A CLOSED FORM SOLUTION TO THE MICROPHONE POSITION SELF-CALIBRATION PROBLEM

Marco Crocco, Alessio Del Bue, Matteo Bustreo and Vittorio Murino

Istituto Italiano di Tecnologia (IIT)

ABSTRACT

This paper presents a novel algorithm for the automatic 3D localization of a set of microphones in an unknown environment. Given the times of arrival at each microphone of a set of sound events, the approach simultaneously estimates the 3D positions of the sensors and the sources that have generated the events. The only assumption made is that the emission time of the sound events must be known in order to measure the time of flight for each event. A closed form solution is also proposed whenever a sound event coincides with a microphone position. Simulated and real experiments show the validity of the approach for different setups of sensors and number of events.

Index Terms— Microphone calibration, source localization, factorization, closed form, bilinear optimization

1. INTRODUCTION

The localisation of a set of michrophones in an unknown environment is a longstanding problem that heavily impacts the practical deployment of acoustic systems. Several applications using microphone arrays require the position of the array element to be precisely known. An unreliable localisation degrades performance of both beamforming and direction of arrival estimation techniques. When a large number of sensors are deployed, their precise localisation in a three dimensional world may be a critical and time consuming task. A possible solution is to calibrate the microphones' positions simply using a set of predefined sound events acquired by each sensor. This *self-calibration* of the microphone's positions using solely sound events is still an open issue of research and a very challenging problem.

Standard solutions revert to the use of manual (e.g. by tape) measured pairwise distances among all the microphone pairs and applying algorithms such as multidimensional scaling (MDS) [1] to recover their spatial locations. However, if the number of microphones is rather large or if they are placed in configurations not easily reachable by a person, such a procedure may become tedious and cumbersome in most applications. In this case, it can be convenient to exploit acoustic sources measuring the time of arrival (TOA) [2], the phase [3, 4] or the intensity [5] of the signals acquired by each microphone. Since also the acoustic sources positions are generally unknown, this approach leads to the minimization of a nonlinear cost functions which can be easily trapped into local minima. Some methods try to overcome this drawback by introducing some additional constraints. For example in [6] the former TOA-based approach of [2] is simplified by assuming that all the acoustic sources are in the far field in respect to the microphones. In such scenario Thrun's intuition [6] was of modelling the sensors and event locations as two bilinear factors in the direction of arrival and microphone positions. This formalisation makes evident a rank constraint in the matrix containing the TOA that can be used to efficiently optimise for the unknown positions.

Even if the framework is elegant and efficient, Thrun's approach is restricted to the far-field case thus limiting its application in practical scenario. Here we present a novel formulation of the microphone position *self-calibration* problem that can deal explicitly with microphones located in the near-field of the sound events. The proposed method finds an approximate solution of a maximum likelihood (ML) problem by transforming the original nonlinear least squares cost function minimisation into a two step procedure. In the first step a Singular Value Decomposition (SVD) is employed to reduce the unknowns from $3(N+M)$ to just nine parameters (N and M being the number of microphones and sources respectively).

In the second step such nine parameters are estimated by solving a nonlinear least square problem, by far much simpler than the original one. Moreover, if the position of just one sound event is coincident with one microphone, the second step can be solved with a single linear least squares procedure, yielding to a full closed form solution. The proposed method uses the whole information of the measured time-of-flight (TOF) and it is based on a simple principle: even if we have measured TOF at different time instants, the microphones do not vary their position in time. This rigidity hypothesis generates rank constraints over the matrix storing the measured TOF and it can be used to find such a solution.

2. PROBLEM FORMULATION

Let us consider N microphones that lay in unknown positions and let us define the 3D coordinates of the *i*-th microphone with $\mathbf{x}_i = (x_{i1},\ x_{i2},\ x_{i3})^\top$. Similarly, let us consider the

M sound events and let us define $\mathbf{a}_j = (a_{j1},\ a_{j2},\ a_{j3})^\top$ the unknown 3D coordinates of the *j*-th event. The difference between the measured arrival time of the event *j* at the sensor *i* and the emission time of the same event (i.e., the measured TOF) can be expressed as:

$$t_{i,j} = c^{-1}|\mathbf{x}_i - \mathbf{a}_j| + n_{ij}, \tag{1}$$

where c is the signal propagation speed and $n_{i,j}$ is a realization of an i.i.d zero mean Gaussian random variable representing the measurement error. The estimated distance between the sensor i and the event j is simply $d_{i,j} = c \cdot t_{i,j}$. It can be demonstrated [2] that the ML estimation of the 3D positions is given by:

$$\min_{\mathbf{x}_i,\mathbf{a}_j} \sum_{i=1}^{N}\sum_{j=1}^{M} \left(|\mathbf{x}_i - \mathbf{a}_j| - d_{i,j}\right)^2. \tag{2}$$

The minimization in (2) is difficult because of the presence of several local minima. The presence of such local minima is common to many ML approaches to sensors position calibration [2, 6, 3] and it is due to the squared roots of the Euclidean distances between sensors and sources in eq. (2). A gradient based method with an initial random guess will work poorly in practice, especially in the case of significant measurement errors and higher number of microphones and/or sound events. Therefore a method for identifying a good initial choice of $\mathbf{x}_i$ and $\mathbf{a}_j$ is mandatory.

2.1. A rank constraint in near-field

In a noiseless case and with the assumption of no TOF measurement errors, the following set of NM equations hold for $i = 1 \dots N$ and $j = 1 \dots M$:

$$|\mathbf{x}_i|^2 + |\mathbf{a}_j|^2 - 2\mathbf{x}_i \cdot \mathbf{a}_j = d_{i,j}^2. \tag{3}$$

From this equation, it is possible to obtain a bilinear form in the sensors and events coordinate vectors if the quadratic terms $|\mathbf{x}_i|^2$ and $|\mathbf{a}_j|^2$ can be eliminated. This can be obtained by subtracting the $(1, j)$-th equation to the (i, j)-th equation in (3) for $i = 2 \dots N$ and $j = 1 \dots M$, giving a set of $(N - 1)M$ equations such that:

$$|\mathbf{x}_i|^2 - |\mathbf{x}_1|^2 - 2(\mathbf{x}_i - \mathbf{x}_1) \cdot \mathbf{a}_j = d_{i,j}^2 - d_{1,j}^2. \tag{4}$$

Similarly, by subtracting the $(i, 1)$-st equation to the (i, j)-th equation in (4) for $i = 2 \dots N$ and $j = 2 \dots M$, we obtain a set of $(N - 1)(M - 1)$ equations as:

$$-2(\mathbf{x}_i - \mathbf{x}_1) \cdot (\mathbf{a}_j - \mathbf{a}_1) = d_{i,j}^2 - d_{1,j}^2 - d_{i,1}^2 + d_{1,1}^2. \tag{5}$$

The terms related to the microphones position can be then organized in a $(N-1)\times 3$ matrix $\mathtt{X}$ such that $\mathtt{X} = \{x_{id} - x_{1d}\}$ where $i = 2, \dots, N$ and $d = 1, 2, 3$ (i.e. the 3D coordinates of the sensor). Likewise, we can form a $(M - 1) \times 3$ matrix such that $\mathtt{A} = \{a_{jd} - a_{1d}\}$ where $j = 2, \dots, N$. We also define the distance differences as

$$\tilde{d}_{ij} = d_{ij}^2 - d_{1j}^2 - d_{i1}^2 + d_{11}^2, \tag{6}$$

that can be stored in a $(N-1)\times(M-1)$ measurement matrix $\mathtt{D} = \{\tilde{d}_{i,j}\}$. We can than write the collection of equations in (5) as a bilinear product in $\mathtt{X}$ and $\mathtt{A}$ as:

$$-2\mathtt{X}\mathtt{A}^\top = \mathtt{D}. \tag{7}$$

The matrix $\mathtt{D}$ has a rank three constraint since $\mathtt{D}$ is a product between the $(N - 1) \times 3$ matrix $-2\mathtt{X}$ and the $3 \times (M - 1)$ matrix $\mathtt{A}^\top$. If we apply a SVD to the matrix $\mathtt{D}$ we have, in case of no noise, that the singular values after the third are equal to zero. Thus we can truncate these SVD components such as:

$$\mathtt{U}\mathtt{V}\mathtt{W} = \mathtt{D}, \tag{8}$$

where $\mathtt{U}$ is an $(N-1)\times 3$ matrix, $\mathtt{V}$ is a 3×3 diagonal matrix and $\mathtt{W}$ is a $3\times(M-1)$ matrix. In a practical situation, in presence of measurement noise, the rank of $\mathtt{D}$ will be higher than three: in this case only the three highest singular values in $\mathtt{V}$ will be considered reducing the size of $\mathtt{U}$, $\mathtt{V}$ and $\mathtt{W}$ according to the noise-free case. From (7) and (8), for whatever invertible 3×3 matrix $\mathtt{C}$, the following holds:

$$\mathtt{X} = \mathtt{U}\mathtt{C} \quad \text{and} \quad -2\mathtt{A}^\top = \mathtt{C}^{-1}\mathtt{V}\mathtt{W}. \tag{9}$$

In order to find the nine elements of the matrix $\mathtt{C}$, we define a non-linear least squares minimization problem using the equations in eq. (4), for $i = 2 \dots N$ and $j = 2 \dots M$, which bears the quadratic terms $|\mathbf{x}_i|^2$ previously discarded:

$$\min_{\mathbf{x}_i,\,\mathbf{a}_j} \sum_{i=2}^{N}\sum_{j=2}^{M} [|\mathbf{x}_i|^2 - |\mathbf{x}_1|^2 - 2(\mathbf{x}_i - \mathbf{x}_1) \cdot \mathbf{a}_j - d_{ij}^2 + d_{1j}^2]^2. \tag{10}$$

Notice that the whole sensors configuration is invariant to any translation, so we can enforce without loss of generality that the coordinates of the first sensor can be set to the global origin of the reference system i.e. $\mathbf{x}_1 = \mathbf{0}$. Moreover, the minimum solution is invariant to any rotation in the 3D-space, so that the first source can be constrained to lay on the x-axis of the reference system, yielding $a_{12} = 0$ and $a_{13} = 0$. Using these five constraints, Eq. (10) can be rewritten as a minimization problem in the entries of $\mathtt{A}$ and $\mathtt{X}$ such that:

$$\begin{aligned}\min_{\mathbf{x}_i,\,\mathbf{a}_j} \sum_{i=2}^{N}\sum_{j=2}^{M} &[|\mathbf{x}_i|^2 + |\mathbf{x}_1|^2 - 2(\mathbf{x}_i - \mathbf{x}_1) \cdot (\mathbf{a}_j - \mathbf{a}_1) + \\ &- 2(x_{i1} - x_{11})a_{11} - d_{ij}^2 + d_{1j}^2]^2.\end{aligned} \tag{11}$$

Finally (11) can be recast as a minimization problem in respect to the entries of $\mathtt{C}$ by substituting into (11) the known values of $\mathtt{X}$ and $\mathtt{A}$ given by the SVD in (9). In particular defining $(\mathtt{U}\mathtt{C})_{ik}$ as the ik-th element of the matrix $\mathtt{U}\mathtt{C}$ and $(\mathtt{U}\mathtt{V}\mathtt{W})_{ij}$ as the ij-th element of the matrix $\mathtt{U}\mathtt{V}\mathtt{W}$ we obtain :

$$\begin{aligned}\min_{\mathtt{C}} \sum_{i=1}^{N-1}\sum_{j=1}^{M-1} &[((\mathtt{U}\mathtt{C})_{i1})^2 + ((\mathtt{U}\mathtt{C})_{i2})^2 + ((\mathtt{U}\mathtt{C})_{i3})^2 + \\ &+ (\mathtt{U}\mathtt{V}\mathtt{W})_{ij} - 2(\mathtt{U}\mathtt{C})_{i1}a_{11} - d_{i+1j+1}^2 + d_{1j+1}^2]^2,\end{aligned} \tag{12}$$

Notice that the only unknown parameter in (12), apart from the matrix $\mathtt{C}$, is a_{11}. This parameter can be substituted with

the estimated distance between the first microphone and the first source, according to equation (3), since the coordinate value is univocally determined given the constraints on $\mathbf{x}_1$ and $\mathbf{a}_1$. Such value can be used as an initialization in order to optimize a_{11} together with the matrix $\mathtt{C}$. Alternatively, if the measurement noise is quite low, the coordinate a_{11} can be fixed to a constant. Even if the proposed procedure reduces the former non-linear optimization problem to a simpler one, the last step of finding the matrix $\mathtt{C}$ still results in a non-linear problem which may get stuck in residual local minima. However, we show that a further constraint on just one source and one sensor may lead to a closed form solution.

2.2. Closed Form Solution

If there is an additional assumption that a source position coincides with a microphone position, an alternative procedure can be derived, which allows to find a completely closed-form solution. In fact, if $a_{11} = x_{11}$, $a_{12} = x_{12}$ and $a_{13} = x_{13}$, and $\mathbf{x}_1 = \mathbf{0}$, the first source coordinates are set to the origin i.e. $\mathbf{a}_{11} = \mathbf{0}$. In this way the matrices of the sources and microphones positions can be redefined respectively as $\mathtt{A} = \{a_{jd}\}$ and $\mathtt{X} = \{x_{id}\}$. The nine elements of $\mathtt{C}$ can now be further reduced to six observing that the minimum solution is invariant to any rotation in the 3D-space. This is an intrinsic indeterminacy of the sensors and sound events localization problem since we can always obtain an orbit of minimal solutions by applying an arbitrary rotation to the sensors position and its inverse to the sound events position. In our case, any real square matrix admits a QR decomposition such that $\mathtt{C} = \mathtt{QR}$ where $\mathtt{Q}$ is a rotation matrix and $\mathtt{R}$ is an upper triangular matrix. Thus, without loss of generality, we can arbitrarily choose the $\mathtt{Q}$ matrix to be the identity matrix i.e. $\mathtt{Q} = \mathtt{I}$. In this way, we can simply substitute $\mathtt{C}$ with $\mathtt{R}$ in (9), obtaining:

$$\mathtt{X} = \mathtt{UR} \quad \text{and} \quad -2\mathtt{A}^{\mathtt{T}} = \mathtt{R}^{-1}\mathtt{VW}. \tag{13}$$

Defining the matrix $\mathtt{R}$ as follows:

$$\mathtt{R} = \begin{pmatrix} r_1 & r_2 & r_3 \\ 0 & r_4 & r_5 \\ 0 & 0 & r_6 \end{pmatrix}, \tag{14}$$

eq. (10) can be expressed in term of $\mathtt{R}$, using eq. (13):

$$\mathtt{R}^* = \arg\min_{\mathtt{R}} \sum_{i=1}^{N-1}\sum_{j=1}^{M-1} [(r_1 u_{i1})^2 + (r_2 u_{i1} + r_4 u_{i2})^2 + \\ + (r_3 u_{i1} + r_5 u_{i2} + r_6 u_{i3})^2 + (\mathtt{UVW})_{ij} - d^2_{i+1j+1} + d^2_{1j+1}]^2, \tag{15}$$

where u_{ik} denotes the ik-th element of the matrix $\mathtt{U}$. Developing the first three squared terms in (15) and grouping together all the other terms we obtain:

$$\mathtt{R}^* = \arg\min_{\mathtt{R}} \sum_{i=1}^{N-1}\sum_{j=1}^{M-1} [u^2_{i1}(r_1^2 + r_2^2 + r_3^2) + u^2_{i2}(r_4^2 + r_5^2) + \\ + u^2_{i,3} r_6^2 + 2u_{i,1}u_{i,2}(r_2 r_4 + r_3 r_5) + 2u_{i,1}u_{i,3}(r_3 r_6) + \\ + 2u_{i2}u_{i3}(r_5 r_6) - k_{ij}]^2, \tag{16}$$

where

$$k_{i,j} = -(\mathtt{UVW})_{ij} + d^2_{i+1j+1} - d^2_{1j+1}. \tag{17}$$

Defining the vector $\mathbf{s}_i$ for $i = 1 \dots N$, the vector $\mathbf{f}$, the $(N-1)(M-1)$ vector $\mathbf{k}$ and the $6 \times (N-1)$ matrix $\mathtt{S}$ respectively as:

$$\mathbf{s}_i = \begin{pmatrix} u^2_{i1} & u^2_{i2} & u^2_{i3} & 2u_{i1}u_{i2} & 2u_{i1}u_{i3} & 2u_{i2}u_{i3} \end{pmatrix}^\top;$$

$$\mathbf{f} = \begin{pmatrix} r_1^2 + r_2^2 + r_3^2 & r_4^2 + r_5^2 & r_6^2 & r_2 r_4 + r_3 r_5 & r_3 r_6 & r_5 r_6 \end{pmatrix}^\top;$$

$$\mathbf{k} = \begin{pmatrix} k_{1,1} & k_{2,1} & \dots & k_{N-1,1} & k_{1,2} & \dots & k_{N-1,M-1} \end{pmatrix}^\top;$$

$$\mathtt{S} = \begin{pmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \dots & \mathbf{s}_{N-1} \end{pmatrix}^\top;$$

and finally the $(N-1)(M-1) \times 6$ matrix $\mathtt{P}$ obtained stacking $M-1$ times the matrix $\mathtt{S}$, we can express eq. (15) as a linear least squares problem in $\mathbf{f}$ as follows:

$$\mathbf{f}^* = \arg\min_{\mathbf{f}} = (|\mathtt{P}\mathbf{f} - \mathbf{k}|)^2. \tag{18}$$

The closed form solution of eq. (18) is finally given by:

$$\mathbf{f}^* = (\mathtt{P}^\top \mathtt{P})^{-1} \mathtt{P}^\top \mathbf{k}. \tag{19}$$

The values of $\mathtt{R}$ can now be easily recovered from the six elements of $\mathbf{f}^*$ as follows:

$$\begin{aligned} r_6 &= \pm\sqrt{f_3}; & r_5 &= f_6/r_6; \\ r_4 &= \pm\sqrt{f_2 - r_5^2}; & r_3 &= f_5/r_6; \\ r_2 &= (f_4 - r_3 r_5)/r_4; & r_1 &= \pm\sqrt{f_1 - r_2^2 - r_3^2}; \end{aligned} \tag{20}$$

where f_i is the $i-$th element of $\mathbf{f}^*$. The sign ambiguities in (20) produce eight different $\mathtt{R}$ matrices corresponding to the combinations of the three specular reflections of the whole coordinate set of sensors and event sources. Given one of the $\mathtt{R}$ found by (20), the matrices $\mathtt{X}$ and $\mathtt{A}$ can be easily recovered. If necessary, one of the eight matrices can be selected using additional information about sensor/events position.

Concerning the computational complexity of the closed form solution, the procedure requires the solution of a least squares system of size MN and 6 using a pseudoinverse, leading to $O(2MN^2)$ if $M > 6$ or $O(N^3)$ if $M \leq 6$.

3. EXPERIMENTS

We use a synthetic setup to compare our approach against the gradient descent (GD) method used to the original maximum likelihood cost function (2). The GD method is applied in a twofold manner, taking as starting point either a random guess or the solution obtained with the proposed method. The experimental setup consists of a fixed number of sensors and 10 audio events randomly placed in a 3D cubic region of side equal to 1 m. A Gaussian random variable with zero mean and std of 0.002 m has been added to each microphone-source distance in order to simulate the Distance Measurement Errors (DME). We run 500 random trials for each configuration and we use the Mean Position Error (MPE), defined as the mean of the Euclidean distances between ground truth and estimated

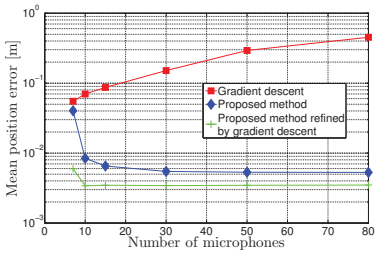

Fig. 1. MPE (logarithmic scale) versus the number of sensors for 10 sources and a standard deviation of 0.002 in the DME.

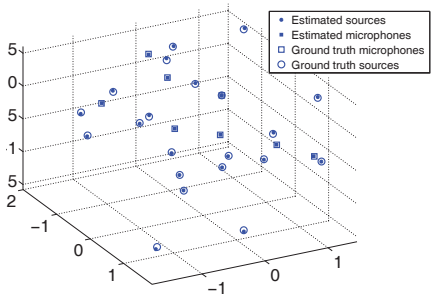

Fig. 2. Real and estimated three dimensional positions of microphones and acoustic sources obtained with the proposed method in a real environment.

positions, averaged over all the trials. In order to account for the intrinsic translational and rotational freedom of the solution, the estimated and real sensor positions were aligned using Procrustes Analysis. We test three methods: GD method, our closed form solution (CF) and the closed form plus an additional GD refinement (CF+GD). Concerning GD method, microphone and source positions were initialized at random coordinates inside the 3D cubic region. The obtained MPEs for each method are displayed in Fig. 1 versus the number of sensors. It can be seen that the proposed CF method outperforms the GD alone, regardless from the number of sensors. Moreover the CF+GD method provides a significant further improvement. Increasing the number of sensors results in a worst performance for the GD due to the increased probability of falling into local minima. Differently, the CF method, with or without the gradient refinement, provides a quasi constant performance with an MPE of about 0.0055 m for the CF method and 0.0035 m for the CF+GD. The gradient refinement improvement is stronger for low numbers of sensors (15 dB for 7 sensors) while it reaches an almost constant value of about 3.5 dB for higher numbers. Further details on the synthetic experiments can be found in [7].

A further experiment has been performed to test the method in a real environment [1]. Eight microphones have been placed in a room of $6 \times 4 \times 3$ m^3. The room is characterised by reflecting walls and significant acoustic noise given mainly by PC fans. An acoustic transducer is moved in 21 different positions, one of them being coincident with one microphone. For each transducer position, a linear sweep chirp pulse, of 5 s duration and about 10 kHz bandwidth is transmitted and acquired by the 8 microphones. Each of the 8 signals acquired is compressed by a matched filter to achieve the best time resolution. To evaluate the TOA, the time instant of the first peak (corresponding to the direct path) exceeding a given threshold is considered. Knowing the emission time the TOF has then been calculated. The ground truth positions is provided by a motion capture system (VICON). The CF method was applied to the matrix of measured distances obtaining a 3D reconstruction of microphone and source positions displayed in Fig. 2 together with the ground truth. The MPE for the microphones was 0.0043 m, while the MPE for the event sources was of 0.0125 m. The difference is mainly due to the fact that acoustic sources were more spread all over the room, while microphones were enclosed in a smaller volume.

[1] Real data setup is available at the website `www.isr.ist.utl.pt/~adb/code/`

4. CONCLUSIONS

We have presented a novel formulation for the microphone position self-calibration problem. Our solution accounts for a closed form-solution with a single additional constraint and it can obtain remarkable results in real acoustic scenario.

5. REFERENCES

[1] S.T. Birchfield and A. Subramanya, "Microphone array position calibration by basis-point classical multidimensional scaling," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 1025 – 1034, September 2005.

[2] R. Biswas and S. Thrun, "A passive approach to sensor network localization," in *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, September 2004, vol. 2, pp. 1544 – 1549 vol.2.

[3] A.J. Weiss and B. Friedlander, "Array shape calibration using sources in unknown locations-a maximum likelihood approach," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 12, pp. 1958 –1966, Dec. 1989.

[4] B. C. Ng and C. M. S. See, "Sensor-array calibration using a maximum-likelihood approach," *Antennas and Propagation, IEEE Trans. on*, vol. 44, no. 6, pp. 827 –835, June 1996.

[5] M. Chen, Z. Liu, L. He, P. Chou, and Z. Zhang, "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, Oct. 2007, pp. 22 –25.

[6] S. Thrun, "Affine structure from sound," in *Proc. of Conf. on Neural Information Processing Systems (NIPS)*. 2005, pp. 1353–1360, MIT Press.

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