DIRECT COMPUTATION OF SOUND AND MICROPHONE LOCATIONS FROM TIME-DIFFERENCE-OF-ARRIVAL DATA

Marc Pollefeys[*] *David Nister*

ETH Zürich and UNC-Chapel Hill
Department of Computer Science

Microsoft
Live Labs

ABSTRACT

In this paper we present a novel approach to directly recover the location of both microphones and sound sources from time-difference-of-arrival measurements only. No approximation solution is required for initialization and in the absence of noise our approach is guaranteed to always recover the exact solution. Our approach only requires solving linear equations and matrix factorization. We demonstrate the feasibility of our approach with synthetic data.

Index Terms— Self-localization, Microphone Array, Sensor Network, TDOA, Factorization

1. INTRODUCTION

The use of microphone arrays is now widespread. Microphone arrays enable beamforming and speaker tracking for applications such as smartrooms, surveillance and event recording. In addition there are also multiple applications for other accoustic sensor networks.

Typically the choice of phase-shifts used to focus on a particular spatial location is governed either by a priori knowledge of the spatial locations of the microphones, or by adaptation based on the signals received by the array. In the latter case, when adaptation takes place based on the sensor data, the approaches used aim to maximize the signal to noise ratio for a particular signal, and hence explicit knowledge of the spatial layout of the microphone array is not necessary.

However, knowledge of the spatial configuration of the microphones is necessary in order to do certain things, such as tracking of a speaker in 3D metric space and in general estimating the 3D locations of sources in space [1, 2]. To do this without estimating the 3D locations of the microphones would require tedious calibration of a mapping between phase-shifts and locations in 3D space.

On the other hand, the task of calibrating the 3D locations of the microphones is non-trivial. Apart from the obvious approach of physically measuring the locations, a few approaches have been suggested that allow more flexibility by augmenting the measurement situation in some way. Sachar et al. [3] suggest using a movable rig with sources in known configuration, which allows triangulation of the microphones. Raykar and Duraiswami [4] assume co-location of microphones and speakers, which simplifies the problem.

While these approaches provides nice workable solutions for many applications, ultimate flexibility would only come with the ability to self-calibrate the microphone array configuration using only the recorded sensor data from an arbitrary array with unknown spatial configuration. Moses et al. [5] described a 2D search to solve the initialization problem for a planar world. To our knowledge, no approach to attacking this problem in general in 3D has yet been proposed.

It is relatively straightforward to devise an algorithm that iteratively optimizes an estimate consisting of microphone spatial configuration and source timings in order to explain the time-difference-of-arrival (TDOA) measurements as well as possible [6]. Rockah and Schultheiss [7] investigate lower bounds on the sensitivity of such a potential approach to noise in the measurements. To achieve optimal accuracy, such an iterative refinement procedure should be used as the final stage of an algorithm. However, this requires a reasonably accurate initialization to converge to the global minimum.

Perhaps most related to our proposed approach is the work of Thrun [8] who presents a rank-3 factorization algorithm for solving for the sensor array, but under the assumption that the sources are far away from the microphone array (so that planar propagation fronts can be assumed), an assumption which does not hold in applications such as for example smart rooms or extended sensor network configurations. Our proposed approach is based on a rank-5 factorization which models spherical propagation fronts needed for the general case.

2. APPROACH

Let m sound sources indexed by i for $1 \leq i \leq m$ be represented by vectors $s_i = \begin{bmatrix} x_i & y_i & z_i \end{bmatrix}^\top$ where x_i, y_i, z_i are spatial coordinates and the t_i represent the (also unknown) times of departure. Similarly, let n microphones indexed by j for $1 \leq j \leq n$ be represented by spatial coordinates $m_j = \begin{bmatrix} X_j & Y_j & Z_j \end{bmatrix}^\top$. Let the measured time of arrival of source i at microphone j be t_{ij}. This would typically be

[*]Thanks to the David and Lucille Packard Foundation for funding and to Frank Dellaert for useful comments.

obtained by correlating the signal recorded by the different microphones. Then we have

$$(x_i - X_j)^2 + (y_i - Y_j)^2 + (z_i - Z_j)^2 = v^2(t_{ij} - t_i)^2, \quad (1)$$

where v is the speed of sound. This can be expanded into

$$S_i^\top M_j = v^2(t_{ij}^2 - 2t_{ij}t_i + t_i^2), \quad (2)$$

where

$$\begin{array}{rcl} S_i & = & [\ s_i^\top s_i \quad -2x_i \quad -2y_i \quad -2z_i \quad 1 \quad]^\top \\ M_j & = & [\ 1 \quad X_j \quad Y_j \quad Z_j \quad m_j^\top m_j \quad]^\top \end{array}$$

It should be noticed that one can also move the term in t_i^2 to the left handside of the equation, i.e.

$$S_i'^\top M_j = v^2(t_{ij}^2 - 2t_{ij}t_i) \quad (3)$$

with $S_i' = \left(S_i - [\ v^2 t_i^2 \quad 0 \quad 0 \quad 0 \quad 0\ ]^\top\right)$. Collecting all yields an $(m \times 5).(5 \times n) = m \times n$ matrix equation.

2.1. Computing the time of departure

In this paragraph we discuss how the times of departure t_i can be obtained when TDOA measurements for 5 sources are available for at least 10 microphones. It is obvious from the above that the $m \times n$ matrix T of coefficients $\{t_{ij}^2 - 2t_{ij}t_i\}$ has to be rank five. Let $A = \{t_{ij}^2\}$, $B = \{-2t_{ij}\}$ and D a diagonal matrix with t_i as entries. Then T can be written

$$T = A + DB, \text{ or alternatively } T = [\ I \quad D\ ] \begin{bmatrix} A \\ B \end{bmatrix} \quad (4)$$

to separate out the unknowns D. Since the first row of M contains only ones, there must exist a linear combination for each set of five independent rows of T that results in a row $[1 \ldots 1]$. Define $\bar{A} = [\ A_{i_1}^\top \quad A_{i_2}^\top \quad A_{i_3}^\top \quad A_{i_4}^\top \quad A_{i_5}^\top\ ]^\top$ and $\bar{B} = [\ B_{i_1}^\top \quad B_{i_2}^\top \quad B_{i_3}^\top \quad B_{i_4}^\top \quad B_{i_5}^\top\ ]^\top$ for a choice of rows i_1, i_2, i_3, i_4, i_5. Thus there must exist a vector C for which

$$C^\top [\ I \quad \bar{D}\ ] \begin{bmatrix} \bar{A} \\ \bar{B} \end{bmatrix} = [1 \ldots 1] \text{ or } X \begin{bmatrix} \bar{A} \\ \bar{B} \end{bmatrix} = [1 \ldots 1] \quad (5)$$

To compute X uniquely, the matrix $\begin{bmatrix} \bar{A} \\ \bar{B} \end{bmatrix}$ has to be of rank 10 which implies that this approach requires at least 10 microphones. The absolute timings can be obtained from X as $t_{i_k} = X_{k+5}/X_k$ and can be solved for five sources at a time. We arbitrarily divide the sources in groups of 5 to solve for t_i. Since for each source i a shift in t_i would result in a corresponding shift in all t_{ij} associated with that source, one is free to apply such a shift. For numerical reasons, it is therefore recommended for each source i to use zero-mean relative timings, i.e. subtract the average value $\sum_{j=1}^m t_{ij}$ from the corresponding t_{ij}. This was shown to improve the accuracy of the results in our experiments.

2.2. Refinement of the absolute timings

Once the complete vector of source timings t_i has been computed, the matrix given on the right-hand side of Eq. (3) is completely determined and can thus be factorized as follows in two rank 5 matrices:

$$v^2(t_{ij}^2 - 2t_{ij}t_i) = \hat{S}_i'^\top \hat{M}_j \quad (6)$$

In practice this can be achieved by using the singular value decomposition to obtain the closest rank-5 approximation of $v^2(t_{ij}^2 - 2t_{ij}t_i)$.

When more than the minimal number of sounds are available, these results can be refined as follows using a simple iterative procedure. Our goal is to minimize

$$\arg\min_{t_i, S_i, M_j} \|t_{ij}^2 - 2t_i t_{ij} - \frac{1}{v^2}\hat{S}_i'^\top \hat{M}_j\| \quad (7)$$

and we do this by alternating between minimizing with respect to variables indexed by i and by j. For each i, $(\hat{S}_i', t_i)$ can be computed using linear least-squares. The vectors $\hat{M}_j$ can be computed as $\hat{M}_j = \{\hat{S}_i'\}^\dagger \left(t_{ij}^2 - 2t_{ij}t_i\right)$ with $\dagger$ representing the Moore-Penrose pseudo inverse. While this iterative procedure is in principle optional, in the presence of noise it significantly improve the quality of the results. Our experiments show that a few iterations are in general sufficient. Notice that since both steps of our iteration minimize the same function, this approach is guaranteed to converge. After refinement $\hat{S}_i^\top$ is easily obtained from $\hat{S}_i'^\top$.

2.3. Computing source and microphone locations

Of course, at this stage the matrices $\hat{S}_i$ and $\hat{M}_j$ are only determined up to an arbitrary non-singular 5×5 matrix, and are related to S_i and M_i by a transformation matrix H as $S_i^\top M_j = \hat{S}_i^\top H^{-1} H \hat{M}_j$. We will compute this transformation as a concatenation of three transformation $H = H_Q H_S H_M$. The first transformation H_M will ensure that the first row of M_j is equal to $[1 \ldots 1]$. The second H_S will ensure that the same is true for the last row of S_i. Finally the transformation H_Q imposes the quadratic consistency constraints on M_j or S_i.

The transformation H_M can be written as $\begin{bmatrix} h_M^\top \\ 0 \quad I \end{bmatrix}$ and h_M can be computed by solving the linear system of equations $h_M^\top \hat{M}_j = 1$. This step requires at least 5 microphones. Similarly, the transformation H_S can be written as $\begin{bmatrix} I & \\ 0 & h_S \end{bmatrix}^{-1}$ and h_S can be computed by solving the linear system of equations $h_S \hat{S}_i = 1$. In fact the first element of h_S has to be zero (to avoid modifying the first row of $H_M \hat{M}_i$) and thus only four or more sources are required for this step.

The remaining constraints are quadratic in nature and ensure that the quadratic term in S_i and M_j are consistent with

the linear terms. The constraint on M_j can be written as:

$$M_j^\top B M_j = 0 \text{ with } B = \begin{bmatrix} 0 & 0 & 0 & 0 & -\frac{1}{2} \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -\frac{1}{2} & 0 & 0 & 0 & 0 \end{bmatrix} \quad (8)$$

Similar constraints can be written down for S_i. Here we will solely use the constraints for M_j as using a mix of constraints on M_j and S_i is non-trivial. Therefore,

$$\hat{M}_j^\top H^\top B H \hat{M}_j = 0 \quad (9)$$

By defining $\hat{M}'_j = H_S H_M \hat{M}_j$ and $Q = H_Q^\top B H_Q$ the following linear equation is thus obtained for the coeffients of the symmetric matrix Q:

$$\hat{M}_j^{'\top} Q \hat{M}'_j = 0 \quad (10)$$

As H_Q should leave the first row of $\hat{M}'_j$ and the last row of $\hat{S}'_i$ unchanged, it must have the following form:

$$H_Q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ . & . & . & . & 0 \\ . & . & . & . & 0 \\ . & . & . & . & 0 \\ . & . & . & . & 1 \end{bmatrix} \text{ and hence } Q = \begin{bmatrix} . & . & . & . & -\frac{1}{2} \\ . & . & . & . & 0 \\ . & . & . & . & 0 \\ . & . & . & . & 0 \\ -\frac{1}{2} & 0 & 0 & 0 & 0 \end{bmatrix}$$

Therefore, taking into account symmetry, Q only has ten degrees of freedom and these coefficients can be computed linearly given ten or more microphones (or alternatively sound sources) using Eq.(10). Then, it can be verified that a valid choice for H_Q is given by

$$H_Q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ & & & & 0 \\ t & & RK & & 0 \\ & & & & 0 \\ t^\top t - Q_{11} & 2\left(t^\top K - \begin{bmatrix} Q_{12} & Q_{13} & Q_{14} \end{bmatrix}\right) & & 1 \end{bmatrix} \quad (11)$$

with K the Cholesky factorization of the middle 3×3 part of Q and R and t representing the rotation and translation associated with the Euclidean and mirroring ambiguity of the reconstruction (e.g. choose $R = I$ and $t^\top = [000]$ for reconstruction). In summary, this approach to absolute arrival timing factorization requires 4/10 or 10/4 microphones and sound sources (where the computation of absolute timings from relative required a minimum of 10/5 microphones and source). It should be noted that a similar algorithm can be set up for any dimension and in particular in 2D where a rank 4 factorization is obtained.

2.4. Non-linear least-squares refinement

We have also implemented a non-linear least-squares approach which allows us to obtain a maximum-likelihood estimation under the assumption of independent Gaussian noise on the TDOA measurements. The approach minimizes the following expression:

$$\arg \min_{s_i, m_j, t_i} \sum_i \sum_j \left(\frac{1}{v} \sqrt{S_i^\top M_j} + t_i - t_{ij} \right)^2 \quad (12)$$

2.5. Planar array - 3D sound

The case of a planar microphone array recording 3D sound sources is very interesting and can be solved effectively up to a per-source mirroring ambiguity about the microphone array plane. Let us assume without loss of generality that the microphone array corresponds to the XY-plane. In this case we have: $(x_i - X_j)^2 + (y_i - Y_j)^2 + z_i^2 = v^2(t_{ij} - t_i)^2$, which can be expanded to

$$S_i^{'\top} M_j = v^2(t_{ij}^2 - 2 t_{ij} t_i), \quad (13)$$

where

$$\begin{array}{rcl} S'_i & = & \begin{bmatrix} s_i^\top s_i - v^2 t_i^2 & -2x_i & -2y_i & 1 \end{bmatrix}^\top \\ M_j & = & \begin{bmatrix} 1 & X_j & Y_j & X_j^2 + Y_j^2 \end{bmatrix}^\top \end{array}$$

The planar structure of the array thus causes the rank of the above matrices to drop to four. Therefore, we can use the approach described in Section 2.1 to compute the arrival timings, but enforcing the lower rank which can be done linearly using 8 or more microphones. Similarly, the approach described in Section 2.3 can be used with the quadratic constraints coming from the microphones. It should be noted that the z_i^2 term in S'_i means that in this case there are no quadratic constraints available for the sources. Once M_j and S'_i have been recovered, the location of microphones and sound sources can be extracted from them as follows:

$$\begin{array}{rl} s_i & = \begin{bmatrix} -\frac{1}{2} S'_{i2} & -\frac{1}{2} S'_{i3} & \pm \sqrt{S'_{i1} - \frac{1}{4}\left(S'^2_{i2} + S'^2_{i3}\right) + v^2 t_i^2} \end{bmatrix}^\top \\ m_j = & \begin{bmatrix} M_{i2} & M_{i3} & 0 \end{bmatrix}^\top \end{array}$$

Notice that the location of the sources can only be recovered up to a mirror ambiguity about the planar microphone array.

3. EXPERIMENTS

We perform several experiments on synthetic data, both for the general 3D case, as well as the case of 3D sources sensed by a planar array. For the 3D case, two different synthetic configurations are used. The first configuration consists of both microphones and sound sources being arbitrarily distributed over a unit cube. The second configuration consists of microphones being placed on a regular grid on three faces of the unit cube, with sound sources arbitrarily distributed in the unit cube. For the planar microphone array experiments, we also distribute the sources randomly over a unit cube. In this case the microphone array consists of a regular grid on one

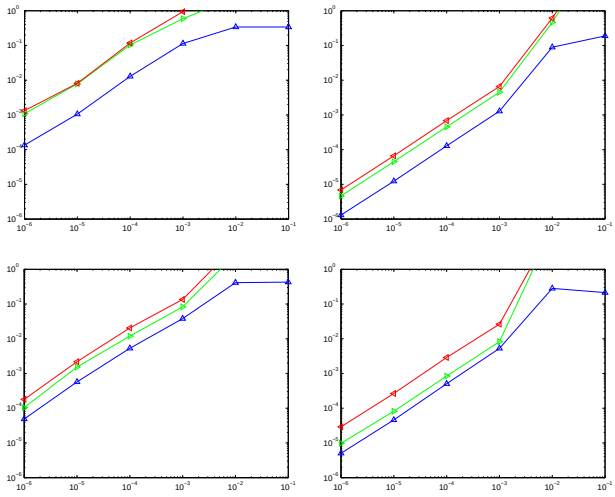

Fig. 1. Figures illustrating the effect on noise on the estimation of absolute time differences $\triangle$, microphone localization $\triangleright$ and source localization $\triangleleft$. $3D$ sound-$3D$ array localization ($5s/10m$ and $20s/20m$) (top) and $3D$ sound-$2D$ array localization ($5s/9m$ and $20s/20m$) (bottom).

of the faces of the cube. The noise corresponds to independent Gaussian noise added to the relative time of arrival data. The average amplitude of the noise is specified relative to the time it takes sound to travel 1m (i.e. $\frac{1}{340}s$). The error corresponds to the mean error after alignment between the localization result and the ground truth. In Fig. 1 we can see that with minimum configurations usable results are obtained for noise levels below 10^{-4} while with more data very good results can already be obtained for a noise level of 10^{-3}. Fig. 2 shows the effect of varying the number of sources and microphones on the accuracy of the results.

4. CONCLUSION

In this paper we have presented what is to our knowledge the first approach for joint source and sensor localization estimation which does not require any initialization. This has significant advantages for applications where source and sensor locations are not approximately known a priori and can enable opportunistic calibration of sensor arrays. We have shown through synthetic experiment that the approach provides reasonable results and is able to successfully initialize non-linear least-squares optimization. Our future work will consist of experiments with real data and exploring applications.

5. REFERENCES

[1] M. S. Brandstein, J. E. Adcock, and H.F. Silverman, "A closed form location estimator for use with room environment microphone arrays," *IEEE Trans. on Speech and Audio Processing*, 5:45–50, 1997.

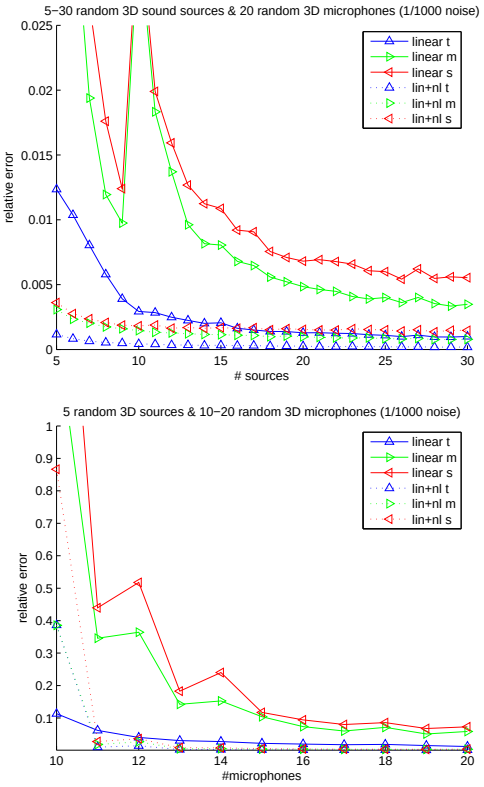

Fig. 2. Experiments with varying numbers of microphones and sources.

[2] J.O. Smith and J.S. Abel, "Closed-form least-squares sources location estimation from range-difference measurements," *Proc. ICASSP*, 1987, 35:1661–1669.

[3] J.M. Sachar, H.F. Silverman, and W.R. Patterson III, "Position calibration of large-aperture microphone arrays," *Proc. ICASSP*, 2002, 2:1797–1800.

[4] V.C. Raykar and R. Duraiswami, "Automatic position calibration of multiple microphones," *Proc. ICASSP*, 2004.

[5] R. Moses, D. Krishnamurthy, and R. Patterson, "A self-localization method for wireless sensor networks," *EURASIP J. on Applied Signal Processing*, 4:348–358, 2003.

[6] A. Weiss and B. Friedlander, "Array shape calibration using sources in unknown locations-a maximum-likelihood approach," *IEEE Trans. on ASSP*, 37:1958–1966, 1989.

[7] Y. Rockah and P. Schultheiss, "Array shape calibration using sources in unknown locations – part i: Far-field sources," *IEEE Trans. on ASSP*, 1987.

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