DOA-BASED MICROPHONE ARRAY POSTION SELF-CALIBRATION USING CIRCULAR STATISTICS

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 absolute geometry of an acoustic sensor network, consisting of spatially distributed microphone arrays, from reverberant speech input. The calibration relies on direction of arrival measurements of the individual arrays. The proposed calibration algorithm is derived from a maximum-likelihood approach employing circular statistics. Since a sensor node consists of a microphone array with known intra-array geometry, we are able to obtain an absolute geometry estimate, including angles and distances. Simulation results demonstrate the effectiveness of the approach.

Index Terms— Geometry calibration, microphone arrays, position self-calibration

1. INTRODUCTION

The usage of distributed microphone arrays instead of a single one, boosts the performance of many speech enhancement and source localization tasks, in particular in large enclosures. With ever more mobile devices, such as laptops or smartphones, being equipped with multiple microphones, ad-hoc configurations can be formed with sensors being at unknown and possibly time-variant positions. However, knowledge of the geometric arrangement is desirable to exploit the full potential of distributed multi-channel signal processing algorithms, or to realize speaker localization and tracking applications.

Several solutions have been proposed for automatic microphone position self-calibration, which, however, come along with certain restrictions. A high positioning accuracy is reached by time of flight (TOF) based algorithms [1], however a tight clock synchronization between transmitter and receiver is essential, thus ruling out the use of a speaker's voice as a calibration signal. TOF measurements often require special calibration hardware [2].

Time difference of arrival (TDOA) based algorithms require a clock synchronization among all sensors, and artificial calibration signals, such as chirps or wide-band noise signals, have been employed to accomplish precise results [3]. The geometry self-calibration problem can be simplified by assuming the acoustic sources to be in the far field with respect to the microphones. This allows to exploit a rank constraint in the matrix containing the TDOA values [4]. Excellent position accuracies were achieved in an anechoic setting. However, to our experience the performance quickly degrades in the presence of reverberation.

If a sensor node consists of a microphone array, direction of arrival (DOA) based approaches can be used for geometry calibration [5]. They avoid the need for an exact clock synchronization among the distributed sensor nodes, but are considered to be able to recover only relative geometries, lacking any absolute distance information [6, 7]. If, however, the intra-array geometry is known, even absolute geometries can be estimated as is shown here.

The objective functions employed for geometry calibration have been mostly derived from geometric considerations. Here, we start from a statistical point of view. Modeling the observed DOAs as draws from a VON MISES probability density function (PDF) we arrive at the same objective function, which has previously been proposed on pure heuristic grounds [8]. A formulation is then developed, which allows for an absolute geometry calibration from unconstrained speech input in a reverberant environment, where a speaker is allowed to continuously move in the room without any halts at specific positions.

This paper is organized as follows. In sec. 2 we derive a calibration algorithm from the maximum-likelihood (ML) approach. Then we illustrate in sec. 3 how to exploit the intra-array configuration of a circular array to obtain absolute geometry information. After describing the simulation framework in sec. 4 we present both relative and absolute calibration results in a reverberant enclosure in sec. 5. The paper concludes with section 6.

2. ML FORMULATION OF GEOMETRY CALIBRATION PROBLEM

Our goal is to determine the geometry of an acoustic sensor network comprising I nodes delivering DOA measurements.

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

A sensor node consists of microphone array, where the location is described in 2D by a position vector $\boldsymbol{p}_i = [x_i, y_i]$ and an orientation θ_i, $i = 1, \ldots, I$.

Now consider a moving speaker located at the unknown positions $\boldsymbol{e}_t = [a_t, b_t]$, $t = 1, \ldots, T$. In the local coordinate system of the i-th node the angle, at which the source is located, is given by, see Fig. 1,

$$\mu_{i,t} := \alpha_{i,t} - \theta_i, \tag{1}$$

while a DOA estimation algorithm may yield a measurement $\varphi_{i,t}$, possibly different from $\mu_{i,t}$, due to reverberation, noise or other imperfections.

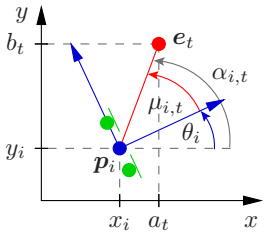

Fig. 1. Geometric relation between 2-element microphone array (green) and speaker location (red).

We assume that the DOA $\varphi_{i,t}$ follows a VON MISES probability density function (PDF):

$$p(\varphi_{i,t}; \mu_{i,t}, \kappa_{i,t}) = \frac{\exp\left(\kappa_{i,t} \cos(\varphi_{i,t} - \mu_{i,t})\right)}{2\pi I_0(\kappa_{i,t})}, \tag{2}$$

where I_0 is the zeroth-order modified Bessel function of the first kind. The VON MISES PDF is the most commonly used density in circular statistics, which defines PDFs on the unit sphere [9]. The parameters μ and κ are analogues to the mean and inverse of the variance (precision) of the normal PDF and are termed mean orientation and concentration parameter.

Now assume that we have independent DOA observations $\varphi_{i,t}$, $i = 1, \ldots I$, $t = 1, \ldots, T$ from I sensor nodes and T speaker positions. Additionally we assume that the sum of terms corresponding to the denominator of eq. (2) is approximately constant, so the log-likelihood function for the mean orientations is then given by

$$\ell(\boldsymbol{\mu}) = \sum_{t=1}^{T} \sum_{i=1}^{I} \kappa_{i,t} \cos(\varphi_{i,t} - \mu_{i,t}), \tag{3}$$

where $\boldsymbol{\mu} = \{\mu_{i,t}; i = 1, \ldots I, t = 1, \ldots, T\}$. Even if we assume the concentration parameters to be constants which are not estimated, this formulation, however, bears the obvious problem that there are as many unknowns $\mu_{i,t}$ as there are observations, which precludes a sensible estimation. The problem could be eased if multiple measurements were taken at a speaker position $\boldsymbol{e}_t$. This, however, would mean that the speaker has to stop moving at a position $\boldsymbol{e}_t$, speak and continue moving only after a sufficient number of DOAs has been taken.

Looking closer at the geometric configuration of Fig. 1, we see that the mean orientation $\mu_{i,t}$ can be written as a function of Cartesian coordinates:

$$\mu_{i,t} = \mu(\boldsymbol{p}_i, \theta_i, \boldsymbol{e}_t) = \operatorname{atan}\left(\frac{b_t - y_i}{a_t - x_i}\right) - \theta_i. \tag{4}$$

If we assume without loss of generality that the first sensor node is located at $\boldsymbol{p}_1 = [0, 0]$ at an orientation $\theta_1 = 0$, there are a total of $3(I-1)$ unknown sensor location parameters and $2T$ source position parameters to be estimated. The number of observations $I \cdot T$ should therefore be at least $3(I-1)+2T$, resulting in $T \geq 3(I-1)/(I-2)$, to have more measurements than unknowns. Note that a single DOA per speaker position is sufficient, as long as this inequality holds.

In this work we considered the concentration parameters to be constants, which are not estimated. A viable heuristic is to assume that they are proportional to the distance between the source and the sensor, i.e., $\kappa_{i,t} \propto d_{i,t} = |\boldsymbol{e}_t - \boldsymbol{p}_i|$. The maximum-likelihood estimates of the sensor positions, sensor orientations and speaker positions are then given by

$$\langle \boldsymbol{p}^*_{2:I}, \theta^*_{2:I}, \boldsymbol{e}^*_{1:T} \rangle = \underset{\boldsymbol{p}_{2:I}, \theta_{2:I}, \boldsymbol{e}_{1:T}}{\operatorname{argmax}} \left\{ \sum_{t=1}^{T} \sum_{i=1}^{I} d_{i,t} \cos\left(\varphi_{i,t} - \mu(\boldsymbol{p}_i, \theta_i, \boldsymbol{e}_t)\right) \right\}, \tag{5}$$

where $(\cdot)_{2:I}$ and $(\cdot)_{1:T}$ denotes all sensor and all speaker position variables respectively.

Indeed, this is the objective function earlier introduced in [8]. While a purely heuristic motivation was given there, we have shown here, that it can be derived from the maximum-likelihood principle assuming a VON MISES PDF for the observation probability.

Eq. (5) constitutes a nonlinear optimization problem. In [8] we proposed to solve it by an iterative Newton algorithm. To do so, the optimization problem was rewritten to

$$\langle \boldsymbol{p}^*_{2:I}, \theta^*_{2:I}, \boldsymbol{e}^*_{1:T} \rangle = \underset{\boldsymbol{p}_{2:I}, \theta_{2:I}, \boldsymbol{e}_{1:T}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{T} \sum_{i=1}^{I} d_{i,t} \left[1 - \cos\left(\varphi_{i,t} - \mu(\boldsymbol{p}_i, \theta_i, \boldsymbol{e}_t)\right)\right] \right\}, \tag{6}$$

and the minimum was found by an iterative Newton root finding algorithm.

Note that the formulation of the objective function contains position coordinates, although the measurements comprise only angles. Cartesian coordinates for sensor and source position have been chosen to arrive at a formulation, where there are more measurements than unknowns. It should, however, not be misinterpreted in the sense that absolute positions can be retrieved. In fact, the position estimates exhibit an arbitrary scaling. To fix the scaling, at least one distance value has to be employed. In the next section we show that sensor nodes consisting of circular microphone arrays and knowledge of the radii of the circular arrays is sufficient to transform the relative geometry to an absolute geometry, including the proper scaling.

3. ABSOLUTE GEOMETRY CALIBRATION WITH CIRCULAR ARRAYS

In the following we assume that a sensor node consists of a circular microphone array comprising K microphones, with known intra-array arrangement. The spatial location of the j-th sensor node is given by the position of the array's center $\boldsymbol{g}_j = [x_j; y_j]$ and it's orientation γ_j (see Fig. 2), $j = 1, \ldots, J$.

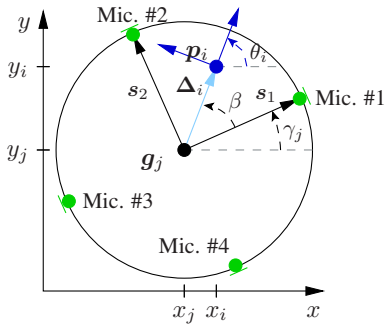

Fig. 2. Structure of a circular microphone array with center $\boldsymbol{g}_j$, orientation γ_j and $K = 4$ microphones equidistantly spaced on a circle of radius r around the center.

The location $\boldsymbol{s}_c$ of the c-th microphone, $c = 1, \ldots, K$, relative to the center of the circular array, is given by

$$\boldsymbol{s}_c = r \left[\cos\left(\gamma_j + 2\pi c/K\right), \sin\left(\gamma_j + 2\pi c/K\right)\right]^T, \quad (7)$$

where r denotes the radius of the array. Each of the $\binom{K}{2}$ microphone pairs within a circular array forms a 2-element array, which can deliver a DOA estimate. Let i enumerate all microphone pairs that can be formed within a circular array, i.e. $i := i(c, d)$, where $c = 1, \ldots, K$ and $d = c + 1, \ldots, K$ are the microphone indices. The geometry of the i-th 2-element array is characterized by the orientation θ_i and its center $\boldsymbol{p}_i$, see Fig. 2. The center is given by

$$\boldsymbol{p}_i = \boldsymbol{g}_j + \boldsymbol{\Delta}_i \;\text{ with }\; \boldsymbol{\Delta}_i := \boldsymbol{\Delta}_{i(c,d)} = \frac{\boldsymbol{s}_c + \boldsymbol{s}_d}{2}. \quad (8)$$

Using angular relationships derived from Fig. 2 the orientation θ_i of a microphone pair formed by microphone #1 and #d can be expressed as

$$\theta_i = \gamma_j + \beta \;\text{ with }\; \beta = \arccos\left(\frac{\boldsymbol{\Delta}_{i(1,d)}^T \boldsymbol{s}_1}{\|\boldsymbol{\Delta}_{i(1,d)}\| \|\boldsymbol{s}_1\|}\right). \quad (9)$$

Similar formulations can be found for the other pairs.

The intra-array relations of eq. (8) and eq. (9) form additional constraints for the optimization problem. These constraints particularly provide the necessary distance information to obtain an absolute calibration, since they incorporate the known radius r to express $\boldsymbol{p}_i$ and θ_i as functions of the parameters of the circular array. This distance information allows to perform an absolute geometry calibration, without measuring any additional distance, such as the distance between two sensor nodes, as it is proposed in [10].

The absolute geometry of the circular sensor network is revealed by plugging eq. (8) and eq. (9) into eq. (6) and optimizing it with respect to the positions and orientations of the circular arrays:

$$\langle \boldsymbol{g}_{2:J}^*, \gamma_{2:J}^*, \boldsymbol{e}_{1:T}^* \rangle = \underset{\boldsymbol{g}_{2:J}, \gamma_{2:J}, \boldsymbol{e}_{1:T}}{\operatorname{argmin}} \quad (10)$$
$$\left\{ \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{i=1}^{I} d_{i,t} \left[1 - \cos\left(\varphi_{i,t} - \mu(\boldsymbol{g}_j, \gamma_j, \boldsymbol{\Delta}_i, \boldsymbol{e}_t)\right)\right] \right\}.$$

Due to reverberations in real environments we cannot assume the DOA estimates to be accurate. Using the DOA estimates obtained from an adaptive beamformer operating on the reverberant microphone signals, see section 5, we observed a bias in the recovered geometry: the estimated distances were consistently smaller than the true ones.

This effect can be explained by looking at eq. (10). In case of accurate DOA estimates and a perfectly recovered geometry the cosine expression equals one. Due to measurement errors the cosine term may differ from one. The term is weighted by the distance between the i-th sensor and the t-th speaker position, $d_{i,t}$. Thus the larger the distance the larger the contribution to the cost function. Consequently, smaller distances are preferred as they lead to smaller contributions to the overall cost.

An obvious remedy to this problem would be to set $d_{i,t} = 1$ and thus obtain a "normalized" cost function. This, however, negatively impacted the numerical stability and convergence properties. If the initial values of the geometry parameters are far off the true ones, divergence of the optimization was frequently encountered. We therefore adopted a two-stage strategy, were we first performed a calibration with the cost function of eq. (10) which includes the distance term. The geometry obtained from this first stage was taken as the initial values for a second optimization with the "normalized" cost function, where $d_{i,t}$ is set to unity. This approach avoided the convergence problems of the normalized cost function and yet delivered geometry estimates that did not favour small distances.

4. SIMULATION FRAMEWORK

In order to evaluate the performance of our proposed calibration algorithm in reverberant environments we compiled an audio-database with reverberation times from $0\,\mathrm{ms}$ to $500\,\mathrm{ms}$, by employing the image method [11]. The setup consists of 15 microphones, within a room of size $8\,\mathrm{m} \times 5\,\mathrm{m}$, arranged in 3 circular arrays with $5\,\mathrm{cm}$ radius each. We used 5 arrangements of the circular arrays with a distance of approximately $1\,\mathrm{m}$ between the individual arrays. An example arrangement is depicted in Fig. 3. The same $6\,\mathrm{min}$ long trajectory, which corresponds to $T \approx 45000$, is used for each reverberation time and microphone configuration.

The DOA estimates, which are input to the geometry cali-

bration algorithm, are obtained from the audio signal by an adaptive beamformer [12, 13]. The beamformer coefficients are given by the eigenvector with the largest eigenvalue of the power spectral density matrix of the microphone signals and are determined by an adaptive eigenvalue decomposition. From the coefficients the DOA can be computed as a side product. The algorithm delivers DOA estimates on blocks of 128 samples (at 16 kHz sampling rate), i.e., every 8 ms, which is much faster than frame sizes typically required for GCC-PHAT [14]. Thus, the speaker can continuously move and is not required to halt for talking.

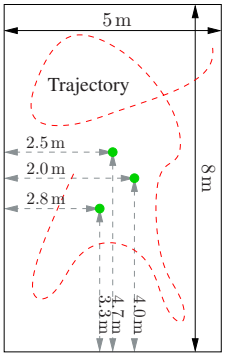

Fig. 3. Example arrangement of 3 circular microphone arrays (green dots).

As mentioned before, imperfections of the input DOA estimates highly influence the precision of the geometry calibration. To reduce their impact on the calibration result, the calibration algorithm is embedded in a "random sample consensus" (RANSAC) framework [15] for outlier rejection. Since the way we employ the RANSAC has been reported earlier [7], we give only a short summary here.

Since we continuously obtain new DOA estimates, due to the movement of the speaker, we usually collect many more observations than are necessary for a single calibration procedure, actually even more than we can handle in a single calibration procedure. To cope with large amounts of observations we split the DOA estimates into subsets, perform a RANSAC embedded calculation on each subset and finally fuse the results into a single geometry estimate.

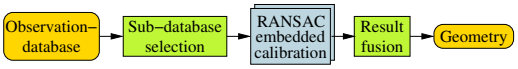

Fig. 4. Block diagram of simulation framework.

5. SIMULATION RESULTS

The accuracy of 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 sensor nodes. To compute the MPE we perform a rigid body transformation [16] to match the calibration result and the reference geometry, since the calibration result has an arbitrary global orientation.

To distinguish between geometric errors and scaling errors, we evaluate the MPE for a relative and an absolute calibration. The relative MPE (rel) is obtained after applying a rigid body transformation, which is allowed to scale the calibration result such that the difference between estimated and scaled position and the true sensor position is minimized. The absolute MPE (abs) is computed without an additional scaling, i.e., employing the scaling estimated by the proposed objective function, hence it identifies scaling errors as well. Fig. 5 shows the relative and absolute calibration error for different reverberation times averaged over all sensor configurations.

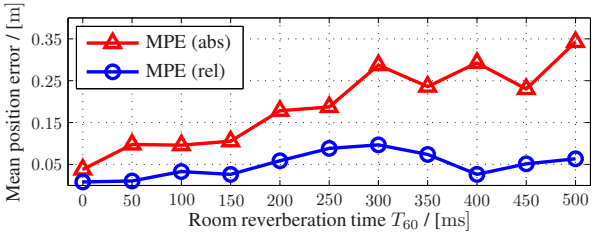

Fig. 5. Comparison of the relative (rel) and absolute (abs) MPE averaged over 5 different sensor arrangements.

For low reverberation times up to $T_{60} = 150\,\mathrm{ms}$ the relative positioning error is smaller than 5 cm and the absolute positioning error is about as twice as large. With increasing reverberation time the gap between both curves increases, since it is more difficult to obtain the right scaling. The MPE does not increase monotonically with increasing reverberation time, probably because certain imperfections of the geometry estimates can be easier compensated for by the rigid body transformation than others, especially if an additional scaling is used. According to our implementation of the TDOA based approach [4], we can state that it outperforms our proposed framework in a non reverberant scenario.However, the approach completely failed in the presence of even moderate reverberation ($T_{60} > 100\,\mathrm{ms}$), while our framework delivers reasonable results, even for large reverberation times.

6. CONCLUSIONS AND RELATION TO PRIOR WORK

We have derived a calibration algorithm from the maximum-likelihood principle using circular statistics to obtain the geometric configuration of distributed sensor nodes. Incorporating the intra-array geometry of circular sensor nodes, we are able to obtain the absolute geometry of the sensor network without measuring any distance between sensor nodes. Additionally we have embedded the proposed calibration algorithm into a RANSAC framework and evaluated its performance in a reverberant environment. Overall we achieved a relative MPE smaller than 12 cm and an absolute MPE smaller than 35 cm for reverberation time up to 500 ms in a simulated room of size $40\,\mathrm{m}^2$.

While we have motivated the objective function earlier from geometric considerations [8], it is shown here that it can be derived as the ML solution using the VON MISES PDF. We also show here that it is sufficient to know an intra-array distance of an array where the microphones must not be arranged collinearly to estimate absolute geometries from DOA estimates, while in earlier work an inter-array distance had to be provided to fix the scaling.

7. REFERENCES

[1] M. Crocco, A. Del Bue, M. Bustreo, and V. Murino, "A Closed Form Solution to the Microphone Position Self-Calibration Problem," in *International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2012, pp. 2597–2600.

[2] J.M. Sachar, H.F. Silverman, and W.R. Patterson, "Microphone position and gain calibration for a large-aperture microphone array," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 42–52, Jan. 2005.

[3] S.D. Valente, M. Tagliasacchi, F. Antonacci, P. Bestagini, A. Sarti, and S. Tubaro, "Geometric calibration of distributed microphone arrays from acoustic source correspondences," in *IEEE International Workshop on Multimedia Signal Processing*, Oct. 2010, pp. 13–18.

[4] S. Thrun, "Affine Structure From Sound," in *Conference on Neural Information Processing Systems*, Cambridge, MA, Jan. 2005, pp. 1353–1360, MIT Press.

[5] A. Redondi, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Geometric calibration of distributed microphone arrays," in *IEEE International Workshop on Multimedia Signal Processing*, Oct. 2009, pp. 1–5.

[6] P.D. Jager, M. Trinkle, and A. Hashemi-Sakhtsari, "Automatic microphone array position calibration using an acoustic sounding source," in *IEEE Conference on Industrial Electronics and Applications*, May 2009, pp. 2110–2113.

[7] J. Schmalenstroeer, F. Jacob, R. Haeb-Umbach, M. Hennecke, and G. Fink, "Unsupervised Geometry Calibration of Acoustic Sensor Networks Using Source Correspondences," in *Interspeech*, Aug. 2011, pp. 597–600.

[8] Florian Jacob, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach, "Microphone Array Position Self-Calibration from Reverberant Speech Input," in *International Workshop on Acoustic Signal Enhancement*, Sept. 2012, pp. 1–4.

[9] Kantilal V. Mardia, Peter E. Jupp, et al., *Directional statistics*, John Wiley & Sons, 2000.

[10] J. Kemper, M. Walter, and H. Linde, "Human-Assisted Calibration of an Angulation Based Indoor Location System," in *Second International Conference on Sensor Technologies and Applications*, Aug. 2008, pp. 196–201.

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

[12] E. Warsitz and R. Haeb-Umbach, "Acoustic filter-and-sum beamforming by adaptive principal component analysis," in *International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2005, pp. iv/797–iv/800 Vol. 4.

[13] E. Warsitz and R. Haeb-Umbach, "Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, July 2007.

[14] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 320–327, Aug. 1976.

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

[16] John H. and Challis, "A procedure for determining rigid body transformation parameters," *Journal of Biomechanics*, pp. 733–737, 1995.