$ into the global coordinate system, the following translation and rotation operation is applied:

$$\begin{aligned} \boldsymbol{s}_k &= \boldsymbol{R}(\theta_l)\boldsymbol{s}_k^{(l)} + \boldsymbol{n}_l & (2)\\ &= d_k^{(l)} \begin{bmatrix} \cos\left(\varphi_k^{(l)} + \theta_l\right) \\ \sin\left(\varphi_k^{(l)} + \theta_l\right) \end{bmatrix} + \boldsymbol{n}_l. & (3) \end{aligned}$$

Here,

$$\boldsymbol{R}(\theta_l) = \begin{bmatrix} \cos(\theta_l) & -\sin(\theta_l) \\ \sin(\theta_l) & \cos(\theta_l) \end{bmatrix} := \boldsymbol{R}_l \tag{4}$$

denotes the rotation matrix corresponding to the rotation angle θ_l.

If all distances and angles were perfectly known, all $\boldsymbol{s}_k^{(l)}$ would map to a unique position $\boldsymbol{s}_k$. Hence, the geometry can be inferred by minimizing the deviation of the projected source positions from an assumed position $\boldsymbol{s}_k$ by minimizing the LS cost function $J(\Omega)$:

$$\widehat{\Omega} = \underset{\Omega}{\operatorname{argmin}} \underbrace{\sum_{l=1}^{L}\sum_{k=1}^{K} \left\| \boldsymbol{s}_k - \left(\boldsymbol{R}_l \boldsymbol{s}_k^{(l)} + \boldsymbol{n}_l\right)\right\|_2^2}_{:=J(\Omega)}, \tag{5}$$

with $\|\cdot\|_2$ denoting the Euclidean norm. Note that at least $K{=}2$ spatially different acoustic source positions have to be observed to arrive at an (over-)determined system of equations which is defined by $\boldsymbol{s}_k = \boldsymbol{R}_l\boldsymbol{s}_k^{(l)} + \boldsymbol{n}_l$ with $l \in \{1, \ldots, L\}$ and $k \in \{1, \ldots, K\}$.

There exists no closed-form solution for the non-linear optimization problem in (5). Thus, (5) has to be solved by an iterative optimization algorithm, e.g., by Newton's method as proposed in [23] or by gradient descent.

Prior works, e.g., [23], have shown that the iterative optimization strongly depends on the initial values. Furthermore, the optimization is computationally demanding and, depending on the number of observed acoustic source positions, very time consuming, which limits its usefulness for WASNs with typically limited computational resources. In the following, we will present a computationally much more reasonable approach.

3.2 Geometry calibration by data set matching

We now interpret the relative acoustic source positions (see (1)) as the vertices of a rigid body. Matching the rigid body shapes as observed by the different sensor nodes will result in an efficient way for geometry calibration as described in [13]. In the following, we shortly recapitulate the concept of efficient geometry calibration based on data set matching [34, 35]. Let

$$\boldsymbol{S}^{(l)} = \left[\boldsymbol{s}_1^{(l)} \; \cdots \; \boldsymbol{s}_K^{(l)} \right]. \tag{6}$$

be the matrix of all K source positions, as measured in the local coordinate system of sensor node l. Similarly, let $\boldsymbol{S}$ be the same matrix of source positions, but now measured in the global coordinate system. The dispersion matrix $\boldsymbol{D}_l$ is defined as follows [35]:

$$\boldsymbol{D}_l = \frac{1}{K}\left(\boldsymbol{S}^{(l)} - \bar{\boldsymbol{s}}^{(l)}\mathbf{1}^{\mathrm{T}}\right)\boldsymbol{W}_l\left(\boldsymbol{S} - \bar{\boldsymbol{s}}\mathbf{1}^{\mathrm{T}}\right)^{\mathrm{T}}, \tag{7}$$

where $\mathbf{1}$ denotes a vector of all ones. $\boldsymbol{W}_l$ is a diagonal matrix with $(W_l)_{k,k} = w_{kl}$, where $(\cdot)_{i,j}$ denotes the i-th row and j-th column element of a matrix. $\bar{\boldsymbol{s}}^{(l)}$ corresponds to the centroid of the observations made by sensor node l and $\bar{\boldsymbol{s}}$ is the centroid of the source positions expressed in the global coordinate system:

$$\bar{\boldsymbol{s}}^{(l)} = \frac{\sum_{k=1}^{K} w_{kl}\boldsymbol{s}_k^{(l)}}{\sum_{k=1}^{K} w_{kl}} \quad \text{and} \quad \bar{\boldsymbol{s}} = \frac{\sum_{k=1}^{K} w_{kl}\boldsymbol{s}_k}{\sum_{k=1}^{K} w_{kl}}. \tag{8}$$

The weights w_{kl} will be introduced in Section 3.3 to control the impact of an individual observation $\boldsymbol{s}_k^{(l)}$ on the geometry estimates.

Carrying out a singular value decomposition (SVD) of the dispersion matrix gives $\boldsymbol{D}_l{=}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$. The estimate $\widehat{\boldsymbol{R}}_l$ of the rotation matrix is then given by [34, 35]

$$\widehat{\boldsymbol{R}}_l = \boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}, \tag{9}$$

and the orientation of the corresponding sensor node by:

$$\widehat{\theta}_l = \operatorname{arctan2}\left(\left(\widehat{\boldsymbol{R}}_l\right)_{1,1}, \left(\widehat{\boldsymbol{R}}_l\right)_{2,1}\right). \tag{10}$$

Here, arctan2 is the four-quadrant arc tangent. Thus, the l-th sensor node position estimate $\widehat{\boldsymbol{n}}_l$ in the reference coordinate system is given by

$$\widehat{\boldsymbol{n}}_l = \bar{\boldsymbol{s}} - \widehat{\boldsymbol{R}}_l\bar{\boldsymbol{s}}^{(l)}. \tag{11}$$

Note that the described data set matching procedure corresponds to minimizing the following cost function [34]:

$$J\left(\boldsymbol{n}_l, \boldsymbol{R}_l\right) = \sum_{k=1}^{K} w_{kl} \left|\left|\boldsymbol{s}_k - \left(\boldsymbol{R}_l \boldsymbol{s}_k^{(l)} + \boldsymbol{n}_l\right)\right|\right|_2^2. \tag{12}$$

3.3 Geometry calibration by iterative data set matching

We now generalize the findings of the last section to an arbitrary number L of sensor nodes. Moreover, we consider the source positions as additional unknowns. The resulting cost function

$$J(\Omega) = \sum_{l=1}^{L} \sum_{k=1}^{K} w_{kl} \left|\left|\boldsymbol{s}_k - \left(\boldsymbol{R}_l \boldsymbol{s}_k^{(l)} + \boldsymbol{n}_l\right)\right|\right|_2^2 \tag{13}$$

is optimized by alternating between the estimation of the set of source positions $\Omega_{\boldsymbol{s}}$ and the estimation of the sensor node parameters Ω_{geo}.

Starting from an initial set of source positions $\Omega_{\boldsymbol{s}}$, the geometry Ω_{geo} can be determined by optimizing (12) for each sensor node $l \in \{1, \ldots, L\}$ by data set matching as outlined in the last section. Note that the estimated positions are given relative to a reference coordinate system. The origin and orientation of this reference coordinate system is a result of the calibration process.

Given a geometry Ω_{geo} the positions $\boldsymbol{s}_k$ can be estimated for each acoustic source $k \in \{1, \ldots, K\}$ via:

$$\hat{\boldsymbol{s}}_k = \underset{\boldsymbol{s}_k}{\text{argmin}} \sum_{l=1}^{L} w_{kl} \left|\left|\boldsymbol{s}_k - \left(\boldsymbol{R}_l \boldsymbol{s}_k^{(l)} + \boldsymbol{n}_l\right)\right|\right|_2^2. \tag{14}$$

For this, a closed-form solution exists, which is given by

$$\hat{\boldsymbol{s}}_k = \frac{\sum\limits_{l=1}^{L} w_{kl} \left(\boldsymbol{R}_l \boldsymbol{s}_k^{(l)} + \boldsymbol{n}_l\right)}{\sum\limits_{l=1}^{L} w_{kl}}. \tag{15}$$

What remains is to describe how the weights w_{kl} are chosen. They should reflect how well the observations $\boldsymbol{s}_k^{(l)}$ fit to the model specified by $\widehat{\Omega}_{\text{geo}}$ and $\widehat{\Omega}_{\boldsymbol{s}}$. This can be achieved by setting

$$w_{kl} = \frac{1}{\left|\left|\hat{\boldsymbol{s}}_k - \left(\widehat{\boldsymbol{R}}_l \boldsymbol{s}_k^{(l)} + \hat{\boldsymbol{n}}_l\right)\right|\right|_2}. \tag{16}$$

With these weights and the ideas of [36], (13) can be interpreted as an iteratively re-weighted least squares (IRLS) algorithm [37] which minimizes the following sum of Euclidean distances:

$$\widehat{\Omega} = \underset{\Omega}{\text{argmin}} \sum_{l=1}^{L} \sum_{k=1}^{K} \left|\left|\boldsymbol{s}_k - \left(\boldsymbol{R}_l \boldsymbol{s}_k^{(l)} + \boldsymbol{n}_l\right)\right|\right|_2. \tag{17}$$

Consequently, the resulting optimization problem is less sensitive to outliers than the optimization problem in (5).

3.4 Implementation details

Algorithm 1 summarizes the iterative data set matching used for geometry calibration. In the beginning the set of observations $\mathcal{S}^{(1)} = \left\{\boldsymbol{s}_1^{(1)}, \boldsymbol{s}_2^{(1)}, \ldots, \boldsymbol{s}_K^{(1)}\right\}$ made by sensor node 1 is used as initial estimate of the acoustic sources' position set $\widehat{\Omega}_{\boldsymbol{s}}$. Experiments on the convergence behavior have shown that the effect of the choice of the sensor node, whose observations are used for initialization, is negligible (see Section 5.2). Due to the fact that at this point no statement can be made about the quality of the observations $\boldsymbol{s}_k^{(l)}$, the initial weights are all set to one: $w_{kl} = 1; \forall k, l$.

Subsequently, a first estimate of the geometry $\widehat{\Omega}_{\text{geo}}$ can be derived by data set matching (line 3) utilizing $\widehat{\Omega}_{\boldsymbol{s}}$ as reference source positions. Then, $\widehat{\Omega}_{\text{geo}}$ is used to estimate the sources' positions $\widehat{\Omega}_{\boldsymbol{s}}$ (line 4) based on (15) with the weights still left as above. In the next iterations, the weights are chosen as described in (16). The iterative weighted data set matching procedure, i.e., lines 3–5 in Algorithm 1, is repeated until $\widehat{\Omega}_{\text{geo}}$ and $\widehat{\Omega}_{\boldsymbol{s}}$ converge. A detailed analysis of the convergence behavior of this part of the algorithm can be found in the Appendix.

Although outliers are already addressed by the weights w_{kl} to some extent, they can still have a detrimental influence on the results of the iterative optimization process if the corresponding errors are very large. Therefore, after convergence, the iterative weighted data set matching procedure is repeated again (lines 7–12); however, only on that subset of observations $\mathcal{S}_{\text{fit}}$ that best fits to the model defined by the current estimates $\widehat{\Omega}_{\text{geo}}$ and $\widehat{\Omega}_{\boldsymbol{s}}$.

There are two criteria that describe how well the observations $\boldsymbol{s}_k^{(l)}$ made by sensor node l fit to the model specified by $\widehat{\Omega}_{\text{geo}}$ and $\widehat{\Omega}_{\boldsymbol{s}}$. First, there are the distances

Algorithm 1: Iterative Geometry Calibration Using Data Set Matching

```
   Data: $\mathcal{S} = \left\{\boldsymbol{s}_1^{(1)}, \boldsymbol{s}_2^{(1)}, \ldots, \boldsymbol{s}_K^{(1)}, \boldsymbol{s}_1^{(2)}, \ldots, \boldsymbol{s}_K^{(L)}\right\}$;
1  Init: $\widehat{\Omega}_{\boldsymbol{s}} = \left\{\boldsymbol{s}_1^{(1)}, \boldsymbol{s}_2^{(1)}, \ldots, \boldsymbol{s}_K^{(1)}\right\}$,
         $\Omega_w = \{w_{11}, \ldots, w_{KL}\} = \{1, \ldots, 1\}$;
2  repeat
3      $\widehat{\Omega}_{\text{geo}} = \texttt{DSM\_Calib}(\mathcal{S}, \widehat{\Omega}_{\boldsymbol{s}}, \Omega_w)$; {Eq. (12)}
4      $\widehat{\Omega}_{\boldsymbol{s}} = \texttt{SRC\_Loc}(\mathcal{S}, \widehat{\Omega}_{\text{geo}}, \Omega_w)$; {Eq. (15)}
5      $\Omega_w = \texttt{Get\_Weights}(\mathcal{S}, \widehat{\Omega}_{\boldsymbol{s}}, \widehat{\Omega}_{\text{geo}})$; {Eq. (16)}
6  until Convergence;
7  repeat
8      $\mathcal{S}_{\text{fit}} = \texttt{Fit\_Select}(\widehat{\Omega}_{\boldsymbol{s}}, \widehat{\Omega}_{\text{geo}}, \mathcal{S})$; {Eq. (20)}
9      $\widehat{\Omega}_{\text{geo}} = \texttt{DSM\_Calib}(\mathcal{S}_{\text{fit}}, \widehat{\Omega}_{\boldsymbol{s}}, \Omega_w)$; {Eq. (12)}
10     $\widehat{\Omega}_{\boldsymbol{s}} = \texttt{SRC\_Loc}(\mathcal{S}, \widehat{\Omega}_{\text{geo}}, \Omega_w)$; {Eq. (15)}
11     $\Omega_w = \texttt{Get\_Weights}(\mathcal{S}, \widehat{\Omega}_{\boldsymbol{s}}, \widehat{\Omega}_{\text{geo}})$; {Eq. (16)}
12 until Convergence;
   Result: $\widehat{\Omega}_{\text{geo}}$;
```

between $\boldsymbol{s}_k^{(l)}$ and the source position estimates $\boldsymbol{s}_k^{(o)}$, $o \in \{1, \dots, L\} \backslash \{l\}$, made by the other sensor nodes:

$$\epsilon_k(l, o) = \left|\left|\left(\widehat{\boldsymbol{R}}_l \boldsymbol{s}_k^{(l)} + \hat{\boldsymbol{n}}_l\right) - \left(\widehat{\boldsymbol{R}}_o \boldsymbol{s}_k^{(o)} + \hat{\boldsymbol{n}}_o\right)\right|\right|_2. \tag{18}$$

Second, there is the distance between the observations after being projected and the estimated source position measured in the global coordinate system:

$$\sigma_k(l) = \left|\left|\hat{\boldsymbol{s}}_k - \left(\widehat{\boldsymbol{R}}_l \boldsymbol{s}_k^{(l)} + \hat{\boldsymbol{n}}_l\right)\right|\right|_2. \tag{19}$$

Note that the choice of $\epsilon_k(l, o)$ and $\sigma_k(l)$ is motivated by the fact that all relative source positions observed by the single sensor nodes would map on the same position in the global coordinate system if the observations are perfect.

Combining the two criteria results in the function

$$C_k(l) = \sigma_k(l) + \frac{1}{L-1} \sum_{\substack{o=1 \\ o \neq l}}^{L} \epsilon_k(l, o), \tag{20}$$

used for the selection of $\mathcal{S}_{\text{fit}}$. The distance and DoA measurements of source k made by a node l are included in $\mathcal{S}_{\text{fit}}$ only if the resulting relative source position belongs to the best γ measurements made by a node l. With $C_k(l)$ outliers can be identified based on the fact that they do not align well with the source position estimates of the other nodes for the current geometry.

In principle, this fitness selection could also be integrated in the first iterative data set matching rounds (lines 3–5). However, initial experiments have shown that this may lead to a degradation of performance if the number of observed source positions K is small. This can be explained by the fact that observations are discarded based on a model which is still not converged.

4 Acoustic distance estimation

To gather distance and, respectively, scaling information that can be used for geometry calibration, we propose to utilize the DNN-based distance estimator which we introduced in [31]. This distance estimator shows state-of-the-art performance and good generalization capabilities to different acoustic environments. In the following, we just concentrate on an adaptation of the distance estimator to directional sources and refer to [31] for a detailed description.

Our approach to acoustic distance estimation considers a microphone pair recording a signal $x(t)$ emitted by a single acoustic source. The reverberant signal, being captured by the ν-th microphone, $\nu \in \{1, 2\}$, is modeled as follows [32]:

$$\begin{aligned} y_\nu(t) &= h_\nu(t) * x(t) + v_\nu(t) \\ &= \underbrace{h_{\nu,e}(t) * x(t)}_{c_\nu(t)} + \underbrace{h_{\nu,\ell}(t) * x(t) + v_\nu(t)}_{r_\nu(t)}, \end{aligned} \tag{21}$$

with $v_\nu(t)$ corresponding to white sensor noise and $h_\nu(t)$ corresponding to the room impulse response which models the sound propagation from the source to the ν-th microphone. The $*$ operator denotes a convolution. $h_\nu(t)$ can be divided into $h_{\nu,e}(t)$ modeling the direct path and the early reflections and $h_{\nu,\ell}(t)$ modeling the late reflections. Thus, $y_\nu(t)$ can be split up into a coherent component $c_\nu(t)$ which corresponds to the direct path and the early reflections and a diffuse component $r_\nu(t)$ produced by the late reflections and the sensor noise.

In [32] it was shown that the CDR, i.e., the power ratio of the coherent signal component $c_\nu(t)$ to diffuse signal component $r_\nu(t)$, is related to the distance between the microphone pair and the acoustic source (the larger the distance the smaller the value of the CDR). The DNN-based distance estimator utilizes a time-frequency representation of the CDR as an input feature.

Due to the large effort needed to measure room impulse responses (RIRs) in various acoustic environments, we here stick to synthetic RIRs for the training of the distance estimator, using the RIR generator of [38]. However, there are a lot of simplifying assumptions for the simulation of RIRs. For example, the room is modeled as a cuboid, and an omnidirectional characteristic is typically assumed for the acoustic sources and microphones.

Especially the omnidirectional characteristic of the acoustic sources is a large deviation from reality, because a real acoustic source, like a speaker, typically exhibits directivity. While an omnidirectional source emits sound waves with equal power in all directions, a directional source emits most of the power into one direction. In both cases, the sound waves are reflected multiple times on the surfaces of the room which mainly causes the late reflections and accumulates to $h_{\nu,\ell}$. Hence, a directional source pointing towards a microphone array causes a less diffuse signal compared to an omnidirectional source that is assumed in the simulated RIRs. Consequently, a distance estimator trained with simulated RIRs and applied to recordings of directional sources, pointing towards the microphone array, would exhibit a systematic error and underestimates the distance. Furthermore, a directional source may cause a more diffuse signal compared to an omnidirectional source if it does not point towards a microphone array, causing a systematic overestimation of the distance. However, this case is not further investigated as such recording conditions are not included in the MIRD database [39] which is used in the experimental section.

We approach this mismatch by applying a recently proposed direct-to-reverberant ratio (DRR) data augmentation technique [40]. The DRR is defined as

$$\eta_\nu = \frac{\sum_t h_{\nu,e}^2(t)}{\sum_t h_{\nu,\ell}^2(t)}. \tag{22}$$

Considering (21), it is obvious that CDR and DRR are equivalent [41] if the influence of the sensor noise is negligible. Consequently, an augmentation of the DRR results into an augmentation of the CDR.

Therefore, during training, a scalar gain α is applied to $h_{\nu,e}(t)$ which contains the direct path and the early reflections of the RIRs. To avoid discontinuities within the RIR caused by the scaling, a window $w_d(t)$ is employed to smooth the product $\alpha \cdot h_{\nu,e}(t)$:

$$\bar{h}_{\nu,e}(t) = w_d(t) \cdot \alpha \cdot h_{\nu,e}(t) + (1 - w_d(t)) \cdot h_{\nu,e}(t). \tag{23}$$

Hereby, $w_d(t)$ corresponds to a Hann window of 5 ms size, which is centered around the time delay t_d corresponding to the direct path. t_d is identified by the location of the maximum of $|h_\nu(t)|$.

Due to the fact that the directivity of the acoustic source is unknown in general there is also no knowledge how α has to be chosen to adapt the simulated RIRs to the real scenario. Nevertheless, it is known that the DRR of the simulated RIRs has to be increased if a directional source pointing towards the center of the microphone pair is considered. Thus, $\alpha \sim \mathcal{U}(1, \alpha_{max})$ is used, where α_{max} corresponds to the fixed upper limit of α and $\sim \mathcal{U}(\min, \max)$ denotes to uniformly draw a value from the interval $[\min, \max]$.

Furthermore, the DRR is only manipulated with probability $Pr(aug)$. Hence, beside manipulated examples, also examples that are not manipulated are presented to the DNN during training. The non-manipulated examples should ease the process of learning that examples being manipulated with different scaling factors α belong to the same distance.

5 Experimental results

In this section, the proposed approach to geometry calibration is evaluated. First, the adaptation of the DNN-based acoustic distance estimation method to directional sources is examined. For deeper insights into acoustic distance estimation see [31]. Afterwards, the proposed approach to geometry calibration is investigated based on simulations of the considered scenario.

5.1 Acoustic distance estimation

In the following, the adaptation of the DNN-based distance estimator to directional sources is evaluated on the MIRD database [39]. This database consists of measured RIRs for multiple source positions on an angular grid at a distance of 1 m and 2 m. The measurements took place in a 6 m×6 m×2.4 m room with a configurable reverberation time T_{60}. From the data we used, the two subsets corresponding to T_{60}=360 ms and T_{60}=610 ms, considering the central microphone pair with inter microphone distance equal to 8 cm.

The setups of the MIRD database are limited w.r.t. the number of source and sensor positions. Nevertheless, the experimental data is sufficient to proof that the approach works for directional acoustic sources and not only on simulated audio data of omnidirectional sources. We refer to [31] for a detailed investigation of a wider range of considered setups using simulated data.

As described in Section 4, the distance estimator is trained utilizing RIRs which are simulated using the implementation of [38]. The training set consists of 100,000 source microphone pair constellations whereby the properties of the considered room and the placement of the microphone pair and acoustic source is randomly drawn for each of these constellations. Table 1 summarizes the corresponding probability distributions. We first draw the position of the microphone pair and then place the acoustic source relative to this position at the same height using the distance d and the DoA φ.

The RIRs are used to reverberate clean speech signals from the TIMIT database [42]. During training, these speech probes are randomly drawn from the database. For the evaluation of the distance estimator on the MIRD database, we utilized R=100 speech probes which were randomly drawn from the TIMIT database and then reverberated by each of the RIRs.

In the following, the configuration and training scheme of the distance estimator are explained. We employ 1 s long speech segments to calculate the CDR which results in a feature map that is passed to the DNN. The short-time Fourier transform (STFT), which is needed to estimate the CDR, utilizes a Blackman window of size 25 ms, and a frame shift of 10 ms. The CDR is calculated for frequencies between 125Hz and 3.5 kHz, which corresponds to the frequency range, where speech has significant power.

Table 2 shows the architecture of the DNN used for distance estimation. The estimator is trained using Adam [43] with a mini-batch size of B=32 and a learning rate of $3{\cdot}10^{-4}$ for 500,000 iterations. Besides, the maximum DRR

Table 1 Description of the training set of the distance estimator used on the MIRD database

Room width	$r_w \sim \mathcal{U}(5\,\text{m}, 7\,\text{m})$
Room length	$r_l \sim \mathcal{U}(5\,\text{m}, 7\,\text{m})$
Room height	$r_h \sim \mathcal{U}(2.2\,\text{m}, 2.6\,\text{m})$
Reverberation time	$T_{60} \sim \mathcal{U}(250\,\text{ms}, 700\,\text{ms})$
Position of the mic. pair	$n_x \sim \mathcal{U}(0.5\,\text{m}, r_w - 0.5\,\text{m})$
	$n_y \sim \mathcal{U}(0.5\,\text{m}, r_l - 0.5\,\text{m})$
	$n_z \sim \mathcal{U}(1\,\text{m}, r_h - 1\,\text{m})$
Orientation of the mic. pair	$\theta \sim \mathcal{U}(0, 2\pi)$
Distance	$d \sim \mathcal{U}(0.3\,\text{m}, 3\,\text{m})$
DoA	$\varphi \sim \mathcal{U}(0, 2\pi)$

Table 2 Architecture of the DNN used for distance estimation on the MIRD database

Block	**Output shape**
CDR	$B \times 1 \times F \times M$
2 × Conv2d(3 × 3; 16)	$B \times 16 \times F \times M$
MaxPool2d(2 × 1)	$B \times 16 \times \lfloor F/2 \rfloor \times M$
2 × Conv2d(3 × 3; 32)	$B \times 32 \times \lfloor F/2 \rfloor \times M$
MaxPool2d(2 × 2)	$B \times 32 \times \lfloor F/4 \rfloor \times \lfloor M/2 \rfloor$
2 × Conv2d(3 × 3; 64)	$B \times 64 \times \lfloor F/4 \rfloor \times \lfloor M/2 \rfloor$
MaxPool2d(2 × 1)	$B \times 64 \times \lfloor F/8 \rfloor \times \lfloor M/2 \rfloor$
2 × Conv2d(3 × 3; 128)	$B \times 128 \times \lfloor F/8 \rfloor \times \lfloor M/2 \rfloor$
MaxPool2d(2 × 2)	$B \times 128 \times \lfloor F/16 \rfloor \times \lfloor M/4 \rfloor$
Conv2d(3 × 3; 256)	$B \times 256 \times \lfloor F/16 \rfloor \times \lfloor M/4 \rfloor$
MaxPool2d(2 × 1)	$B \times 256 \times \lfloor F/32 \rfloor \times \lfloor M/4 \rfloor$
Reshape	$B \times 256 \cdot \lfloor F/32 \rfloor \times \lfloor M/4 \rfloor$
3 × Conv1d(3; 512)	$B \times 512 \times \lfloor M/4 \rfloor$
2 × GRU(512)	$B \times 512$
$\mathrm{fc}_{\mathrm{ReLU}}(512)$	$B \times 512$
$\mathrm{fc}_{\mathrm{Softmax}}(D_c)$	$B \times D_c$

Each conv{1,2}d layer includes ReLU as activation and batch normalization. Only the last output vector of the gated recurrent unit (GRU) is forwarded to the fully connected layers (fc). Dropout with a dropout probability equal to 0.5 is used on the output of all GRU and fully connected layers except for the last fully connected layer. D_c denotes the number of distance classes (distance estimation is formulated as a classification problem; see [31].). For simplicity, we write, e.g., $\lfloor M/4 \rfloor$ instead of $\lfloor\lfloor M/2 \rfloor/2\rfloor$

augmentation factor α_{max} is chosen to be equal to 3. After training, we utilize the best performing checkpoint w.r.t. the mean-absolute error (MAE) of the distance estimates on an independent validation set.

The influence of the DRR manipulation probability $Pr(aug)$ can be seen in Table 3. Thereby, the MAE

$$e_d = \frac{1}{2 \cdot A \cdot R} \sum_{c=1}^{2} \sum_{a=1}^{A} \sum_{r=1}^{R} |d(c,a) - \widehat{d}_r(c,a)| \qquad (24)$$

is used as metric. Here, $d(1,a)$=1 m and $d(2,a)$=2 m correspond to the ground truth distance at DoA-candidate a. $\widehat{d}_r(c,a)$ denotes the corresponding estimate using the r-th speech sample and A the number of DoA in the angular grid of the MIRD database. Furthermore, results for distance estimation on a simulated version of the RIRs of the MIRD database with omnidirectional sources are provided (see Table 3).

Table 3 MAE e_d/ m on the MIRD database and the corresponding simulated RIRs

	sim. RIRs		**MIRD**	
***Pr*(*aug*)**	**360ms**	**610ms**	**360ms**	**610ms**
0	0.18	0.23	0.45	0.53
0.5	0.24	0.3	0.32	0.32
0.8	0.24	0.29	0.25	0.26
0.9	0.26	0.33	0.28	0.26
1	0.28	0.33	0.26	0.24

Without DRR augmentation, i.e., for $Pr(aug)$=0, the distance estimation error is large compared to the error on simulated RIRs. This can be explained by the systematic error resulting from the fact that the simulated RIRs used during the training include more diffuse signal parts than the recorded RIR. With DRR augmentation the error of the distance estimates on the MIRD database can be reduced and the best performance is achieved if the DRR of all examples is manipulated during training. However, DRR augmentation makes the learning process more difficult, which increases the error on the simulated RIRs.

5.2 Geometry calibration

To evaluate the proposed approach to geometry calibration, we generated a data set consisting of G=100 simulated scenarios. Thereby, each scenario corresponds to a WASN with L=4 sensor nodes. Furthermore, each scenario contains acoustic sources at a fixed amount of K=100 spatially independent positions within the room. This number can be justified by the fact that in realistic environments, e.g., living rooms, acoustic sources like speakers will move over time such that the amount of observed acoustic source positions will also grow over time. All rooms have a random width $r_w \sim \mathcal{U}(6\,\mathrm{m}, 7\,\mathrm{m})$, random length $r_l \sim \mathcal{U}(5\,\mathrm{m}, 6\,\mathrm{m})$, and a fixed height r_h of 3 m. In the experiments, we investigate reverberation times T_{60} from the set $\{300\,\mathrm{ms}, 400\,\mathrm{ms}, 500\,\mathrm{ms}, 600\,\mathrm{ms}\}$.

Both, the nodes and the acoustic sources, are placed at a height of 1.4 m, whereby the sensor nodes are equipped with a circular array with six microphones and a diameter of 5 cm. The way how the sensor nodes and the acoustic sources are placed within the room is exemplarily shown in Fig. 3.

We assume that at each of the possible K=100 source positions, a 1 s long speech signal is emitted, whereby the speech signals are randomly drawn from the TIMIT database. The speech samples are reverberated by RIRs gathered from the RIR generator of [38]. Subsequently, the reverberant signals are used for distance and DoA estimation.

We employ the convolutional recurrent neural network (CRNN) which we proposed in [31] to compute the distance estimates used for geometry calibration. Feature extraction, training set, and training scheme mainly coincide with the ones described in Section 5.1. The description of the corresponding training set which consists of 10,000 source node constellations can be found in Table 4. During training, DRR augmentation is used with a manipulation probability of $Pr(aug)$=0.5.

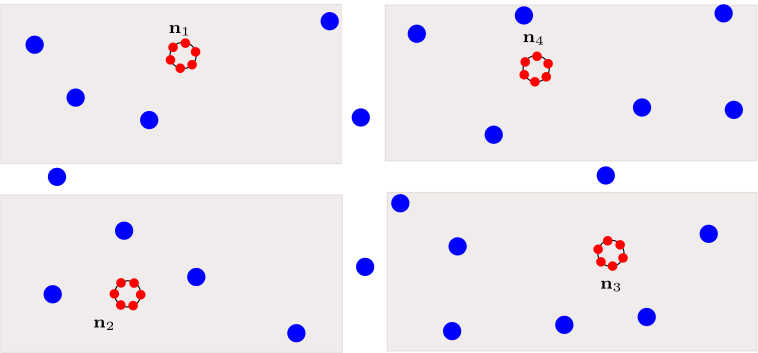

Fig. 3 Simulated setup; red: microphones; blue: acoustic sources; gray area: possible area to randomly place sensor nodes (microphone arrays); all sensor nodes and acoustic sources have a minimum distance of 0.1 m to the closest wall; 1 m spacing between the gray areas

We take the three microphone pairs formed by the opposing microphones of the considered circular microphone array for distance estimation. The CDR is estimated for each of these microphone pairs and the three resulting feature maps are jointly passed to the CRNN.

DoA estimation is done using the complex Watson kernel method introduced in [44] , where it was shown that this estimator is competitive to state-of-the-art estimators. The considered DoA candidates have an angular resolution of 1ř and the concentration parameter of the complex Watson probability density function is chosen to be κ=5.

The fitness selection contained in our approach to geometry calibration always selects the best 50% relative source positions for each sensor node.

Figures 4 and 5 show the cumulative distribution function (CDF) of the distance and DoA estimation errors. The majority of distance and DoA estimates exhibits only small errors, so in general there will be enough reliable estimates for geometry calibration. But in both cases, there is also a non-negligible amount of estimates exhibiting large errors which have to be considered as outliers. It can also be observed that the amount of outliers increases with increasing reverberation time T_{60}. We refer to [31, 44] for a comparison of the used estimators to alternative estimators.

Table 4 Description of the training set of the distance estimator used for geometry calibration

Room width	$r_w \sim \mathcal{U}(5\text{ m}, 7\text{ m})$
Room length	$r_l \sim \mathcal{U}(5\text{ m}, 7\text{ m})$
Room height	$r_h = 3\text{ m}$
Reverberation time	$T_{60} \sim \mathcal{U}(250\text{ ms}, 700\text{ ms})$
Position of the node	$n_x \sim \mathcal{U}(0.5\text{ m}, r_w - 0.5\text{ m})$
	$n_y \sim \mathcal{U}(0.5\text{ m}, r_l - 0.5\text{ m})$
	$n_z = 1.4\text{ m}$
Orientation of the node	$\theta \sim \mathcal{U}(0, 2\pi)$
Distance	$d \sim \mathcal{U}(0.3\text{ m}, 6\text{ m})$
DoA	$\varphi \sim \mathcal{U}(0, 2\pi)$

After the geometry calibration process is started, more and more observed relative source positions $\boldsymbol{s}_k^{(l)}$ will become available. The resulting effect on the geometry calibration results can be seen in Fig. 6, which displays the MAE of the sensor nodes' position

$$e_p = \frac{1}{G \cdot L} \sum_{g=1}^{G} \sum_{l=1}^{L} \left|\left|\boldsymbol{n}_{l,g} - \widehat{\boldsymbol{n}}_{l,g}\right|\right|_2 \tag{25}$$

and orientation

$$e_o = \frac{1}{G \cdot L} \sum_{g=1}^{G} \sum_{l=1}^{L} \left| \angle \left(e^{j(\theta_{l,g} - \widehat{\theta}_{l,g})} \right) \right|, \tag{26}$$

where $\angle(\cdot)$ denotes the phase of a complex-valued number. Further, $\boldsymbol{n}_{l,g}$ and $\theta_{l,g}$ are the ground truth values of the location parameters of the l-th node in the g-th scenario and $\widehat{\boldsymbol{n}}_{l,g}$ and $\widehat{\theta}_{l,g}$ denote the corresponding estimates. Note that the geometry estimates are projected into the coordinate system of the ground truth geometry using data set matching to align the sensor node positions before the errors are calculated.

Figure 6 shows that the geometry estimation error gets smaller when more source positions have been observed and thus more relative source position estimates exhibiting a small error are available. Hence, the estimate of the geometry will improve over time. However, reasonable results can already be achieved with a small amount of observed source positions. This especially holds for scenarios with small reverberation times T_{60} where the estimates of the relative source positions are less error-prone.

In addition to the MAE of the geometry estimates, the distribution of the corresponding error is displayed in

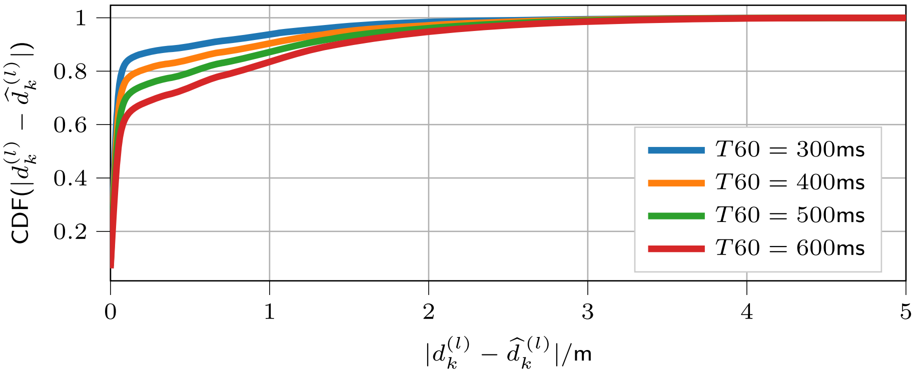

Fig. 4 CDF of the distance estimation error

Figs. 7 and 8 for K=20 and K=100 observed source positions. For a small number of observed source positions, i.e., K=20, the majority of node position and node orientation estimates shows acceptably small errors. As can be seen, there are still outliers exhibiting large errors, despite the used error-model-based re-weighting method and the fitness selection method.

If more source positions are observed, e.g., K=100, the probability increases that a sufficient amount of good relative source position estimates is available, thus improving the average calibration accuracy and also decreasing the number of outliers.

Table 5 shows the influence of the individual outlier rejection and error handling steps of our approach to geometry calibration, namely the weighting in data set matching (WLS), the weighting in source localization ($\mathrm{WLS_{SRC}}$), and the fitness selection (Select). If all weights are set to w_{kl}=1; $\forall\, k, l$, and fitness selection is omitted, the geometry estimates are clearly worse compared to the other cases depicted in the table. Introducing weighting factors in data set matching and source localization improves the results. However, the experiment with active data selection reveals that the weighting is not powerful enough to completely suppress the detrimental effect of outliers, which can only be achieved by removing these outliers from the processed data via fitness selection.

Figures 9 and 10 show the effect of fitness selection on the distribution of the DoA and distance estimation errors. Fitness selection causes larger errors to occur less frequently for both quantities, removing a large portion of the outliers. This especially holds for the distance estimates.

These outliers are often caused by strong early reflections of sound on surfaces in the room, e.g., when a sensor node is placed near to a wall, resulting in poor distance and DoA estimates. However, outliers can also occur if a source is too close to a sensor node, i.e., the far-field assumption for DoA estimation is not met, or the distance between a sensor node and an acoustic source is too large which leads to a challenging situation for distance estimation. Because of the large number of possible reasons for outliers in the DoA and distance estimates, we refer the reader to the relevant literature for a more detailed discussion [31, 44, 45].

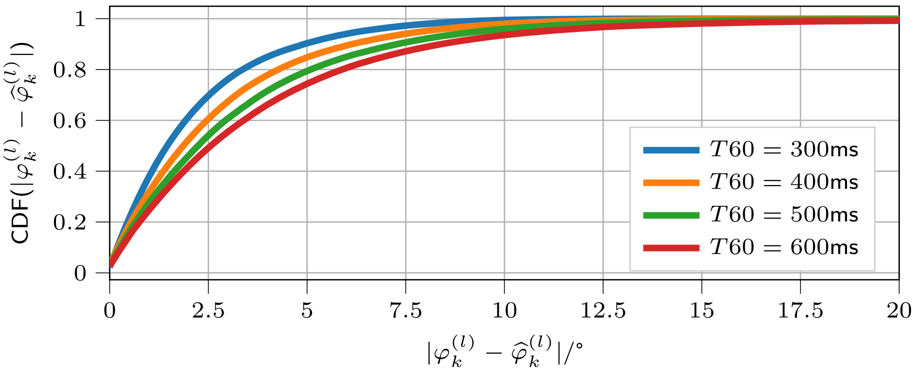

Fig. 5 CDF of the DoA estimation error

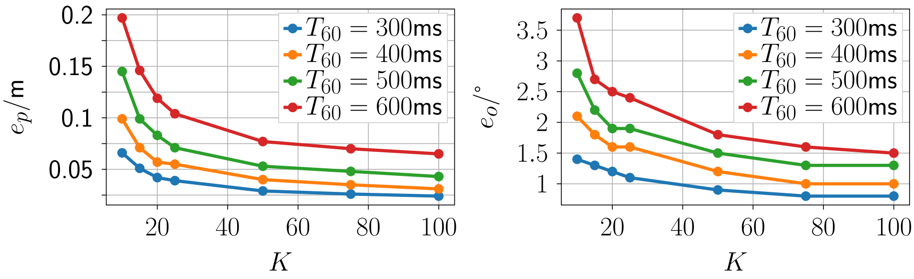

Fig. 6 Influence of number of source positions on calibration performance

The convergence behavior of the sensor nodes' positions is shown in Fig. 11 based on the CDF of the average spread of the sensor node position estimates

$$\zeta_{\boldsymbol{n}_l} = \frac{1}{I}\sum_{i=1}^{I} \left|\left|\widehat{\boldsymbol{n}}_{l,i} - \mu_{\boldsymbol{n}_l}\right|\right|_2, \tag{27}$$

whereby $\widehat{\boldsymbol{n}}_{l,i}$ denotes the estimate of the position of the l-th sensor node resulting from the i-th of the I considered initializations of $\widehat{\Omega}_s$ and $\mu_{\boldsymbol{n}_l}=\frac{1}{I}\sum_{i=1}^{I}\widehat{\boldsymbol{n}}_{l,i}$ the corresponding mean.

We compare two initialization strategies, namely the proposed initialization using the observed source positions of one sensor node and a random initialization. For the proposed initialization scheme, the geometry was estimated using the observations of each of the sensor nodes as initial values resulting in I=L=4 different initializations. In the random case, all values of $\widehat{\Omega}_s$ are drawn from a normal distribution and I=100 initialization were considered.

It can be seen that the proposed initialization scheme leads to smaller deviations in the results. In most cases, the spread of the sensor node positions is even vanishingly small. Consequently, the choice of the sensor node whose source position estimates were used as initial values is not critical for the proposed initialization scheme. Moreover, the experiments showed that the spread of the estimated node orientations is in the order of magnitude of $(10^{-13})°$ and can therefore be neglected.

In addition to geometry, our approach also provides estimates of the positions of the sound sources. The MAE of these estimates

$$e_s = \frac{1}{G\cdot K}\sum_{g=1}^{G}\sum_{k=1}^{K}\left|\left|\boldsymbol{s}_{k,g} - \hat{\boldsymbol{s}}_{k,g}\right|\right|_2 \tag{28}$$

is given in Table 6. Again, the coordinate system of the geometry estimates is aligned with the coordinate system of the ground truth geometry using data set matching before the errors are calculated. These results are compared to the results of source localization, i.e., solving (14) for each acoustic source, using the ground truth geometry. It is shown that for small reverberation times T_{60}, the proposed iterative geometry calibration procedure yields comparable results to source localization using the ground truth geometry of the sensor network. As the reverberation time increases and thus the observation errors increase, the geometry calibration error increases and consequently the source localization error increases.

Moreover, the effect of fitness selection is shown in Table 6. Calculating the MAE e_s only for the subset of observed source positions selected by the fitness selection always leads to a smaller error. Thus, the algorithm succeeds in selecting a set of observations with smaller errors.

Finally, in Table 7, we compare the proposed approach to geometry calibration to state-of-the-art approaches solely using distance [46] or DoA estimates [29]. Hereby, the DoA-based approach utilizes the optional Maximum Likelihood refinement procedure which was proposed in [29]. Note that the considered distance-based approach called GARDE only delivers estimates for the positions of the sensor nodes and no orientations. Furthermore, the DoA-based approach estimates a relative geometry which has to be scaled subsequently. To this end, we employed the ground truth source node distances to fix the scaling as described in [31].

Table 7 shows that our approach is able to outperform both approaches by far. This can be explained by the additional information which results from the combined usage of distance and DoA information. In addition to that, the considered DoA-based approach contains no outlier handling while GARDE suffers from the outliers in the distance estimates.

Table 5 Influence of the weighting of the proposed geometry calibration procedure for K=20 and T_{60}=500 ms

WLS	WLS$_{\text{SRC}}$	Select	e_p/m	e_o/°
			0.26	2.9
✓			0.15	1.9
✓	✓		0.13	1.8
✓	✓	✓	0.08	1.9

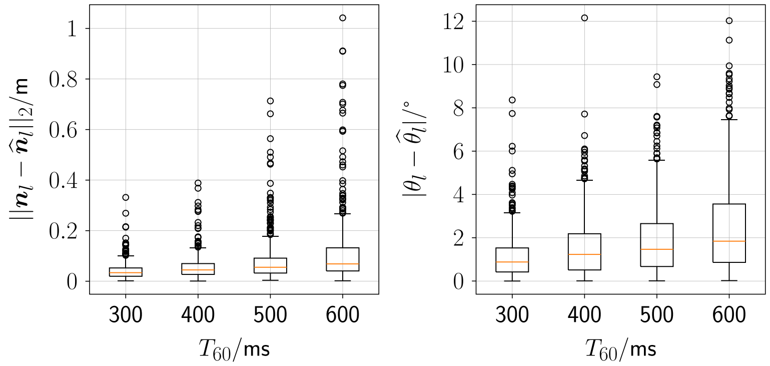

Fig. 7 Distribution of the geometry calibration error for K=20

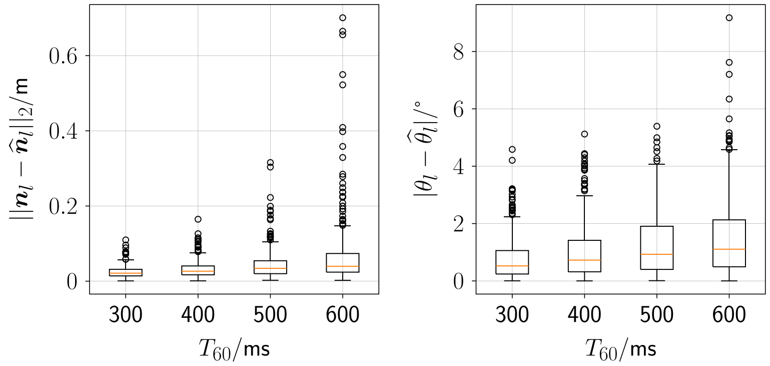

Fig. 8 Distribution of the geometry calibration error for K=100

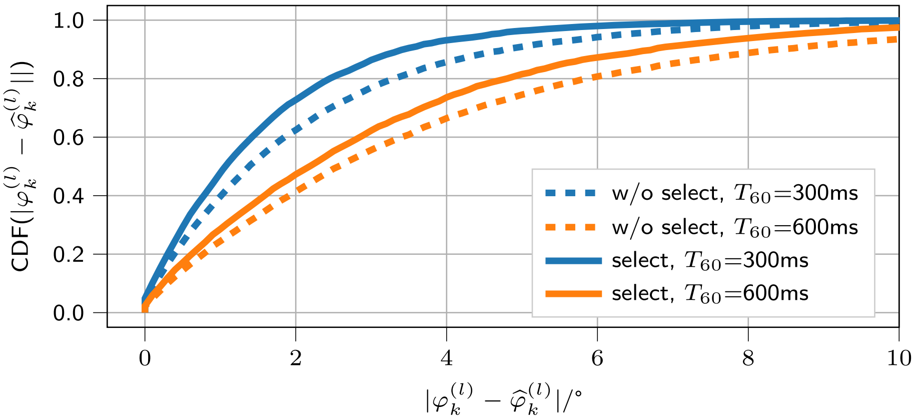

Fig. 9 Effect of fitness selection on the distribution of DoA estimation errors for K=20

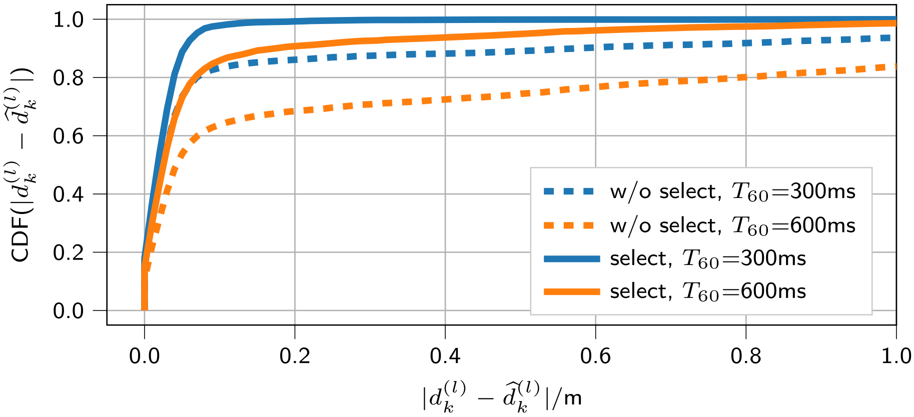

Fig. 10 Effect of fitness selection on the distribution of distance estimation errors for K=20

The proposed approach also compares favorably in terms of computational effort, when looking at the average computing time $\overline{T}_c$, i.e., the average time which is needed to estimate the geometry once. The average computing time for distance estimation (47 ms) and the average computing time for DoA estimation (545 ms) are not included in $\overline{T}_c$. Note that the DoA-based approach utilizes a Fortran accelerated implementation [47] to optimize the underlying cost function while all other approaches are based on a Python implementation. Moreover, Table 7 provides the average computing time required to solve the optimization problem in (5) by the Broyden-Fletcher–Goldfarb-Shanno (BFGS) method and the average computing time of the proposed approach if the weighting and the fitness selection is omitted which also can be interpreted as solving (5). Thereby, the latter leads to the same results as the BFGS method while being 70 times faster. This leaves room for the additional computing time required for the weighting and fitness selection in our approach. Consequently, despite its iterative character the proposed approach shows competitive computing time compared to the other considered approaches while providing better geometry estimates.

6 Conclusions

In this paper, we proposed an approach to geometry calibration in a WASN using DoA and distance information. The DoA and distances are estimated from the microphone signals and are interpreted as estimates of the relative positions of acoustic sources w.r.t. the coordinate system of the sensor node. Our approach uses these observations to alternatingly estimate the geometry and the acoustic sources' positions. Hereby, geometry calibration is formulated as an iterative data set matching problem which can be efficiently solved using a SVD.

In order to improve robustness against outliers and large errors contained in the observations, we integrate the iterative geometry estimation and source localization procedure into an error-model-based weighting and observation selection scheme. Simulations show that the proposed approach delivers reliable estimates of the geometry while being computationally efficient. Furthermore, it

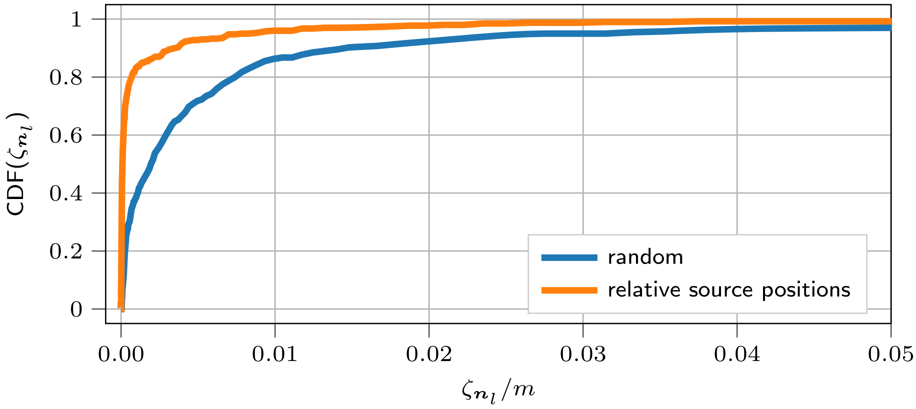

Fig. 11 Effect of the initialization on the convergence behavior of the sensor nodes' positions for K=20 and T_{60}=500 ms

Table 6 MAE $e_{\boldsymbol{s}}$/ m of source positions with and without fitness selection (Select) for K=20

Geometry	Select	300 ms	400 ms	500 ms	600 ms
Ground truth		0.04	0.06	0.07	0.08
Estimate		0.09	0.12	0.16	0.21
Estimate	✓	0.08	0.07	0.12	0.16

requires only a coarse synchronization between the sensor nodes.

Appendix

Convergence analysis of geometry calibration using iterative data set matching

We now analyze the convergence behavior of the iterative data set matching procedure, following the ideas of [48]. Therefore, we consider the part of iterative data set matching procedure where fitness selection is not used as shown in Algorithm 2. In the following, the superscript $[\eta]$ denotes the value after the update in the η-th iteration. Thus, the sets of quantities resulting from the η-th iteration of the alternating optimization procedure are defined as $\Omega_{\text{geo}}^{[\eta]}=\left\{\boldsymbol{n}_1^{[\eta]},\dots,\boldsymbol{n}_L^{[\eta]},\theta_1^{[\eta]},\dots,\theta_L^{[\eta]}\right\}$, $\Omega_{\boldsymbol{s}}^{[\eta]}=\left\{\boldsymbol{s}_1^{[\eta]},\dots,\boldsymbol{s}_K^{[\eta]}\right\}$, and $\Omega_{\text{w}}^{[\eta]}=\left\{w_{11}^{[\eta]},\dots,w_{KL}^{[\eta]}\right\}$. $\boldsymbol{R}_l^{[\eta]}$ denotes the rotation matrix corresponding to $\theta_l^{[\eta]}$. Furthermore, the cost function is now interpreted as a function of $\Omega_{\text{geo}}^{[\eta]}$, $\Omega_{\boldsymbol{s}}^{[\eta]}$ and $\Omega_{\text{w}}^{[\eta]}$:

$$J\left(\Omega_{\text{geo}}^{[\eta]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_{w}^{[\eta]}\right)=\sum_{l=1}^{L}\sum_{k=1}^{K} w_{kl}^{[\eta]}\left\|\boldsymbol{s}_k^{[\eta]}-\left(\boldsymbol{R}_l^{[\eta]}\boldsymbol{s}_k^{(l)}+\boldsymbol{n}_l^{[\eta]}\right)\right\|_2^2. \tag{29}$$

Considering the $(\eta+1)$-th iteration of the alternating optimization the following monotonicity property of the cost function holds:

```
Algorithm 2: Part of Algorithm 1 considered for convergence analysis
  Data: $\mathcal{S}$;
1 Init: $\Omega_{\boldsymbol{s}}^{[0]}$, $\Omega_{w}^{[0]}$, $\eta=0$;
2 repeat
3     $\Omega_{\text{geo}}^{[\eta+1]}$ = DSM_Calib($\mathcal{S}, \Omega_{\boldsymbol{s}}^{[\eta]}, \Omega_{w}^{[\eta]}$);
4     $\Omega_{\boldsymbol{s}}^{[\eta+1]}$ = SRC_Loc($\mathcal{S}, \Omega_{\text{geo}}^{[\eta+1]}, \Omega_{w}^{[\eta]}$);
5     $\Omega_{w}^{[\eta+1]}$ = Get_Weights($\mathcal{S}, \Omega_{\boldsymbol{s}}^{[\eta+1]}, \Omega_{\text{geo}}^{[\eta+1]}$);
6     $\eta=\eta+1$;
7 until Convergence;
  Result: $\Omega_{\text{geo}}^{[\eta]}$;
```

Table 7 Comparison of the calibration results and average computing time $\overline{T}_c$

	e_p/m	e_o/°	$\overline{T}_c$/ ms
DoA based [29] + Scaling [31]	0.19	1.7	338
GARDE [46]	0.17	-	1864
BFGS solving (5)	0.22	2.0	70
Proposed (w/o FitSelect/Weights)	0.22	2.0	1
Proposed	0.04	1.3	83

K=100; T_{60}=500 ms; Single core on an [Intel(R) Xeon(R) CPU E3-1240 v6 @ 3.70GHz, 32GB RAM]

Lemma 6.1 *The inequality*

$$J\left(\Omega_{geo}^{[\eta+1]},\Omega_{\boldsymbol{s}}^{[\eta+1]},\Omega_{w}^{[\eta+1]}\right)\le J\left(\Omega_{geo}^{[\eta]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_{w}^{[\eta]}\right) \tag{30}$$

holds for all $\eta > 0$, *i.e., each iteration monotonically decreases the considered cost function.*

Proof Inserting the definition of the weights

$$w_{kl}^{[\eta]}=\frac{1}{\left\|\boldsymbol{s}_k^{[\eta]}-\left(\boldsymbol{R}_l^{[\eta]}\boldsymbol{s}_k^{(l)}+\boldsymbol{n}_l^{[\eta]}\right)\right\|_2} \tag{31}$$

into (29) leads to

$$\begin{aligned}J\left(\Omega_{\text{geo}}^{[\eta]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_{w}^{[\eta]}\right)&=\sum_{l=1}^{L}\sum_{k=1}^{K} w_{kl}^{[\eta]}\left\|\boldsymbol{s}_k^{[\eta]}-\left(\boldsymbol{R}_l^{[\eta]}\boldsymbol{s}_k^{(l)}+\boldsymbol{n}_l^{[\eta]}\right)\right\|_2^2\\&=\sum_{l=1}^{L}\sum_{k=1}^{K}\frac{\left\|\boldsymbol{s}_k^{[\eta]}-\left(\boldsymbol{R}_l^{[\eta]}\boldsymbol{s}_k^{(l)}+\boldsymbol{n}_l^{[\eta]}\right)\right\|_2^2}{\left\|\boldsymbol{s}_k^{[\eta]}-\left(\boldsymbol{R}_l^{[\eta]}\boldsymbol{s}_k^{(l)}+\boldsymbol{n}_l^{[\eta]}\right)\right\|_2}\\&=\sum_{l=1}^{L}\sum_{k=1}^{K}\left\|\boldsymbol{s}_k^{[\eta]}-\left(\boldsymbol{R}_l^{[\eta]}\boldsymbol{s}_k^{(l)}+\boldsymbol{n}_l^{[\eta]}\right)\right\|_2\end{aligned} \tag{32}$$

for the costs at the end of the η-th iteration.

Firstly, data set matching is used to update the geometry Ω_{geo} (see line 3 in Algorithm 2). As described in [34] data set matching minimizes the cost function

$$J_\eta\left(\boldsymbol{n}_l,\boldsymbol{R}_l\right)=\sum_{k=1}^{K} w_{kl}^{[\eta]}\left\|\boldsymbol{s}_k^{[\eta]}-\left(\boldsymbol{R}_l\boldsymbol{s}_k^{(l)}+\boldsymbol{n}_l\right)\right\|_2^2 \tag{33}$$

for each of the L sensor nodes. Considering all L sensor nodes together results in

$$\Omega_{\text{geo}}^{[\eta+1]}=\underset{\Omega_{\text{geo}}}{\operatorname{argmin}}\sum_{l=1}^{L}J_\eta(\boldsymbol{n}_l,\boldsymbol{R}_l)=\underset{\Omega_{\text{geo}}}{\operatorname{argmin}}\,J\left(\Omega_{\text{geo}},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_{w}^{[\eta]}\right). \tag{34}$$

Consequently,

$$J\left(\Omega_{\text{geo}}^{[\eta+1]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_{w}^{[\eta]}\right)\le J\left(\Omega_{\text{geo}}^{[\eta]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_{w}^{[\eta]}\right) \tag{35}$$

holds.

The next step, i.e., the update of the source positions $\boldsymbol{s}_k$ (see line 4 in Algorithm 2), is done by minimizing

$$J_\eta\left(\boldsymbol{s}_k\right)=\sum_{l=1}^{L} w_{kl}^{[\eta]}\left\|\boldsymbol{s}_k-\left(\boldsymbol{R}_l^{[\eta+1]}\boldsymbol{s}_k^{(l)}+\boldsymbol{n}_l^{[\eta+1]}\right)\right\|_2^2 \qquad (36)$$

for all K source positions. Note that $J_\eta\left(\boldsymbol{s}_k\right)$ corresponds to a sum of squared Euclidean distances, i.e, a convex function of $\boldsymbol{s}_k$, and, thus, is convex. Consequently, the resulting linear least squares solution (see (15)) corresponds to the global minimum of $J_\eta\left(\boldsymbol{s}_k\right)$. Summarizing this step for all K acoustic sources gives

$$\Omega_{\boldsymbol{s}}^{[\eta+1]}=\underset{\Omega_{\boldsymbol{s}}}{\operatorname{argmin}}\sum_{k=1}^{K} J_\eta\left(\boldsymbol{s}_k\right)=\underset{\Omega_{\boldsymbol{s}}}{\operatorname{argmin}}\, J\left(\Omega_{\text{geo}}^{[\eta+1]},\Omega_{\boldsymbol{s}},\Omega_w^{[\eta]}\right). \qquad (37)$$

So it follows that

$$J\left(\Omega_{\text{geo}}^{[\eta+1]},\Omega_{\boldsymbol{s}}^{[\eta+1]},\Omega_w^{[\eta]}\right)\le J\left(\Omega_{\text{geo}}^{[\eta+1]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_w^{[\eta]}\right) \qquad (38)$$

and with (35) it holds:

$$J\left(\Omega_{\text{geo}}^{[\eta+1]},\Omega_{\boldsymbol{s}}^{[\eta+1]},\Omega_w^{[\eta]}\right)\le J\left(\Omega_{\text{geo}}^{[\eta]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_w^{[\eta]}\right). \qquad (39)$$

Finally, the influence of the weight update has to be discussed (see line 5 in Algorithm 2). Applying Titu's lemma to $J\left(\Omega_{\text{geo}}^{[\eta+1]},\Omega_{\boldsymbol{s}}^{[\eta+1]},\Omega_w^{[\eta]}\right)$ gives

$$\begin{aligned}
&J\left(\Omega_{\text{geo}}^{[\eta+1]},\Omega_{\boldsymbol{s}}^{[\eta+1]},\Omega_w^{[\eta]}\right)\\
&=\sum_{l=1}^{L}\sum_{k=1}^{K}\frac{\left\|\boldsymbol{s}_k^{[\eta+1]}-\left(\boldsymbol{R}_l^{[\eta+1]}\boldsymbol{s}_k^{(l)}+\boldsymbol{n}_l^{[\eta+1]}\right)\right\|_2^2}{\left\|\boldsymbol{s}_k^{[\eta]}-\left(\boldsymbol{R}_l^{[\eta]}\boldsymbol{s}_k^{(l)}+\boldsymbol{n}_l^{[\eta]}\right)\right\|_2}\\
&\ge\frac{\left(\sum_{l=1}^{L}\sum_{k=1}^{K}\left\|\boldsymbol{s}_k^{[\eta+1]}-\left(\boldsymbol{R}_l^{[\eta+1]}\boldsymbol{s}_k^{(l)}+\boldsymbol{n}_l^{[\eta+1]}\right)\right\|_2\right)^2}{\sum_{l=1}^{L}\sum_{k=1}^{K}\left\|\boldsymbol{s}_k^{[\eta]}-\left(\boldsymbol{R}_l^{[\eta]}\boldsymbol{s}_k^{(l)}+\boldsymbol{n}_l^{[\eta]}\right)\right\|_2}\\
&=\frac{\left(J\left(\Omega_{\text{geo}}^{[\eta+1]},\Omega_{\boldsymbol{s}}^{[\eta+1]},\Omega_w^{[\eta+1]}\right)\right)^2}{J\left(\Omega_{\text{geo}}^{[\eta]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_w^{[\eta]}\right)}.
\end{aligned} \qquad (40)$$

With (39) and (40) it follows:

$$\frac{\left(J\left(\Omega_{\text{geo}}^{[\eta+1]},\Omega_{\boldsymbol{s}}^{[\eta+1]},\Omega_w^{[\eta+1]}\right)\right)^2}{J\left(\Omega_{\text{geo}}^{[\eta]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_w^{[\eta]}\right)}\le J\left(\Omega_{\text{geo}}^{[\eta]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_w^{[\eta]}\right). \qquad (41)$$

Since $J\left(\Omega_{\text{geo}}^{[\eta]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_w^{[\eta]}\right)>0$ holds this results in

$$\left(J\left(\Omega_{\text{geo}}^{[\eta+1]},\Omega_{\boldsymbol{s}}^{[\eta+1]},\Omega_w^{[\eta+1]}\right)\right)^2\le\left(J\left(\Omega_{\text{geo}}^{[\eta]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_w^{[\eta]}\right)\right)^2 \qquad (42)$$

and, finally, in

$$J\left(\Omega_{\text{geo}}^{[\eta+1]},\Omega_{\boldsymbol{s}}^{[\eta+1]},\Omega_w^{[\eta+1]}\right)\le J\left(\Omega_{\text{geo}}^{[\eta]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_w^{[\eta]}\right). \qquad (43)$$

Due to the fact that $J\left(\Omega_{\text{geo}}^{[\eta]},\Omega_{\boldsymbol{s}}^{[\eta]},\Omega_w^{[\eta]}\right)$ is monotonically decreasing and has the lower bound $J_\infty\ge 0$ it converges to $J_\infty\ge 0$ for $\eta\to\infty$.

Abbreviations

BFGS: Broyden-Fletcher–Goldfarb-Shanno; CDF: Cumulative distribution function; CDR: Coherent-to-diffuse power ratio; CRNN: Convolutional recurrent neural network; DNN: Deep neural network; DoA: Direction of arrival; DRR: Direct-to-reverberant ratio; GP: Gaussian process; GRU: Gated recurrent unit; IRLS: Iteratively re-weighted least squares; LS: Least squares; MAE: Mean-absolute error; ppm: Parts per million; RANSAC: Random sample consensus; RIR: Room impulse response; STFT: Short-time Fourier transform; SVD: Singular value decomposition; TDoA: Time difference of arrival; ToA: Time of arrival; ToF: Time of flight; WASN: Wireless acoustic sensor network

Acknowledgements

We would like to thank Mr. Andreas Brendel for the fruitful discussions on distance estimation.

Authors' contributions

DNN model development and training: TG. Geometry calibration software and experiments: TG and JS. Writing paper: TG, JS, and RH. The authors read and approved the final manuscript.

Authors' information

Reinhold Haeb-Umbach received the Dipl.-Ing. and Dr.-Ing. degrees from RWTH Aachen University of Technology in 1983 and 1988, respectively. He is currently a professor of Communications Engineering at Paderborn University, Germany. His main research interests are in the fields of statistical signal processing and machine learning, with applications to speech enhancement, acoustic beamforming and source separation, as well as automatic speech recognition and unsupervised learning from speech and audio. He is a fellow of the International Speech Communication Association(ISCA) and of the IEEE.
Joerg Schmalenstroeer received the Dipl.-Ing. and Dr.-Ing. degree in electrical engineering from the University of Paderborn in 2004 and 2010, respectively. Since 2004, he has been a Research Staff Member with the Department of Communications Engineering of the University of Paderborn. His research interests are in acoustic sensor networks and statistical speech signal processing.
Tobias Gburrek is a Ph.D. student at Paderborn University since 2019 where he also pursued his Bachelor's and Masters's degree in Electrical Engineering. His research interests include acoustic sensor networks with a focus on geometry calibration and signal processing with deep neural networks.

Funding

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project 282835863. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The datasets and Python software code supporting the conclusions of this article are available in the paderwasn repository, https://github.com/fgnt/paderwasn. The MIRD database [39] is available under the following link: https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/multi-channel-impulse-response-database/.

Declarations

Consent for publication

All authors agree to the publication in this journal.

Competing interests

The authors declare that they have no competing interests.

Received: 14 December 2020 Accepted: 21 May 2021
Published online: 02 July 2021

References

1. A. Bertrand. Applications and trends in wireless acoustic sensor networks: a signal processing perspective, (2011). https://doi.org/10.1109/SCVT.2011.6101302
2. V. Potdar, A. Sharif, E. Chang, in *Proc. International Conference on Advanced Information Networking and Applications Workshops (AINA)*. Wireless Sensor Networks: A Survey (IEEE, Bradford, 2009), pp. 636–641. https://doi.org/10.1109/WAINA.2009.192
3. N. Ono, H. Kohno, N. Ito, S. Sagayama, in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Blind alignment of asynchronously recorded signals for distributed microphone array (IEEE, New Paltz, 2009). https://doi.org/10.1109/ASPAA.2009.5346505
4. S. Wozniak, K. Kowalczyk, Passive Joint Localization and Synchronization of Distributed Microphone Arrays. IEEE Signal Proc. Lett. **26**(2), 292–296 (2019). https://doi.org/10.1109/LSP.2018.2889438
5. B. Laufer-Goldshtein, R. Talmon, S. Gannot, Semi-supervised source localization on multiple manifolds with distributed microphones. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(7), 1477–1491 (2017). https://doi.org/10.1109/TASLP.2017.2696310
6. B. Laufer-Goldshtein, R. Talmon, S. Gannot, Semi-supervised source localization on multiple manifolds with distributed microphones. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(7), 1477–1491 (2017). https://doi.org/10.1109/TASLP.2017.2696310
7. A. Plinge, F. Jacob, R. Haeb-Umbach, G. A. Fink, Acoustic Microphone Geometry Calibration: an overview and experimental evaluation of state-of-the-art algorithms. IEEE Signal Proc. Mag. **33**(4), 14–29 (2016). https://doi.org/10.1109/MSP.2016.2555198
8. H. Afifi, J. Schmalenstroeer, J. Ullmann, R. Haeb-Umbach, H. Karl, in *Proc. ITG Fachtagung Sprachkommunikation (Speech Communication)*. MARVELO - A Framework for Signal Processing in Wireless Acoustic Sensor Networks, (Oldenburg, Germany, 2018)
9. G. Miller, A. Brendel, W. Kellermann, S. Gannot, Misalignment recognition in acoustic sensor networks using a semi-supervised source estimation method and Markov random fields (2020). http://arxiv.org/abs/arXiv:2011.03432
10. J. Elson, K. Roemer, in *Proc. ACM Workshop on Hot Topics in Networks (HotNets)*. Wireless sensor networks: a new regime for time synchronization (Association for Computing Machinery, Princeton, 2002)
11. R. Lienhart, I. V. Kozintsev, S. Wehr, M. Yeung, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. On the importance of exact synchronization for distributed audio signal processing (IEEE, Hong Kong, 2003), p. 840. https://doi.org/10.1109/ICASSP.2003.1202774
12. I.-K. Rhee, J. Lee, J. Kim, E. Serpedin, Y.-C. Wu, Clock synchronization in wireless sensor networks: an overview. Sensors. **9**(1), 56–85 (2009). https://doi.org/10.3390/s90100056
13. M. Hennecke, T. Plotz, G. A. Fink, J. Schmalenstroeer, R. Haeb-Umbach, in *Proc. IEEE/SP Workshop on Statistical Signal Processing (SSP 2009)*. A hierarchical approach to unsupervised shape calibration of microphone array networks, (2009), pp. 257–260. https://doi.org/10.1109/SSP.2009.5278589
14. L. Wang, T. Hon, J. D. Reiss, A. Cavallaro, Self-localization of ad-hoc arrays using time difference of arrivals. IEEE Trans. Signal Process. **64**(4), 1018–1033 (2016). https://doi.org/10.1109/TSP.2015.2498130
15. M. H. Hennecke, G. A. Fink, in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. Towards acoustic self-localization of ad hoc smartphone arrays, (Edinburgh, United Kingdom, 2011), pp. 127–132. https://doi.org/10.1109/HSCMA.2011.5942378
16. V. C. Raykar, I. V. Kozintsev, R. Lienhart, Position calibration of microphones and loudspeakers in distributed computing platforms. IEEE Trans. Speech Audio Proc. **13**(1), 70–83 (2005). https://doi.org/10.1109/TSA.2004.838540
17. D. Mills, Internet Time Synchronization: The Network Time Protocol. IEEE Trans. Commun. **39**, 1482–1493 (1991). https://doi.org/10.1109/TSA.2004.838540
18. M. Maróti, B. Kusy, G. Simon, A. Lédeczi, in *Proceedings of the 2nd international conference on Embedded networked sensor systems*. The flooding time synchronization protocol, (Baltimore, 2004), pp. 39–49. https://doi.org/10.1145/1031495.1031501
19. M. Maroti, B. Kusy, G. Simon, A. Ledeczi, in *Proc. International Conference on Embedded Networked Sensor Systems (SenSys)*. The flooding time synchronization protocol (Association for Computing Machinery, Baltimore, 2004). https://doi.org/10.1145/1031495.1031501
20. M. Leng, Y.-C. Wu, Distributed clock synchronization for wireless sensor networks using belief propagation. IEEE Trans. Signal Process. **59**(11), 5404–5414 (2011). https://doi.org/10.1109/TSP.2011.2162832
21. A. Plinge, G. A. Fink, S. Gannot, Passive online geometry calibration of acoustic sensor networks. IEEE Signal Proc. Lett. **24**(3), 324–328 (2017). https://doi.org/10.1109/LSP.2017.2662065
22. Y. Dorfan, O. Schwartz, S. Gannot, Joint speaker localization and array calibration using expectation-maximization. EURASIP Journal on Audio, Speech, and Music Processing. **2020**(9), 1–19 (2020). https://doi.org/10.1186/s13636-020-00177-1
23. J. Schmalenstroeer, F. Jacob, R. Haeb-Umbach, M. Hennecke, G. A. Fink, in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. Unsupervised geometry calibration of acoustic sensor networks using source correspondences (ISCA, Florence, 2011), pp. 597–600
24. F. Jacob, J. Schmalenstroeer, R. Haeb-Umbach, in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*. Microphone array position self-calibration from reverberant speech input (VDE, Aachen, 2012)
25. F. Jacob, J. Schmalenstroeer, R. Haeb-Umbach, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DOA-based microphone array postion self-calibration using circular statistics (IEEE, Vancouver, 2013), pp. 116–120. https://doi.org/10.1109/ICASSP.2013.6637620
26. F. Jacob, R. Haeb-Umbach, in *Proc. ITG Fachtagung Sprachkommunikation (Speech Communication)*. Coordinate mapping between an acoustic and visual sensor network in the shape domain for a joint self-calibrating speaker tracking (VDE, Erlangen, 2014)
27. F. Jacob, R. Haeb-Umbach, Absolute Geometry Calibration of Distributed Microphone Arrays in an Audio-Visual Sensor Network. ArXiv e-prints, abs/1504.03128 (2015)
28. R. Wang, Z. Chen, F. Yin, DOA-based three-dimensional node geometry calibration in acoustic sensor networks and its Cramér–Rao bound and sensitivity analysis. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(9), 1455–1468 (2019). https://doi.org/10.1109/TASLP.2019.2921892
29. S. Wozniak, K. Kowalczyk, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Exploiting rays in blind localization of distributed sensor arrays (IEEE, Barcelona, 2020), pp. 221–225. https://doi.org/10.1109/ICASSP40776.2020.9054752
30. M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications ACM. **24**(6), 381–395 (1981). https://doi.org/10.1145/358669.358692
31. T. Gburrek, J. Schmalenstroeer, A. Brendel, W. Kellermann, R. Haeb-Umbach, in *Proc. European Signal Processing Conference (EUSIPCO)*. Deep neural network based distance estimation for geometry calibration in acoustic sensor networks, (Amsterdam, The Netherlands, 2021)
32. A. Brendel, W. Kellermann, Distributed source localization in acoustic sensor networks using the coherent-to-diffuse power ratio. IEEE J. Sel. Top. Signal Proc. **13**(1), 61–75 (2019). https://doi.org/10.1109/JSTSP.2019.2900911
33. A. Brendel, A. Regensky, W. Kellermann, in *Proc. International Congress on Acoustics*. Probabilistic modeling for learning-based distance estimation (Deutsche Gesellschaft für Akustik (DEGA e.V.), Aachen, 2019)
34. J. M. Sachar, H. F. Silverman, W. R. Patterson, Microphone position and gain calibration for a large-aperture microphone array. IEEE Trans. Speech Audio Proc. **13**(1), 42–52 (2005). https://doi.org/10.1109/TSA.2004.834459
35. O. Sorkine-Hornung, M. Rabinovich, Least-squares rigid motion using svd. Computing. **1**(1), 1–5 (2017)
36. K. Aftab, R. Hartley, J. Trumpf, Generalized weiszfeld algorithms for lq optimization. IEEE Trans. Pattern Anal. Mach. Intell. **37**(4), 728–745 (2015). https://doi.org/10.1109/TPAMI.2014.2353625
37. I. Daubechies, R. DeVore, M. Fornasier, C. S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery. Commun. Pur. Appl. Math. **63**(1), 1–38 (2010). https://doi.org/10.1002/cpa.20303
38. E. A. Habets, Room impulse response generator. Technische Universiteit Eindhoven, Tech. Rep. **2**(2.4), 1 (2006)
39. E. Hadad, F. Heese, P. Vary, S. Gannot, in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*. Multichannel audio database in various acoustic environments (IEEE, Antibes, 2014), pp. 313–317. https://doi.org/10.1109/IWAENC.2014.6954309

40. N. J. Bryan, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation (IEEE, Barcelona, 2020), pp. 1–5. https://doi.org/10.1109/ICASSP40776.2020.9052970
41. A. Schwarz, W. Kellermann, Coherent-to-diffuse power ratio estimation for dereverberation. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(6), 1006–1018 (2015). https://doi.org/10.1109/TASLP.2015.2418571
42. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM. NIST (1993). https://doi.org/10.6028/nist.ir.4930
43. D. Kingma, J. Ba, in *Proc. International Conference on Learning Representations (ICLR)*. Adam: a method for stochastic optimization, (Banff, Canada, 2014). http://arxiv.org/abs/arXiv:1412.6980v9
44. L. Drude, F. Jacob, R. Haeb-Umbach, in *Proc. European Signal Processing Conference (EUSIPCO)*. DOA-estimation based on a complex Watson kernel method (IEEE, Nice, 2015). https://doi.org/10.1109/EUSIPCO.2015.7362384
45. J. R. Jensen, J. K. Nielsen, R. Heusdens, M. G. Christensen, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DOA estimation of audio sources in reverberant environments, (2016), pp. 176–180. https://doi.org/10.1109/ICASSP.2016.7471660
46. T. Gburrek, J. Schmalenstroeer, R. Haeb-Umbach, in *Accepted for Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Iterative geometry calibration from distance estimates for wireless acoustic sensor networks, (2021). http://arxiv.org/abs/arXiv:2012.06142
47. R. H. Byrd, P. Lu, J. Nocedal, C. Zhu, A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**(5), 1190–1208 (1995). https://doi.org/10.1137/0916069
48. P. V. Giampouras, A. A. Rontogiannis, K. D. Koutroumbas, Alternating iteratively reweighted least squares minimization for low-rank matrix factorization. IEEE Trans. Signal Process. **67**(2), 490–503 (2019). https://doi.org/10.1109/TSP.2018.2883921

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.