MICROPHONE ARRAY POSITION SELF-CALIBRATION FROM REVERBERANT SPEECH INPUT

Florian Jacob, Joerg Schmalenstroeer, Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn, Germany

{jacob, schmalen, haeb}@nt.uni-paderborn.de

ABSTRACT

In this paper we propose an approach to retrieve the geometry of an acoustic sensor network consisting of spatially distributed microphone arrays from unconstrained speech input. The calibration relies on Direction of Arrival (DoA) measurements which do not require a clock synchronization among the sensor nodes. The calibration problem is formulated as a cost function optimization task, which minimizes the squared differences between measured and predicted observations and additionally avoids the existence of minima that correspond to mirrored versions of the actual sensor orientations. Further, outlier measurements caused by reverberation are mitigated by a Random Sample Consensus (RANSAC) approach. The experimental results show a mean positioning error of at most $25\,\mathrm{cm}$ even in highly reverberant environments.

***Index Terms*—** Unsupervised, geometry calibration, microphone arrays, position self-calibration

1. INTRODUCTION

With the availability of inexpensive microphone hardware, ever more devices will be equipped with acoustic sensors. This allows for the possibility to form ad-hoc microphone networks by the sensors finding themselves in the same enclosure. In such scenarios the sensor placement will be unknown and may even change over time, but many acoustic signal processing applications, such as source localization and tracking, require the position of the microphones to be known.

Since the manual measurement of the microphone positions is a tedious and error-prone task techniques have been developed to automate this process. One possibility is to manually measure only the pairwise distances between the sensors and apply multi-dimensional scaling for recovering the positions [1]. Other authors have proposed a fully automatic spatial calibration, either based on TDoA [2] or by time of arrival (ToA) measurements [3]. They employed artificial calibration signals, such as chirps, to achieve high positioning accuracy.

While ToA and TDoA measurements require a clock synchronization among the sensor nodes, this is avoided by DoA-based algorithms [4, 5]. The DoA can be estimated using general cross correlation (GCC) approaches like GCC-PHAT [6] or by beamforming methods [7]. For DoA estimation, a sensor node must consist of a microphone array with known (intra-array) geometry. Note that DoA-based algorithms can only reveal the relative geometry and require an additional distance information e.g. from a TDoA measurement.

Irrespective of which measurements are used, the position self-calibration problem leads in general to a non-convex cost function, which may suffer from many local minima, such that gradient descent approaches often lead to unsatisfactory results. If additional restrictions apply, simplified formulations have been found. For example, in [8] it has been assumed, that the source signal impinges on all sensors at the same angle, resulting in the solution to reside in a low-dimensional affine subspace, which can be found by Singular Value Decomposition. Here, a DoA-based calibration is developed, which allows for arbitrary sensor configurations and yet avoids the existence of minima corresponding to mirrored version of the true sensor orientation, thus overcoming the deficiencies of earlier formulations [9, 10].

To minimize the calibration effort it is further preferable if the position self-calibration can be carried out with natural sound events, such as speech, as it does not require loudspeakers for the playback of the calibration signals. Using speech in a reverberant enclosure, asks for measures to avoid the impact of erroneous measurements due to speech being a non-ideal probing signal and due to reverberation. In [10] we have therefore proposed to embed the cost function optimization in an outlier rejection scheme based on the RANSAC from [11]. Here we propose an improved method for fusing the output of different RANSACs to an overall geometry estimate.

The paper is organized as follows. In Section 2 we present the new formulation of the cost function and compare it with an earlier approach in Sec. 3. After giving an overview of the complete calibration system and a description of the retrieval of the final geometry estimate in Sec. 4, Section 5 presents the experimental results, and we finish with some conclusions in Sec. 6.

2. DEVELOPMENT OF A COST FUNCTION

We consider geometry calibration in 2-dimensional space and assume that the sensor nodes deliver DoA estimates. Our

goal is to determine the position $[x_j^S, y_j^S]$ and orientation Θ_j, $j = 1, \ldots, K$ of the K sensor nodes. Without loss of generality the first sensor's location is positioned at the origin of the coordinate system ($[x_1^S, y_1^S] = [0, 0]$) with an orientation along the x-axis ($\Theta_1 = 0$). This coordinate system will be called "global coordinate system" in the following.

Further, the speaker positions $\boldsymbol{P}_i = [x_i^P, y_i^P], i = 1, \ldots, N$, are also unknown. All $3(K-1)+2N$ unknowns are gathered in the vector

$$\boldsymbol{\Omega} = [x_2^S, y_2^S, \Theta_2, \ldots, x_K^S, y_K^S, \Theta_K, x_1^P, y_1^P, \ldots, x_N^P, y_N^P].$$

These unknowns are to be determined from the $K \cdot N$ measurements $\phi_{ij}, 1 \le i \le N, 1 \le j \le K$, where ϕ_{ij} is the DoA measured by sensor node j for the i-th speaker position, relative to the sensor's own (local) coordinate system. At least $N \ge \frac{3(K-1)}{K-2}$ independent observations are required to solve the calibration problem.

Each DoA measurement ϕ_{ij} can be expressed as a unit-length direction vector

$$\boldsymbol{v}_{ij} = \begin{bmatrix} \cos(\phi_{ij}) & \sin(\phi_{ij}) \end{bmatrix}^T \tag{1}$$

within the coordinate system of the respective sensor, pointing to the observation. Here, $(\cdot)^T$ denotes vector transposition.

The measurement will be compared with a prediction. According to Fig. 1, the DoA predicted by the assumed sensor and source locations $[x_j^S, y_j^S]$ and $[x_i^P, y_i^P]$, respectively, is given by

$$\tilde{\phi}_{ij} = \operatorname{atan}\left(\frac{\Delta y_{ij}}{\Delta x_{ij}}\right) = \operatorname{atan}\left(\frac{y_i^P - y_j^S}{x_i^P - x_j^S}\right), \tag{2}$$

where $\operatorname{atan}(\cdot)$ is the four-quadrant extension of the arctan function.

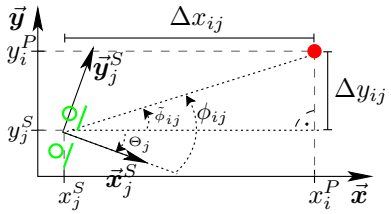

Fig. 1. Geometric relation between microphone array and observation.

Note that the predicted DoA $\tilde{\phi}_{ij}$ is defined in the global coordinate system, whereas the measurement ϕ_{ij} is made relative to the orientation Θ_j of the j-th sensor. A predicted direction vector $\hat{\boldsymbol{v}}_{ij}$ within the coordinate system of the particular sensor is thus obtained by

$$\hat{\boldsymbol{v}}_{ij} = \begin{bmatrix} \cos\left(\tilde{\phi}_{ij} - \Theta_j\right) \\ \sin\left(\tilde{\phi}_{ij} - \Theta_j\right) \end{bmatrix} = \begin{bmatrix} \cos\left(\operatorname{atan}\left(\frac{\Delta y_{ij}}{\Delta x_{ij}}\right) - \Theta_j\right) \\ \sin\left(\operatorname{atan}\left(\frac{\Delta y_{ij}}{\Delta x_{ij}}\right) - \Theta_j\right) \end{bmatrix} \tag{3}$$

which can be written as

$$\hat{\boldsymbol{v}}_{ij} = \underbrace{\begin{bmatrix} \cos(\Theta_j) & \sin(\Theta_j) \\ -\sin(\Theta_j) & \cos(\Theta_j) \end{bmatrix}}_{\boldsymbol{R}(-\Theta_j)} \underbrace{\frac{1}{\sqrt{\Delta x_{ij}^2 + \Delta y_{ij}^2}}}_{1/|\tilde{\boldsymbol{v}}_{ij}|} \underbrace{\begin{bmatrix} \Delta x_{ij} \\ \Delta y_{ij} \end{bmatrix}}_{\tilde{\boldsymbol{v}}_{ij}}. \tag{4}$$

Here, $\tilde{\boldsymbol{v}}_{ij}$ is the unnormalized direction vector from the position of sensor j to source position i. Eq. (4) describes the normalization and the rotation of the vector $\tilde{\boldsymbol{v}}_{ij}$ from the local coordinate system of the microphone array into the global coordinate system by the rotation matrix $\boldsymbol{R}(-\Theta_j)$.

Our goal is to determine the unknowns $\boldsymbol{\Omega}$ such that $\hat{\boldsymbol{v}}_{ij}$ is as close to $\boldsymbol{v}_{ij}$ as possible, for all i, j. Since these are unit length direction vectors, an Euclidian distance measure is, however, inappropriate. We therefore propose to use a cosine distance instead:

$$1 - \cos\left(\sphericalangle\left(\boldsymbol{v}_{ij}, \hat{\boldsymbol{v}}_{ij}\right)\right) = 1 - \boldsymbol{v}_{ij}^T \hat{\boldsymbol{v}}_{ij} = 1 - \boldsymbol{v}_{ij}^T \boldsymbol{R}(-\Theta_j) \frac{\tilde{\boldsymbol{v}}_{ij}}{|\tilde{\boldsymbol{v}}_{ij}|}.$$

The distance is minimal if $\boldsymbol{v}_{ij}$ and $\hat{\boldsymbol{v}}_{ij}$ point in the same direction and maximal if they point in opposite directions.

Assuming that the DoA of speakers farer away from the sensor can be estimated with a lower angle error than close-by speakers, it is reasonable to employ a weighted distance measure, where the weight corresponds to the distance between the speech source and the sensor. Thus the previous equation is multiplied by $|\tilde{\boldsymbol{v}}_{ij}|$ to obtain

$$g_{ij}(\boldsymbol{\Omega}) := |\tilde{\boldsymbol{v}}_{ij}|\left(1 - \cos\left(\sphericalangle\left(\boldsymbol{v}_{ij}, \hat{\boldsymbol{v}}_{ij}\right)\right)\right). \tag{5}$$

In summary, we want to seek position and orientation estimates such that the cost function

$$J_{\mathrm{C}}(\boldsymbol{\Omega}) = \sum_{i=1}^{N} \sum_{j=1}^{K} \left(g_{ij}(\boldsymbol{\Omega})\right)^2 \tag{6}$$

is minimized.

3. COMPARISON TO PREVIOUS APPROACH

To understand the benefits of the above objective function we compare it to the approach of [10], which was an improved version of a formulation originally published in [9].

Expressing the DoA measurement in the global coordinate system, we obtain (see Fig. 1)

$$\begin{aligned} \tan\left(\Theta_j + \phi_{ij}\right) &= \frac{\sin\left(\Theta_j + \phi_{ij}\right)}{\cos\left(\Theta_j + \phi_{ij}\right)} = \frac{y_i^P - y_j^S}{x_i^P - x_j^S} = \frac{\Delta y_{ij}}{\Delta x_{ij}} \\ \Leftrightarrow \underbrace{\sin\left(\Theta_j + \phi_{ij}\right) \Delta x_{ij} - \cos\left(\Theta_j + \phi_{ij}\right) \Delta y_{ij}}_{f_{ij}(\boldsymbol{\Omega})} &= 0, \end{aligned} \tag{7}$$

which can be expressed as follows, after some elementary mathematical manipulations,

$$\begin{aligned} f_{ij}(\boldsymbol{\Omega}) &= \underbrace{\begin{bmatrix} \cos(\phi_{ij}) \\ \sin(\phi_{ij}) \end{bmatrix}^T}_{\boldsymbol{v}_{ij}^T} \underbrace{\begin{bmatrix} \sin(\Theta_j) & -\cos(\Theta_j) \\ \cos(\Theta_j) & \sin(\Theta_j) \end{bmatrix}}_{\boldsymbol{R}(-\Theta_j + \frac{\pi}{2})} \underbrace{\begin{bmatrix} \Delta x_{ij} \\ \Delta y_{ij} \end{bmatrix}}_{\tilde{\boldsymbol{v}}_{ij}} \\ &= \boldsymbol{v}_{ij}^T \boldsymbol{R}(-\Theta_j + \frac{\pi}{2}) \tilde{\boldsymbol{v}}_{ij}. \end{aligned} \tag{8}$$

Using eq. (4) this can be written as

$$\begin{aligned} f_{ij}(\boldsymbol{\Omega}) &= |\tilde{\boldsymbol{v}}_{ij}| \boldsymbol{v}_{ij}^T \boldsymbol{R}(\frac{\pi}{2}) \hat{\boldsymbol{v}}_{ij} \\ &= -|\tilde{\boldsymbol{v}}_{ij}| \sin\left(\sphericalangle\left(\boldsymbol{v}_{ij}, \hat{\boldsymbol{v}}_{ij}\right)\right). \end{aligned} \tag{9}$$

The overall cost function to be minimized is then [10]

$$J_{\mathrm{S}}(\boldsymbol{\Omega}) = \sum_{i=1}^{N} \sum_{j=1}^{K} \left(f_{ij}(\boldsymbol{\Omega})\right)^2 . \tag{10}$$

Fig. 2 compares $f_{ij}^2(\boldsymbol{\Omega})$ of eq. (9) with $g_{ij}^2(\boldsymbol{\Omega})$ of eq. (5). While both achieve a minimum for $\boldsymbol{v}_{ij} = \hat{\boldsymbol{v}}_{ij}$ the marked difference is that within the range $[-\pi, \pi]$, eq. (5) has a single global minimum at $\boldsymbol{v}_{ij} = \hat{\boldsymbol{v}}_{ij}$, while eq. (9) has two more minima at the boundaries of the interval, which correspond to wrong orientation estimates of the sensors. For an iterative optimization, eq. (5) is thus much more suitable as it avoids being stuck in unwanted local minima.

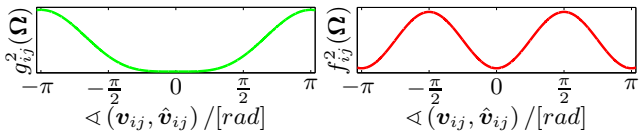

Fig. 2. Comparison of $g_{ij}^2(\boldsymbol{\Omega})$, eq. (5), with $f_{ij}^2(\boldsymbol{\Omega})$, eq. (9), as a function of the angle between $\boldsymbol{v}_{ij}$ and $\hat{\boldsymbol{v}}_{ij}$.

Different techniques can be employed to optimize eq. (6) and eq. (10). For the results presented here, we employed the iterative root finding method by Newton, which may be applied here, since the optimum is $J(\hat{\boldsymbol{\Omega}}) = 0$.

4. SYSTEM OVERVIEW

The precision of the automatic geometry calibration highly depends on the quality of the DoA estimates. This is particularly an issue in highly reverberant enclosures. To reduce the impact of erroneous DoA estimates, the position self-calibration is embedded into a RANSAC algorithm for outlier rejection.

In a typical setup with a person speaking and walking in the room for some time, many more DoA measurements are obtained than the minimum number required to solve eq. (6) via Newton's root finding method. Multiple RANSACs, each operating on a different sub-database ('sub-database selection'), are therefore run in parallel, as is depicted in Fig. 3, which shows a block diagram of the system.

In summary, an individual RANSAC consists of the following steps (for a detailed discussion see [10]): First, a subset of observations is selected at random ('Subset selection'). These observations are used to obtain a first estimate of the geometry parameters ('Model calculation'). Subsequently, all observations are scored against the current geometry estimate. All observations which comply, up to small deviations, to the geometry estimate form the new consensus set ('Consensus determination'). If the consensus set has increased compared to the previous iteration, the estimation of the geometry parameters is repeated using the observations from the new consensus set ('Iterate'). If the size of the consensus set did not increase compared to the previous iteration, the RANSAC starts all over again and a new set of initial observations is selected ('Fail'). The RANSAC is terminated, if the consensus set reaches a predefined amount of observations ('Success').

An obvious approach to combine the results of the individual RANSACs to a common geometry would be the calculation of the median or mean of each parameter in $\boldsymbol{\Omega}$. Superior results, however, are obtained if complete geometries rather than individual position estimates are reconciled. To this end we randomly select a reference geometry out of the RANSACs results and apply the rigid body transformations from [12] to the other geometries to match them with the reference geometry. Afterwards we compute the median of the matched geometries.

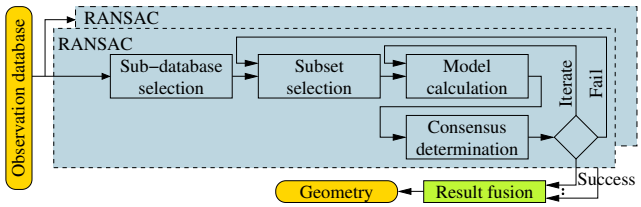

Fig. 3. Block diagram of the system.

5. EXPERIMENTAL RESULTS

In order to compare the performance of our purposed algorithm to the approach of our earlier publication [10] we used the same audio database as there in a first set of experiments. It contains recordings for 4 microphone arrays with 2 microphones per array at an inter-microphone distance of $0.05\,\mathrm{m}$, within a room of size $3.5\,\mathrm{m} \times 4\,\mathrm{m}$. The audio data for reverberation times T_{60} from $50\,\mathrm{ms}$ to $450\,\mathrm{ms}$ has been computed using the image method. For each reverberation time a separate trajectory of 90 seconds length was simulated, corresponding to a speaker walks through the room while continuously speaking.

The accuracy the calibration algorithm is quantified by the "mean position error" (MPE), which is the average distance between the real and the estimated positions of the sensors. To obtain absolute geometry estimates we match the calibration results with the actual geometry before computing the MPEs. Tab. 1 compares the MPEs obtained with the formulations of a cost function according to eq. (10) (J_{S}) of our previous publication (Baseline) with the purposed approach (J_{C}) from eq. (6).

T_{60} [ms] / Method	50	100	150	200	250	300	350	400	450
J_{S} (from [10])	0.05	0.14	0.16	0.49	0.74	0.50	0.69	0.96	0.36
J_{C} (proposed)	0.05	0.09	0.10	0.16	0.15	0.48	0.23	0.28	0.29

Table 1. Comparison of the mean position errors between the approach from [10] and the proposed method.

The results indicate a consistent improvement by the new formulation, but the MPE does not monotonically increase with T_{60}, as one could have expected. This is an artefact of the database and can be explained by the fact that different speaker trajectories were used for each reverberation time.

To better compare the MPEs at different reverberation

times the subsequent investigations are carried out on a new database, where the same speaker trajectory was used for all reverberation times. The size of the room and the microphone arrays are the same as before. The placement of the sensors within the new scenario is shown in Fig. 4. Additionally we increased the duration of the recordings to 6 minutes. We estimated the DoA values every $8\,\mathrm{ms}$ using an acoustic beamformer from [7], so we end up with 45000 observations per sensor. During the preselection phase of each of the 64 RANSACs we chose a sub-database of 450 observations to limit the computational complexity.

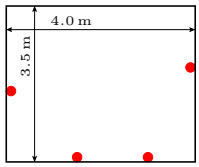

Fig. 4. Sensor placement (red dots).

Fig. 5 shows the MPE of the earlier cost function (J_S) and the proposed cost function (J_C) for different reverberation times. The color of the bars indicate the histogram of the individual RANSAC results, where lighter colors represent less frequent and darker colors more frequent MPEs. The black and cyan boxes at each bar represent the error of the geometry obtained from the different fusion techniques. Using the rigid body transformation technique (cyan) leads to a moderate reduction of the total error compared to the simple median operation (black).

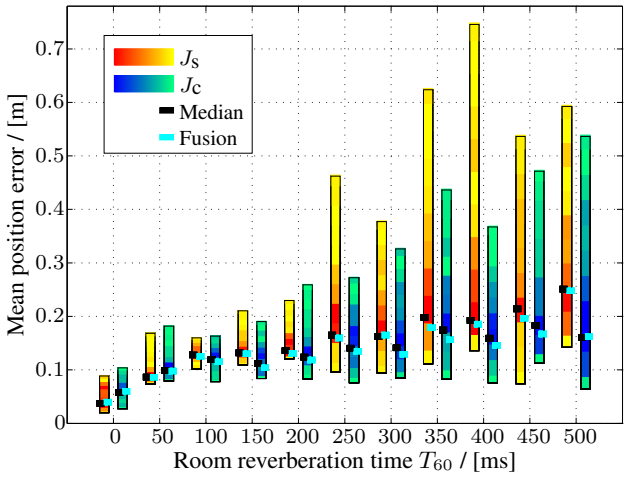

Fig. 5. Comparison of the mean positioning error (MPE) between the existing cost function (J_S) and the purposed cost function (J_C) for different reverberation times.

Comparing the results after the fusion step, we can state that the proposed cost function clearly outperforms the existing version, except for the very low reverberation times smaller than $100\,\mathrm{ms}$. Additionally, the plot shows that the MPEs of J_C vary less than those of J_S.

In case of very low reverberation, i.e., up to $50\,\mathrm{ms}$, it is possible to obtain geometry estimates without applying a RANSAC, but for higher reverberation times Newton's root finding method failed nearly for a hundred percent without a RANSAC.

6. CONCLUSIONS

We have presented a new formulation of the geometry calibration problem of distributed microphone arrays employing a cost function that avoids solutions that correspond to mirrored (wrong) sensor orientations. This cost function minimizes the squared difference between the observed and the predicted DoA. Additionally, we used the RANSAC to increase the robustness against reverberation. Multiple RANSACs were run in parallel on different subsets of the database and a rigid body transformation was employed to fuse the results. Overall a mean positioning error was obtained, which was at most $0.25\,\mathrm{m}$ for reverberation times up to $500\,\mathrm{ms}$.

7. ACKNOWLEDGMENTS

This work has been supported by Deutsche Forschungsgemeinschaft (DFG) under contract no. Ha3455/7-1.

8. REFERENCES

[1] S.T. Birchfield and A. Subramanya, "Microphone Array Position Calibration by Basis-Point Classical Multidimensional Scaling," *Speech and Audio Processing, IEEE Trans. on*, 2005.

[2] V. C. Raykar and R. Duraiswami, "Automatic Position Calibration of Multiple Microphones," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2004.

[3] Marco Crocco, Alessio Del Bue, Matteo Bustreo, and Vittorio Murino, "A closed form solution to the microphne position self-calibration problem," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2012.

[4] A. Redondi, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Geometric calibration of distributed microphone arrays," in *Multimedia Signal Processing, 2009. IEEE Int. Workshop on*.

[5] M. Hennecke, T. Ploetz, G.A. Fink, J. Schmalenstroer, and R. Haeb-Umbach, "A hierarchical approach to unsupervised shape calibration of microphone array networks," in *Statistical Signal Processing, 2009. 15th Workshop on*.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay'," *IEEE Trans. Acoust., Speech, Signal Process.*, 1976.

[7] E. Warsitz and R. Haeb-Umbach, "Acoustic filter-and-sum beamforming by adaptive principal component analysis," in *Int. Conf. Acoustics, Speech, and Signal Processing, 2005.*

[8] S. Thrun, "Affine structure from sound," in *In NIPS*, 2005.

[9] J. Kemper, H. Linde, and M. Walter, "Human-Assisted Calibration of an Angulation based Indoor Location System," in *2nd Int. Conf. on Sensor Technologies and Applications*, 2008.

[10] J. Schmalenstroeer, F. Jacob, R. Haeb-Umbach, M. Hennecke, and G. Fink, "Unsupervised Geometry Calibration of Acoustic Sensor Networks Using Source Correspondences," in *Proc. Interspeech*, 2011.

[11] M. Fischler and R. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Comm. of the ACM*, 1981.

[12] John H. and Challis, "A procedure for determining rigid body transformation parameters," *Journal of Biomechanics*, 1995.