AN AUTOMATIC SYSTEM FOR MICROPHONE SELF-LOCALIZATION USING AMBIENT SOUND

Simayijiang Zhayida, Fredrik Andersson, Yubin Kuang and Kalle Åström

Centre for Mathematical Sciences, Lund University
{zhayida,fa,yubin,kalle}@maths.lth.se

ABSTRACT

In this paper, we develop a system for microphone self-localization based on ambient sound, without any assumptions on the 3D locations of the microphones and sound sources. We aim at developing a system capable of dealing with multiple moving sound sources. We will show that this is possible given that there are instances where there are only one dominating sound source. In the first step of the system we employ a feature detection and matching strategy. This produces TDOA data, possibly with missing data and with outliers. Then we use a robust and stratified approach for the parameter estimation. We use robust techniques to calculate initial estimates on the offsets parameters, followed by non-linear optimization based on a rank criterion. Sequentially we use robust methods for calculating initial estimates of the sound source positions and microphone positions, followed by non-linear Maximum Likelihood estimation of all parameters. The methods are tested and verified using anechoic chamber sound recordings.

1. INTRODUCTION

Time-of-arrival (TOA) and time-difference-of-arrival (TDOA) measurements are used in applications ranging from radio based positioning to beamforming and audio sensing. Although such problems have been studied extensively in the literature in the form of localization of e.g. a sound source using a calibrated detector array, see e.g. [1–4], the problem of self-calibration of a sensor array is still an open problem.

Several previous contributions of solving self-calibration problem rely on prior knowledge or extra assumptions of locations of the sensors to initialize the problem [5–10]. Iterative methods exist for TOA or TDOA based self-calibration [11, 12]. However, such methods are dependent on initialization and can get stuck in local minima. For a general graph structure, one can relax the TOA-based calibration problem as a semi-definite program [13].

Calculation of initial estimate for calibration of TOA sensor networks using only measurements without any prior estimate on the locations has been studied in [14, 15], where solutions to the minimal cases of three senders and three receivers in the plane, or six senders and four receivers in 3D are given. Initialization of TDOA networks is studied in [16] and refiend in [17] where a solution to non-minimal case of 9 receivers and 4 speakers in 3D was derived. In [18] and refined in [19] a far field approximation was utilized to initialize both TOA and TDOA problems. [14–19] attempt to solve the self-calibration problem with either minimal or close to minimal data. Studying minimal cases is both of theoretical importance and essential to develop fast stable algorithms suitable in random sample consensus (RANSAC) [20] schemes.

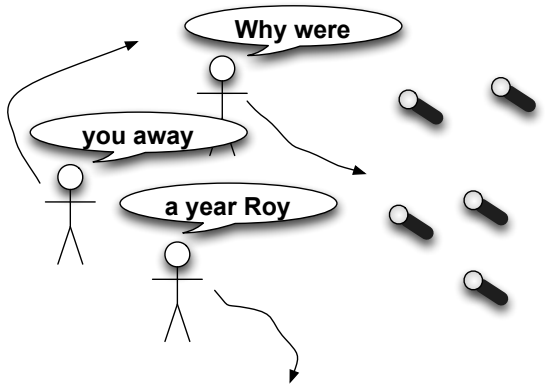

Fig. 1: The paper presents a automatic system for microphone self-localization based on ambient sound. In the experiments we use several moving sound sources as depicted in the figure.

In this paper we focus on a system approach. The input to the system is a sound recording with M channels (one from each microphone), see Figure 1. These are first processed to find the time-difference vectors. The estimated time-difference vectors typically contain noise, missing data and possible outliers. We then follow a stratified approach, where we first estimate offsets using a robust method. This is followed by a robust method for finding the 3D positions of all the microphones and the 3D positions of all sound sources.

We emphasize the need for further research within this system approach. Therefore we intend to publish our dataset. This dataset includes ground truth for time-difference shifts, for offsets and for 3D positions of microphones and sound sources. We hope that such data can be used by other researchers to empirically test algorithms for feature detection and matching as well as for estimating offsets and geometry of the microphone-speaker setup.

2. SYSTEM DESIGN

The input to the system consists of sound recordings with M channels $(\mathbf{x_1}, \ldots, \mathbf{x_M})$. The microphones are at unknown positions $(\mathbf{m_1}, \ldots, \mathbf{m_M})$. We assume that among the sounds there are one or several, possibly moving sound sources. This means that at several time instances along the sound channels there are one or several matchings. Each such match correspond to a set of time instants t_i of arrival times to the microphones. Each such time vector $(t_1, \ldots, t_M)$ correspond to a sound made at instant t_0 at 3D position $\mathbf{s}$ fulfilling

$$c(t_i - t_0) = ||\mathbf{m_i} - \mathbf{s}||,$$

where $||\cdot||$ is Euclidean norm, c is the speed of sound, assumed to be known and constant. Without loss of generality, we will in the sequel assume that all time differences are measured against channel 1. We introduce $u_i = c(t_i - t_1)$, which can be interpreted as

$$u_i = c(t_i - t_0) - c(t_1 - t_0) = ||\mathbf{m_i} - \mathbf{s}|| - ||\mathbf{m_1} - \mathbf{s}||. \quad (1)$$

In the sequel we will use the *matching vector* for time matchings of 'same signal' at some time instant in each channels, which is denoted as $(u_1, u_2, \ldots, u_M)^T$. We will allow missing data for such vectors, i.e. there might be one or several indices in a vector that has unknown values. The components of the vector might contain outliers. Also introduce $o = c(t_1 - t_0) = ||\mathbf{m_1} - \mathbf{s}||$ as the offset. This can be interpreted as the distance from the sound to microphone 1. Using this notation the measurement equation (1) becomes $u_i = ||\mathbf{m_i} - \mathbf{s}|| - o$. Let j be used as an index for different sounds. The key idea is that using a number of such measurements u_{ij} it is possible to estimate the unknown parameters $(\mathbf{m_i}, \mathbf{s_j}, o_j)$ so that

$$u_{i,j} = ||\mathbf{m_i} - \mathbf{s_j}|| - o_j.$$

The system has three components, see Figure 2.

- In signal processing step where the sound channels $\mathbf{x_1}, \ldots, \mathbf{x_M}$ are analyzed to extract number of time-matching vectors $\mathbf{u}_j = (u_{1j}, u_{2j}, \ldots u_{Mj})^T$.
- Robust estimation of the offsets o_j. Here we utilize the fact that the double compaction matrix of $U = (\mathbf{u}_1, \ldots, \mathbf{u}_N)$, where N is number of matching vectors which is find in signal processing step, with respect to the correct offset has rank 3, [21] and utilize minimal solvers from [17] in a RANSAC fashion, cf. [20] followed by non-linear optimization.
- Robust estimation of the remaining parameters $\mathbf{m_i}$ and $\mathbf{s_j}$. Here we follow the reconstruction techniques from [15].

2.1. Time-Difference Estimation

Experiments were made in different environments (normal and echo-free). Different types of sound sources were used (claps, voices, continuous songs) and have tried single moving sound sources and multiple moving sound sources. The current system uses two types of signal processing components. For claps we use flank detectors that detect the onset times t_{ij} of such claps. We then use a simple matching scheme to match the onset times along the different sound channels. Here we assume that the different claps do not occur too frequently. For other sound sources we search for time differences between channel 1 and channel i using different errors measures. We have used GCC-PHAT (Generalized Cross Correlation with Phase Transform), normalized cross-correlation and similar techniques. In the current system we are analyzing channel 1 vs channel i for $i = 2, \ldots, M$ at 1000 positions along the track. At each such time instant we compare channel 1 and channel i with shifts from -500 sample points to +500 sample points. See Figure 3 , where the matching score is shown for each position (x-axis) and for each shift (y-axis). If there are sufficient number of confident matching scores for most channels at a position, a matching vector $\mathbf{u} = (u_1, u_2, \ldots, u_M)^T$ is generated.

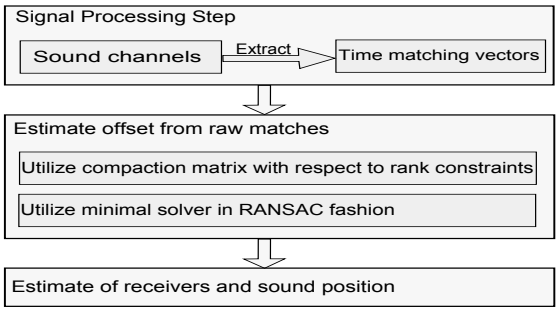

Fig. 2: Block diagram.

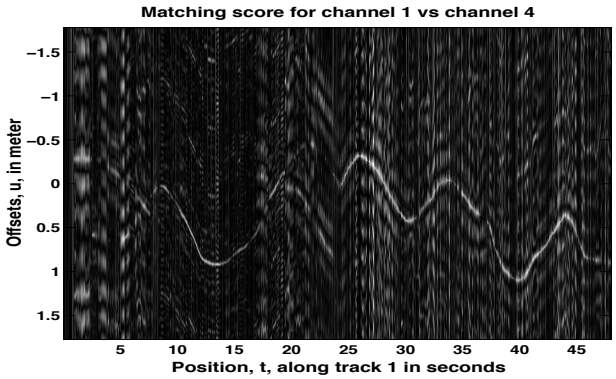

Fig. 3: The figure shows the matching score for different shifts (y-axis) at 1000 equally spaced time instants of the 48 seconds long recording.

2.2. Offset Estimation

In this section we derive the rank constraint, which can be used for estimating the offsets o_j from elements of the matching vectors u_{ij}, where i ranges from 1 to M (number of chan-

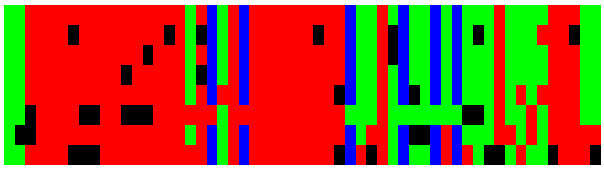

Fig. 4: The robust parameter estimation is based on the RANSAC paradigm. In the illustration there are 8 channels (rows) and 56 (out of 129) matching vectors (columns), we choose the smaller set of matching vectors only for better display reason. A random subset of 7 channels and 6 matching vectors (illustrated in blue) are used to estimate the parameters. The remaining data is used to verify (or falsify) the estimate. For this starting point there are a substantial number of inliers (illustrated with green) indicating that this is indeed a promising estimate. Outliers are shown in red and missing data are shown in black.

nels) and j is ranges from 1 to N (number of matching vectors) . The derivation follows that of [17, 21].

Notice that we have

$$(u_{ij} - o_j)^2 = ||\mathbf{m_i} - \mathbf{s_j}||^2.$$

By constructing the vectors $\tilde{\mathbf{M}}_i = \begin{bmatrix} 1 & \mathbf{m}_i^T & \mathbf{m}_i^T\mathbf{m}_i \end{bmatrix}^T$ and $\tilde{\mathbf{S}}_j = \begin{bmatrix} \mathbf{s}_j^T\mathbf{s}_j - o_j^2 & -2\mathbf{s}_j^T & 1 \end{bmatrix}^T$, we obtain $u_{i,j}^2 - 2u_{i,j}o_j = \tilde{\mathbf{M}}_i^T\tilde{\mathbf{S}}_j$. By collecting $\tilde{\mathbf{M}}_i$ and $\tilde{\mathbf{S}}_j$ into matrix $\tilde{\mathbf{M}} \in \mathbb{R}^{5\times M}$ and $\tilde{\mathbf{S}} \in \mathbb{R}^{5\times N}$, we have $\mathbf{D} = \tilde{\mathbf{M}}^T\tilde{\mathbf{S}}$, where $\mathbf{D}$ is the $M \times N$ matrix with elements $d_{i,j} = u_{i,j}^2 - 2u_{i,j}o_j$. This suggests that matrix $\mathbf{D}$ is at most of rank 5 as we increase M and N.

Now form the matrix $\mathbf{F} = \mathbf{C}_M^T\mathbf{D}\mathbf{C}_N$, where $\mathbf{C}_M = \begin{bmatrix} -\mathbb{1}_{M-1} & \mathbf{I}_{M-1} \end{bmatrix}^T$, $\mathbf{C}_N = \begin{bmatrix} -\mathbb{1}_{N-1} & \mathbf{I}_{N-1} \end{bmatrix}^T$, and where $\mathbb{1}_{N-1}$ is a $(N-1)\times 1$ vector with 1 as entries and $\mathbf{I}_{N-1}$ is identity matrix of size $(N-1)$. This matrix $\mathbf{F}$ has size $(M-1)\times(N-1)$. The elements are $f_{i,j} = d_{i,j} - d_{i,1} - d_{1,j} + d_{1,1}$. It is relatively easy to see that $\mathrm{rank}(\mathbf{F}) \leq 3$. In fact

$$\mathbf{F} = \mathbf{M}^T\mathbf{S},$$

where $\mathbf{M} = -2\begin{bmatrix} \mathbf{m}_2 - \mathbf{m}_1, \ldots, \mathbf{m}_M - \mathbf{m}_1 \end{bmatrix}$ and $\mathbf{S} = \begin{bmatrix} \mathbf{s}_2 - \mathbf{s}_1, \ldots, \mathbf{s}_N - \mathbf{s}_1 \end{bmatrix}$. The matrix $\mathbf{F}$, here called the *double compaction matrix*, depends on the matching vectors $\mathbf{u_j}$ and on the offsets o_j.

There are three minimal problems for determining offsets so that the the double compaction matrix $\mathbf{F}(\mathbf{U}, \mathbf{o})$ has rank 3, cf. [17]. The minimal problems are

- 9 microphones and 5 sounds (unique solutions).
- 7 microphones and 6 sounds (five solutions).
- 6 microphones and 8 sounds (14 solutions).

For all of these problems there are efficient closed form algorithms for finding all solutions.

Offsets and inlier set is estimated using the following RANSAC paradigm.

1. Randomly select a subset of 7 channels and 6 matching vectors. Use the closed form algorithm for finding the offsets o_j for these 6 matching vectors. The relevant (valid) solutions should have offsets that are real, and since $||\mathbf{m_i} - \mathbf{s_j}|| = u_{ij} - o_j$, the offsets should fulfill the constraints $u_{ij} \geq o_j$ for $i = 1, \ldots, M$ and $j = 1, \ldots, N$. Ignore solutions that do not fulfill these constraints.
2. For each solution study how many of the remaining matching vectors that fulfill the geometric constraint.
3. Repeat (1) and (2) a fixed number of times and choose the solution with the maximum number of inlier matching vectors.

The output of the RANSAC loop is a selection of a subset of the data that is considered to be inliers together with an initial estimate of offsets o_j.

Rank-based Nonlinear Optimization

In this section, we derive a iterative nonlinear optimization method for improving the estimate of the unknown offsets o_j. While the non-iterative schemes presented above apply only to specific number of M and N with no missing data, the method present in this section cope with such cases naturally.

Given the knowledge that the measurement matrix after compaction is of rank K, we can derive another scheme based nonlinear optimization to estimate the offsets $\mathbf{o} = (o_1, o_2, \ldots, o_N)$. The idea is to find the offset such that the measurement matrix after compaction is as close to a rank-K matrix as possible. Thus, we have the following minimization problem:

$$\begin{aligned} \min_{\mathbf{o},\mathbf{A}} \; & ||\mathbf{F}(\mathbf{U}, \mathbf{o}) - \mathbf{A}||_{F,\Omega} \\ s.t. \; & \mathrm{rank}(\mathbf{A}) = K, \end{aligned} \tag{2}$$

where $\mathbf{F}$ is the matrix resulting from the compaction operators as in the previous section, $\mathbf{U}$ and $\mathbf{o}$ are introduced in section 2, $\mathbf{A} \in \mathbb{R}^{(M-1)\times(N-1)}$, and $||\cdot||_{F,\Omega}$ is the Frobenious norm on the matrix entries that are observed specified by the set Ω.

Similar formulation of the minimization problem (2) has been proposed in [22] to utilize the rank constraints. Given that rank constraint on $\mathbf{A}$, the minimization problem is non-convex. In [22], an alternating scheme is proposed. To be more specific, one first fixes the offsets $\mathbf{o}$, and solve for the optimal $\mathbf{A}$ using SVD. Then, one fixes $\mathbf{A}$, the problem of finding the optimal $\mathbf{o}$ is convex. However, the rate of convergence of this alternative scheme is very slow. Thus, an additional regularization term on $\mathbf{A}$ is introduced to speed up the convergence. Here, we used a gradient descent scheme that utilize a local parameterization of the rank constraints on $\mathbf{A}$ directly.

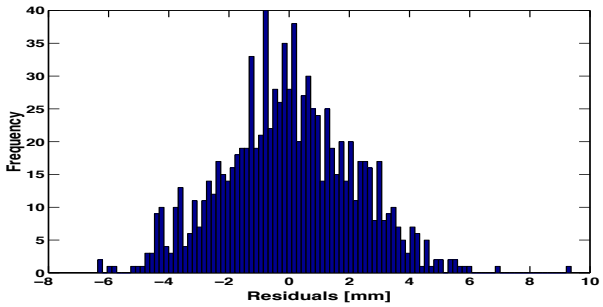

Fig. 5: Histogram of the residuals between the measured data u_{ij} and the fit $||\mathbf{m}_i - \mathbf{s}_j||_2 + o_j$.

2.3. Microphone and sound source position estimation

Once we have calibrated the measurement matrix with the offsets $\{o_j\}$, we proceed to solve the locations of $\{\mathbf{m}_i\}$ and $\{\mathbf{s}_j\}$ as a TOA problem. We follow the two-step technique in [15]. Specifically, after a factorizing the rank 3 matrix $\mathbf{F}$ above, one obtains $N-1$ linear equations and $M-1$ polynomial equations. Here we use the linear equations only to obtain parameters that then can be used as initial estimates to local optimization of the non-linear least squares

$$\min_{\mathbf{m}_i, \mathbf{s}_j, o_j} \sum_{ij} \left(u_{ij} - (||\mathbf{m}_i - \mathbf{s}_j||_2 + o_j)\right)^2 \qquad (3)$$

using standard techniques (Levenberg-Marquart) in order to obtain the maximal likelihood estimate of the parameters. Here it is also useful to improve the estimates u_{ij} and also to estimate the variance according to [23].

3. EXPERIMENTAL VALIDATION

We have made several experiments with 8 microphones (Shure SV100). These are connected to an audio interface (M-Audio Fast Track Ultra 8R) connected to a laptop. The microphones were positioned in a room with approximate distance 0-2 meters from each other. We generated sounds in several scenarios which is:

- Random distinct sound bursts made by banging two spoons together. This produces a set of discrete sound events that are relatively easy to detect and match.
- One continuously moving sound source playing part of a song. This produces a set of smoothly changing time-differences. If this is known, tracking techniques (Kalman filter, Particle filter) could be used to track the changes.
- Several continuously moving sound sources.
- Mixture of several people talking, clapping, walking around in the room.

The 8 sound channels were sampled at 96000 Hz.

We illustrate some of the steps of the automatic system with one of the experiments. In this case we have two moving sound sources in 3D. Figure 3 shows a plot of the matching score for different shifts between the channel 4 and channel 1, i.e. $u_4 = c(t_4 - t_1)$ on the y-axis at 1000 equally spaced positions along the 48 seconds long recording. The matching algorithm produces 129 matching vectors. There are 83 missing data among these $129 \times 8 = 1032$ time difference measurements. The RANSAC algorithm finds an inlier set of 75 (out of the 129) matching vectors. This is illustrated in Figure 4 where each dot corresponds to a measurement. Missing data are indicated as absence of a dot. The RANSAC algorithms selects random subsets of 7 rows and 6 columns. One such random selection is illustrated with blue dots. The inlier data from the algorithm are illustrated with green dots.

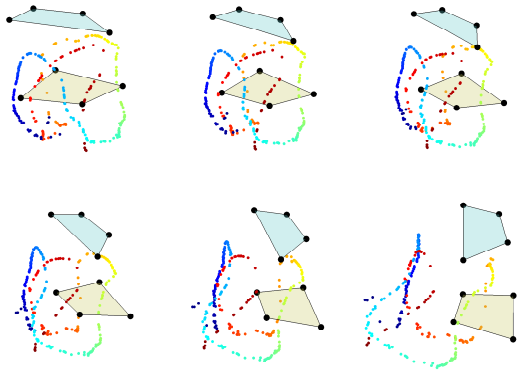

Fig. 6: Three-dimensional reconstruction of the microphone setup (black dots) as well as sound source positions (coloured dots). Note that the microphones in the experiment setup are located in two planes (also indicated in the figure).

These 75 inlier matching vectors are then used to estimate the 3D positions for the senders and receivers. A histogram of the residuals $u_{ij} - (||\mathbf{m}_i - \mathbf{s}_j||_2 + o_j)$ is shown in Figure 5. The errors are in the order of a few millimeters. The final 3D reconstruction of the microphones and of the sound source paths for one of the experiments are shown in Figures 6. In this experiment we have microphones in two planes (four in each). The moving sound source starts outside the convex hull of the microphones, then moves inside the microphone cluster and then out again.

To validate the method we have used several independent recordings. These have different sound source positions, but identical microphone setup. The error between the three reconstructions and the mean has a standard deviation of about 1 cm, indicating the accuracy of the system.

4. CONCLUSION

In this paper, we have developed an automatic system for microphone self-localization using ambient sound. The system does not put any constraints on the motion of the sound sources in relation to the microphone array setup. The system is based on a first finding several time-difference matching vectors for the recording. These are then used as input

to robust geometric algorithms based on minimal solvers and RANSAC to provide initial estimates of the unknown parameters, i.e. the offsets and the 3D positions of the sound sources and the receivers. These estimates are then improved by non-linear optimization to obtain the maximum likelihood estimate of the parameters. The components of the system as well as the system as a whole has been tested on real data with single and multiple moving sound sources. The automatic sound matching works well with a single moving sound source. For the case of multiple sound sources the current system works well, provided that there are instances where there is only one dominating sound source. In the case where there is a dominating sound source we can successfully reconstruct the locations of the microphone. This could possibly be used as a calibration step in the case of general multiple sound sources to further improve the detection part of the matching vectors.

REFERENCES

[1] M.S. Brandstein, J.E. Adcock, and H.F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 1, pp. 45 –50, Jan. 1997.

[2] A. Cirillo, R. Parisi, and A. Uncini, "Sound mapping in reverberant rooms by a robust direct method," in *Acoustics, Speech and Signal Processing, IEEE International Conference on*, April 2008, pp. 285 –288.

[3] M. Cobos, A. Marti, and J.J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *Signal Processing Letters, IEEE*, vol. 18, no. 1, pp. 71 –74, Jan. 2011.

[4] Hoang Do, H.F. Silverman, and Ying Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction(SRC) on a large-aperture microphone array," in *ICASSP 2007*, April 2007, vol. 1, pp. 121–124.

[5] S. T. Birchfield and A. Subramanya, "Microphone array position calibration by basis-point classical multidimensional scaling," *IEEE transactions on Speech and Audio Processing*, vol. 13, no. 5, 2005.

[6] D. Niculescu and B. Nath, "Ad hoc positioning system (aps)," in *GLOBECOM-01*, 2001.

[7] V. C. Raykar, I. V. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE transactions on Speech and Audio Processing*, vol. 13, no. 1, 2005.

[8] M. Crocco, A. Del Bue, M. Bustreo, and V. Murino, "A closed form solution to the microphone position self-calibration problem," in *ICASSP*, March 2012.

[9] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE transactions on Signal Processing*, vol. 50, 2002.

[10] P. Pertila, M.S. Hamalainen, and M. Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 11, pp. 2393–2402, Nov. 2013.

[11] R. Biswas and S. Thrun, "A passive approach to sensor network localization," in *IROS 2004*, 2004.

[12] J. Wendeberg, F. Hoflinger, C. Schindelhauer, and L. Reindl, "Anchor-free TDOA self-localization," in *Indoor Positioning and Indoor Navigation (IPIN), 2011 International Conference on*, Sept. 2011, pp. 1 –10.

[13] P. Biswas, T.C. Lian, T.C. Wang, and Y. Ye, "Semidefinite programming based algorithms for sensor network localization," *ACM Transactions on Sensor Networks (TOSN)*, vol. 2, no. 2, pp. 188–220, 2006.

[14] H. Stewénius, *Gröbner Basis Methods for Minimal Problems in Computer Vision*, Ph.D. thesis, Lund University, April 2005.

[15] Y. Kuang, S. Burgess, A. Torstensson, and K. Åström, "A complete characterization and solution to the microphone position self-calibration problem," in *ICASSP*, 2013.

[16] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data," in *Proc. of ICASSP*, 2008.

[17] Yubin Kuang and Kalle Åström, "Stratified sensor network self-calibration from tdoa measurements," in *21st European Signal Processing Conference 2013*, 2013.

[18] S. Thrun, "Affine structure from sound," in *Proc. of NIPS*, 2005.

[19] Y. Kuang, E. Ask, S. Burgess, and K. Åström, "Understanding toa and tdoa network calibration using far field approximation as initial estimate," in *ICPRAM*, 2012.

[20] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–95, 1981.

[21] F. Jiang, Y. Kuang, and K. Åström, "Time delay estimation for tdoa self-calibration using truncated nuclear norm," in *Proc. of ICASSP*, 2013.

[22] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *ICASSP*, 2013.

[23] K. Åström and A. Heyden, "Stochastic analysis of image acquisition, interpolation and scale-space smoothing," *Advances in Applied Probability*, vol. 30, no. 1, 1999.