Self-Calibration and Geometry Inference with Distributed Compact Spherical Microphone Arrays

Thomas Wilding[1], Christian Schörkhuber[2]

[1] *Email: thomas.wilding@outlook.com*

[2] *Institut für Elektronische Musik und Akustik, Email: schoerkhuber@iem.at*

Introduction

Distributed spherical microphone arrays are frequently used for large scale acoustic scene analysis, spatial sound recordings, and room acoustic analysis. For all these applications the positions and orientations of the arrays must be known; however, when the arrays are distributed over a large area, measuring their positions is often infeasible or cumbersome at best. Blind estimation of microphone positions from the recorded signals, referred to as position *self-calibration*, has been studied in [1, 2, 3], and usually these methods rely on the estimated time difference of arrival (TDOA) of sound events picked up by the microphones.

For some applications, e.g. room acoustic analysis, also a floor plan of the recording venue needs to be available; more generally, the positions and orientations of reflective surfaces need to be known. Again, manually measuring these room properties can be cumbersome and time consuming. The problem of estimating these properties from recorded signals or room impulse responses is referred to as *geometry inference*, and several solutions assuming known microphone positions have been proposed [4, 5]. Methods for jointly solving the self-calibration and geometry inference problem have been proposed in [6, 7].

In this paper we propose a practical solution to both problems using distributed spherical microphone arrays equipped with 4 cardioid microphones. The unknown positions and orientations of the arrays as well as the positions and orientations of reflective room boundaries are estimated by recording several impulse-like sounds (hand-claps) at arbitrary unknown positions. The proposed approach is based on two sets of parameters: (i) the estimated direction-of-arrivals (DOAs) of the direct sounds and first-order reflections, and (ii) the TDOAs between sound events picked up by different arrays (inter-array TDOAs) as well as the TDOAs between the direct sound and first-order reflections at each array (intra-array TDOAs). From these parameters, we firstly estimate the positions and orientations of the arrays as well as the source positions using the direct sound events only; the minimum number of microphone arrays and sources required is 2 and 3 for the 2-dimensional and 3-dimensional case, respectively. Secondly, we estimate the positions and orientations of room boundaries using the estimated DOAs and TDOAs of first order reflections. The proposed solution for the geometry inference problem can be applied to arbitrary room geometries, however, here we consider only rectangular ones.

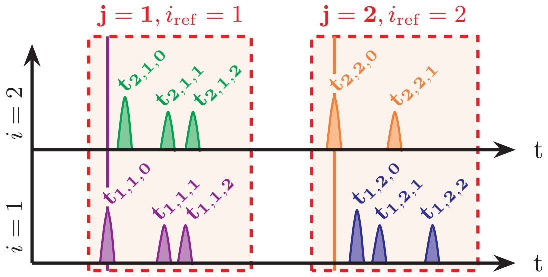

Figure 1: Visualization of signals from two different sources (with reflections) arriving at two synchronized arrays, arrival times are $t_{i,j,r}$.

The performance of the proposed algorithm is evaluated using measurement data from two different environments.

Signal Model and Notation

We model the signal of the j-th source (i.e. the j-th hand clap) arriving at the i-th array as

$$\mathbf{y}_{i,j}(n) = \sum_{r=0}^{N_r} \mathbf{a}(\Omega_{i,j,r}) x_{i,j,r}(n) + \mathbf{v}_i(n), \qquad (1)$$

where $x_{i,j,r}(n)$ is a sound event ($r = 0$ refers to the direct sound and $r > 0$ to the r-th reflection) and $\mathbf{v}_i(n)$ the measurement noise. $\mathbf{a}(\Omega_{i,j,r})$ is a weighting depending on the DOA Ω. With $t_{i,j,0}$ denoting the time of arrival of the direct sound of source j at array i, the inter-array TDOA is defined as

$$\Delta t_{i,j,0} = t_{i,j,r} - t_{i_{\text{ref}},j,r}, \qquad (2)$$

where i_{ref} is the index of the array that detected the first direct sound arrival (see Figure 1).

Similarly, we define the intra-array TDOA $\Delta t_{i,j,r}$ with $r > 0$ as

$$\Delta t_{i,j,r} = t_{i,j,r} - t_{i,j,0}, \qquad (3)$$

where $t_{i,j,r}$ is the time of arrival of the r-th reflection of the j-th source at array i.

By $\Omega_{i,j,r} = (\varphi_{i,j,r}, \vartheta_{i,j,r})$ we denote the DOA of the r-th sound event created by source j as observed by the i-th array. Note that since the orientations of the arrays are unknown, $\Omega_{i,j,r}$ refers to the *local* coordinate system of the i-th array.

The signal model is visualized in Figure 1, indicating the times of arrival of sound events and possible signal windows during which only the j-th source, as well as the reference array i_{ref} for each source.

Direction of Arrival Estimation

Many DOA estimators need a search over a large parameter space, for example steered response power (SRP) or similar approaches. An efficient estimator is proposed in [8], based on a direct weighting of the capsule look directions of a microphone array (fulfilling certain restrictions) by the recorded spectrum.

An extension thereof is described in this section in form of an alternative weighting using an eigendecomposition of the array covariance matrix performed in the frequency domain, computed as

$$\mathbf{R}(k) = \mathbb{E}\left[\mathbf{Y}(k,n)\mathbf{Y}^H(k,n)\right] = \mathbf{U}(k)\mathbf{D}(k)\mathbf{U}(k)^H. \quad (4)$$

In upper equation $\mathbb{E}\left[\cdot\right]$ denotes the expectation operator, $\mathbf{Y}(k,n)$ the STFT of the array output $\mathbf{y}(n)$ and $\mathbf{U}(k)$ and $\mathbf{D}(k)$ the eigenvector and eigenvalue matrices. As estimate for the true covariance matrix $\mathbf{R}(k)$ the sample covariance matrix $\hat{\mathbf{R}}(k,m)$ is used, computed over a short signal window centered around a sample m at frequency bin k. As alternative weighting of the capsule-look-directions (in the columns of $\mathbf{N}$) the eigenvector $\tilde{\mathbf{u}}(k,m)$ corresponding to the largest eigenvalue $\tilde{\lambda}(k,m)$ is used.

This DOA estimation is performed separately for all microphone arrays, yielding a DOA vector

$$\mathbf{d}_{i,j}(k,m) = \mathbf{N}\cdot|\tilde{\mathbf{u}}_{i,j}(k,m)| \quad (5)$$

at each time-frequency bin.

An instantaneous DOA estimate $\hat{\Omega}_{i,j}(m)$ at time m is then found by computing a histogram over the azimuth and elevation angles of all frequencies k of $\mathbf{d}_{i,j}(k,m)$ as

$$\mathrm{H}_{i,j}(\alpha,\beta,m) = \underset{\varphi,\vartheta,k}{\text{histogram}}\ \mathbf{d}_{i,j}(k,m) \quad (6)$$

and picking the angular direction of the maximum in the histogram as final DOA estimate

$$\hat{\Omega}_{i,j}(m) = \underset{\alpha,\beta}{\text{argmax}}\ \mathrm{H}_{i,j}(\alpha,\beta,m). \quad (7)$$

α and β are the angles of the histogram bins.

Time of Arrival Estimation

For TOA estimation the broadband character of the calibration signals can be exploited by performing peak picking over time of the largest eigenvalue $\tilde{\lambda}_{i,j}(k,m)$ at each frequency. This results in N_p possible TOAs according to

$$t_{i,j,p}(k) = \mathbb{PP}_{r=1}^{N_p}\left[\tilde{\lambda}_{i,j}(k,m)\right] \quad (8)$$

where $\mathbb{PP}_{r=1}^{N}\left[x(n)\right]$ finds the locations n of the N_p largest peaks of $x(n)$ (indexed by p). On these TOAs a histogram is used to derive an RIR-like function

$$h_{i,j}(m) = \underset{p,k}{\text{histogram}}\ t_{i,j,p}(k). \quad (9)$$

Locations of maxima in $h_{i,j}(m)$ then indicate the time-of-arrival $t_{i,j,r}$ of a broadband sound event r of source j at microphone i.

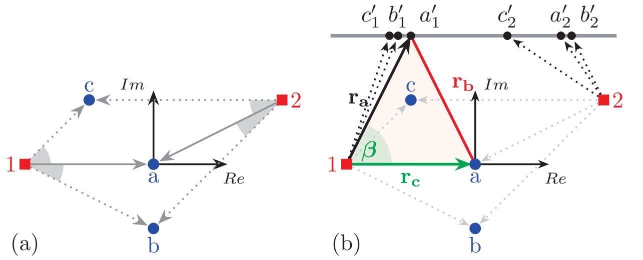

Figure 2: Example for self-calibration (a) and room inference (b) with three sources $j = \{a,b,c\}$, two microphone arrays $i = \{1,2\}$ and reflection points j_i' of single reflector.

Self-calibration

For the self-calibration (described in two dimensions for simplicity) only the direct sound DOAs and TOAs are needed. As the orientation of each microphone array is unknown, the DOAs of different direct sound events are used as direction-differences-of-arrival (DDOAs, $\Delta\varphi_{i,j}$), referenced to the DOA of an arbitrary reference source. Using these parameters, phasor systems containing all sources in a local microphone coordinate system can be constructed (shown in Figure 2a). The points of the phasor systems are computed according to

$$z_i = -(\Delta t_{i,j,0} + \tau_{i_{\mathrm{ref}},j}) \quad (10)$$

$$z_{i,j} = \left(z_i + (\Delta t_{i,j,0} + \tau_{i_{\mathrm{ref}},j})\cdot e^{i\Delta\varphi_{i,j}}\right)\cdot e^{i\phi_i} \quad (11)$$

where $z_{i,j}$ is the position of source j relative to microphone point z_i, ϕ_i are the unknown phasor system rotations (due to the unknown rotation of each array) and $\tau_{i_{\mathrm{ref}},j}$ the unknown times sound travels from the j-th source to the closest microphone i_{ref}. The optimal parameters for $\tau_{i_{\mathrm{ref}},j}$ and ϕ_i are found by minimizing the cost function

$$J(\tau_{i_{\mathrm{ref}},j},\phi_i) = \sum_{j=1}^{N_j}\sum_{i=1}^{N_i}\sum_{\substack{i'=1\\ i'\neq i}}^{N_i}|z_{i,j} - z_{i',j}| \quad (12)$$

which implies minimizing the cumulative distances between all relative source estimates. We obtain the optimal parameters using an iterative method: starting with a random initialization, we keep $\tau_{i_{\mathrm{ref}},j}$ fixed and find the optimal ϕ_i using a simple line search. Then we keep ϕ_i fixed and find the optimal $\tau_{i_{\mathrm{ref}},j}$ by solving the now convex problem. This procedure is repeated until convergence.

As the resulting points z_i and $z_{i,j}$ are in complex coordinates and have the unit seconds they have to be converted to meters and Cartesian coordinates using

$$\mathbf{z} = c\cdot\left(\mathrm{Re}\left[z\right]\quad \mathrm{Im}\left[z\right]\right)^T \quad (13)$$

where c is the speed of sound.

Geometry Inference

With the self-calibration results, the estimated DDOAs and TDOAs of detected first order reflections, reflection points can be computed using

$$r_c + \Delta t_{i,j,r} = r_a + r_b \quad (14)$$

$$r_b^2 = r_a^2 + r_c^2 - 2r_ar_c\cos\beta, \quad (15)$$

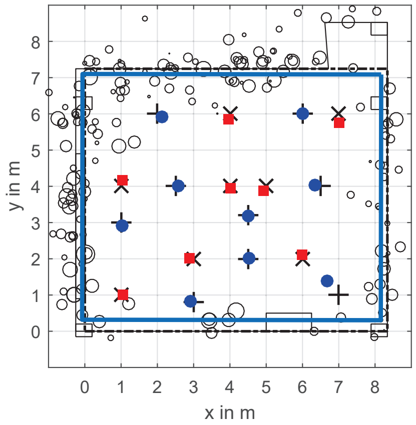

Figure 3: Reflection points, resulting room estimate and self-calibration results. Estimated reflection points are drawn as circles with size corresponding to weights indicating the similarity of the DOA estimates over all frequencies.

with r_a, r_b and r_c as the sides of a triangle and β as the reflection DDOA $\Delta\varphi_{i,j,r}$ (all indicated in Figure 2b for an exemplary reflection point). r_a is the quantity of interest. Inserting Eq. 14 into Eq. 15 results in

$$r_a = \frac{2r_c\Delta t_{i,j,r} + \Delta t_{i,j,r}^2}{2(r_c + \Delta t_{i,j,r}) - 2r_c\cos\beta}, \qquad (16)$$

which allows direct computation of first order reflection points relative to the corresponding source-microphone pairs. The computed reflection points need to be converted using Equation 13 as well.

Using the point clouds consisting of all estimated reflection points (see Figure 3), different methods for estimating corresponding arbitrary or rectangular room shapes can be used for geometry inference. For arbitrary room shapes the Hough transform (used in [6]) or a simple clustering of the reflection points by the reflector angle can be used. Rectangular room shapes can be found by fitting a rectangle to the reflection points (for example using a modified ellipse equation $\left(\frac{x}{a}\right)^{2\eta} + \left(\frac{y}{b}\right)^{2\eta} = 1$ with $\eta \geq 2$) or by projecting the points onto their principal components and computing histograms. For the results presented here we use a rectangular fit based on the modified ellipse equation.

Results

The algorithm performance is evaluated using data from two measurements, conducted in an absorptive measurement room and a box shaped lecture hall. A panorama view of the lecture hall can be seen in Figure 7, illustrating the microphone and calibration source positions and the room edges. All calibration sources (hand-claps) and microphones (B-format arrays) were located on the same height at measured positions in both measurements. The self-calibration results are evaluated as *mean position error* ϵ_s and ϵ_r (in m) for source and microphone positions and as *mean absolute orientation error* ϵ_ρ (in degrees) for the array orientations. Tables 1 and 2 show the numerical results for the measurement room and the lecture hall for six repetitions (one repetition corresponds to a single clap at every source position). For both scenes the best case results for microphone and source positions are below 10 *cm*. By combining all six measurements, we achieve results below the average error, with the position error of the microphones significantly lower than that of the calibration sources (the rows labelled *comb.*). The combination is achieved by optimally aligning all self-calibration results to a chosen result. The plots for these combined results are shown in Figures 4 and 5. The microphone orientations are indicated as lines in the estimated direction, the real orientation was in direction of the positive y-axis.

Table 1: Self-calibration results for the measurement room.

rep.	Calibration Error		
	ϵ_s/m	ϵ_r/m	ϵ_ρ/deg
1	0.1054	0.0949	3.78
2	0.1447	0.1037	5.06
3	0.0870	0.0597	2.55
4	0.0819	0.1230	2.55
5	0.2400	0.3521	9.30
6	0.3297	0.1468	5.12
comb.	**0.0927**	**0.0831**	**1.21**

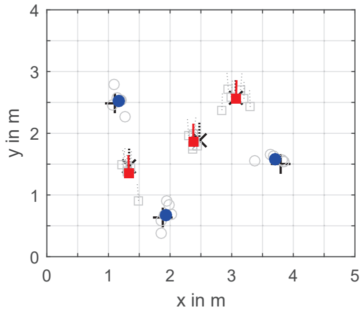

Figure 4: Self-calibration results for the measurement room, microphone estimates are indicated by red squares, source estimates by blue circles, the real positions by × and +. Repetition results are indicated by grey squares and circles.

The numerical results for the room inference are shown in Table 2 in the two rightmost columns as *average distance error* ϵ_d and *orientation error* ϵ_a of the estimated to the real walls. The combined results are again close to the best case results of 8 *cm* distance and less than 1° orientation error. For the combined results only repetitions 1 to 5 are used, as repetition 6 contributes all the large outliers in the self-calibration results (see Figure 5).

Conclusion and Future Work

A complete solution for acoustic scene parameter estimation is described, using simple ways to estimate the most important parts (microphone positions and reflective boundary) of a scene. The proposed self-calibration algorithm only needs a minimum number of microphones and calibration sources, equal to the dimension of the attempted scene map. The room inference procedure also allows simple estimation of arbitrary geometries.

Table 2: Calibration and inference results for the lecture hall.

rep.	Calibration Error			Inference Error	
	ϵ_s/m	ϵ_r/m	ϵ_ρ/deg	ϵ_d/m	ϵ_a/deg
1	0.1267	0.1318	2.80	0.1799	0.1
2	0.1674	0.1687	3.44	0.1744	2.14
3	0.2399	0.1327	1.81	0.0843	0.31
4	0.0874	0.1387	3.27	0.2136	0.81
5	0.1225	0.0962	3.43	0.1660	2.87
6	0.2042	0.3165	5.19	0.6134	8.17
comb.	**0.1240**	**0.0907**	**2.99**	**0.1077**	**0.89**

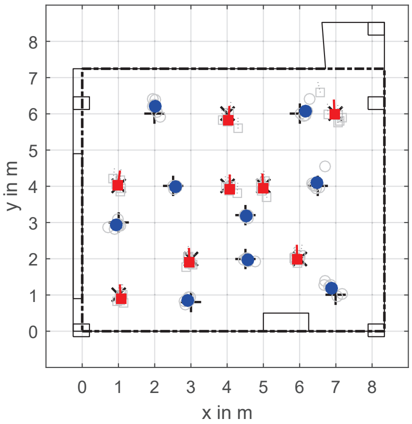

Figure 5: Self-calibration results for the lecture room, microphone estimates are indicated by red squares, source estimates by blue circles, the real positions by × and + respectively. Repetition results are indicated by grey squares and circles.

Work that still needs to be done is the evaluation of the estimated scene in terms of beamforming algorithms for localization or tracking of actual sources, as well as from an auditory point of view by comparing a model to the real counterpart. To simplify the calibration procedure it is also interesting to examine the influence of positions and numbers of calibration sources on the results, attempting to minimize the effort.

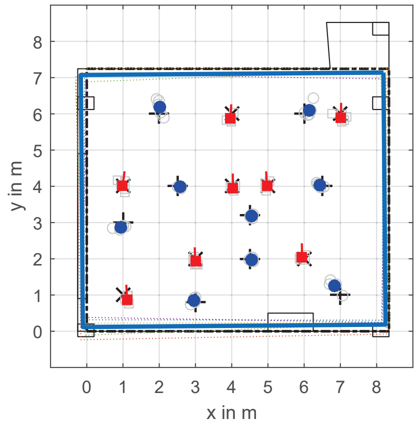

Figure 6: Final results combining 5 repetitions. The averaged room is indicated as a blue rectangle, sources and microphones as in Figures 4 and 5.

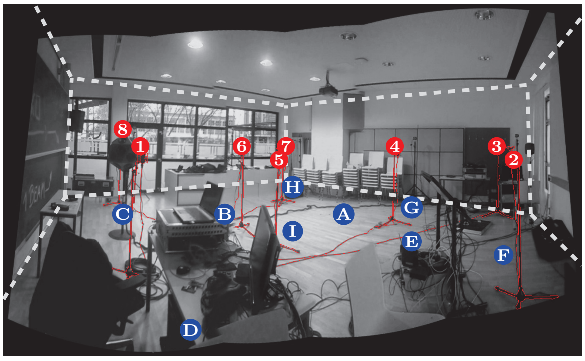

Figure 7: Panorama view of the measurement setup in the rectangular lecture hall. Microphones are indicated by red numbers, source positions as blue letters (projected onto the ground). The picture was taken in the top left corner in Figure 5 in direction of the bottom right corner.

References

[1] 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 *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*, p. 13–18, IEEE, 2010.

[2] M. Crocco, A. Del Bue, and V. Murino, "A bilinear approach to the position self-calibration of multiple sensors," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, p. 660–673, 2012.

[3] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 106–110, IEEE, 2013.

[4] S. Tervo and T. Korhonen, "Estimation of reflective surfaces from continuous signals," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 153–156, IEEE, 2010.

[5] S. Tervo, T. Korhonen, and T. Lokki, "Estimation of reflections from impulse responses," *Building Acoustics*, vol. 18, no. 1-2, p. 159–173, 2011.

[6] J. Filos, *Inferring Room Geometries.* PhD thesis, Imperial College London, 2013.

[7] I. Dokmanic, L. Daudet, and M. Vetterli, "From acoustic room reconstruction to slam," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6345–6349, IEEE, 2016.

[8] A. Politis, S. Delikaris-Manias, and V. Pulkki, "Direction-of-arrival and diffuseness estimation above spatial aliasing for symmetrical directional microphone arrays," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, p. 6–10, IEEE, 2015.