Proceedings of the 4th International Symposium on Communications, Control and Signal Processing, ISCCSP 2010, Limassol, Cyprus, 3-5 March 2010

Self-calibration of two microphone arrays from volumetric acoustic maps in non-reverberant rooms

S.D.Valente, F.Antonacci, M.Tagliasacchi, A.Sarti, S.Tubaro

Abstract— **In this paper we present a methodology for the self-calibration of two microphone arrays based on the localization of acoustic sources from volumetric acoustic maps, one for each array. A set of correspondences are obtained moving the acoustic source at different locations in space. The proposed algorithm estimates the rigid motion that brings the coordinate system of the second microphone array to the first one through the solution of a least squares problem. The approach presented here enables the self-calibration even when the acoustic sources are in the near-field of the microphone arrays, thus extending the methodology presented by the authors in another work.**

I. Introduction

Microphone arrays enable the acquisition of the space-time structure of an acoustic field. Thus, they have been widely used to solve many tasks in computational auditory scene analysis, ranging from blind source separation to de-reverberation, localization and tracking. In some cases, e.g. in acoustic source localization and tracking, the location and pose of the arrays with respect to the environment needs to be available or it needs to be somehow estimated.

In many scenarios, the analysis of the auditory scene can potentially take advantage of multiple microphone arrays distributed across the environment. Ideally, all the microphones of the various arrays might be thought of as composing a single array, whereby all signals are synchronous with respect to a centralized clock. Unfortunately, this scenario is unrealistic with the current technology. Professional equipment able to acquire simultaneously more than 8-16 channels can be costly. Thus, in a resource constrained scenario, the alternative of deploying distinct, asynchronous microphone arrays, each governed by its own acquisition device, represents a more viable option. Nevertheless, there is the need for devising a procedure for determining the location and the pose of each array with respect to a selected coordinate reference system.

Microphone array calibration has been studied in the recent literature. In [1], [2], [3] the authors address the problem of retrieving the location of microphones within the same array. In [4] the authors address the problem using Multi-Dimensional Scaling (MDS). The formulation used accounts for non-ideal synchronization among the devices.

The authors are with Dipartimento di Elettronica ed Informazione Politecnico di Milano p.zza Leonardo da Vinci, 32, 20133 Milano valente@elet.polimi.it

The work presented in this paper is realized under the funding of the SCENIC project. The SCENIC project acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 226007

978-1-4244-6287-2/10/$26.00 ©2010 IEEE

On the other hand, the problem of inter-array calibration is rather unexplored. In [5] we presented a solution, partially inspired by [6], that addresses the problem of self-calibration using probing signals produced by a loudspeaker moved at different unknown locations. Acoustic images, which represent the energy measured along different directions, are measured for each array using a Delay and Sum Beamformer [7]. Under the hypothesis that a single source is active at each time instant, the three-dimensional Directions Of Arrival obtained by the pair of acoustic images refer to the same acoustic source. Inter-array calibration is finally performed exploiting computer vision tools by analyzing the set of correspondences extracted from the acoustic images.

In [8] the authors address a similar problem in a two stages approach: first each array is internally calibrated using the methodology originally presented in [9] and then an algorithm that is related to [5] addresses the inter-array calibration.

The algorithm in [5] proves to be efficient when sources are in the far-field of the microphone arrays. However, when far-field condition does not apply, a distortion on the acoustic images appears, thus impairing a successful self-calibration. In this paper we present an alternative solution based on the acquisition of volumetric acoustic images using Steered Response Power [10], which describes the distribution of acoustic energy in space. Source localization is performed by estimating the location of the global maxima of the volumetric acoustic maps. Each map is referred to a local coordinate system centered in the reference microphone of the array. In order to infer the mutual positions of the array, we have to estimate the rigid motion (i.e. a rotation matrix and a translation vector) that brings the local coordinate system of the second array to the coordinate system of the first one. We formulate the problem using least squares, which allows us to find the mutual location and pose using a non-iterative algorithm.

The rest of the paper is structured as follows: in Section II we present the acquisition of volumetric acoustic maps using Steered Response Power. Section III formulates the self-calibration problem and addresses it using the rigid motion estimation. Section IV describes some experimental results that compare the far-field methodology in [5] with the present algorithm. Finally, Section V draws some conclusions.

II. Acquisition of volumetric acoustic maps

In this Section we present the Steered Response Power (SRP) [10], the algorithm used to localize acoustic sources in space. The idea behind localization algorithms based on

acoustic maps (see [7], [10], [11]) is to make an hypothesis $\boldsymbol{X}$ on the source location and then verify the coherence of the data with the hypothetical location using a suitable coherence measure. In [5] we formulated the problem of self-calibration using projective acoustic maps extracted using Delay And Sum Beamformer [7], which retrieves the position of the source in terms of direction of arrival rather than source position. In order to obtain a description of acoustic images in terms of location in space rather than Direction Of Arrival, in this paper we make use of Steered Response Power. In the next few lines we present the data model, followed by a description of the localization algorithm.

A. Data model

Let us consider the presence of an acoustic source located at $\boldsymbol{X}_P : [x_{1_P}, x_{2_P}, x_{3_P}]^T$ and of a synchronized cluster of microphones, located at $\boldsymbol{X}_m = [x_{1_m}, x_{2_m}, x_{3_m}]^T, m = 1, \ldots, M$. Figure 1 shows the geometry of the setup.

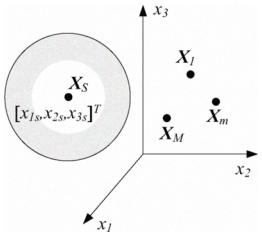


Fig. 1. Geometric notation: receivers are at $\boldsymbol{X}_m = [x_{1_m}, x_{2_m}, x_{3_m}]^T, m = 1, \ldots, M$ and the source is at $\boldsymbol{X}_S = [x_{1_S}, x_{2_S}, x_{3_S}]^T$.

Microphone signals are organized in the column vector

$$\mathbf{x}(t, \boldsymbol{X}_S) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_M(t) \end{bmatrix}, \quad (1)$$

where

$$x_m(t) = s\left(t - \tau_{0m}(\boldsymbol{X}_S)\right) + \nu_m(t) \quad m = 1, ..., M \quad (2)$$

is the signal acquired by the m-th sensor; $s(t)$ is the source signal at the reference microphone and $\nu_m(t)$ is an additive noise. The delay $\tau_{mp}(\boldsymbol{X}_S)$ accounts for the Time Difference of Arrival between sensors m and p and is given by

$$\tau_{mp}(\boldsymbol{X}_S) = \frac{1}{c}\left(d_{\boldsymbol{X}_S, \boldsymbol{X}_m} - d_{\boldsymbol{X}_S, \boldsymbol{X}_p}\right) . \quad (3)$$

The term $d_{\boldsymbol{X}_S, \boldsymbol{X}_m}$ is the distance between the source and the m-th microphone in the array and c is the sound speed. Notice that in eq.(2) we are assuming that reverberations are not present.

After time sampling, the signal acquired by sensors is

$$x_m(k) = x_m(kT), \quad m = 1, ..., M , \quad (4)$$

where T is the sampling period. As usual with microphone arrays, we work with frames composed by W samples. The time-discrete TDOA corresponding to τ_{mp} is represented, by the symbol i_{mp}.

B. Steered Response Power

We compute the cross-correlation $R_{mp}(k)$ between $x_m(k)$ and $x_p(k)$. In the absence of noise, $R_{mp}(k)$ exhibits a peak at the lag $k = i_{mp}$.

Notice that, if the receivers location are given, the time-discrete TDOA i_{mp} depends only on the source position.

The Steered Response Power function for the hypothesis source location $\boldsymbol{X}'$ is

$$SRP(\boldsymbol{X}') = \sum_{m=1}^{M-1} \sum_{p=m+1}^{M} R_{mp}(i_{mp}(\boldsymbol{X}')) \, , m \neq p \, , \quad (5)$$

where $i_{mp}(\boldsymbol{X}')$ is the Time Difference of Arrival between microphones m and p for the position $\boldsymbol{X}'$. If $\boldsymbol{X}'$ is actually the source location, peaks sum coherently in $SRP(\boldsymbol{X}')$ and a global maximum appears.

In order to localize sources, the 3D space is sampled with a regular grid. More specifically, a volume of points $\boldsymbol{X}_{i,j,k} = [x_{1i}, x_{2j}, x_{3k}]^T, i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K$ is defined.

The source location is estimated as the global maximum of $SRP(\boldsymbol{X}_{i,j,k})$:

$$\hat{\boldsymbol{X}}_S = \arg \max_{\boldsymbol{X}_{i,j,k}} SRP(\boldsymbol{X}_{i,j,k}) \, . \quad (6)$$

When reverberations are present, multiple peaks, related to image sources, are superimposed in the acoustic map. In order to address reverberations, therefore, the design of the array should ensure sufficient resolution, so that peaks do not overlap and it is possible to distinguish the actual source from the image ones.

Notice also that the full-search on the volume may be computationally demanding. As an example, if we are localizing a source on a volume of 1 m^3 with a resolution of 0.001 m, the volume turns out to be composed of 10^9 points. In order to address the computational cost problem, we adopt a multi-scale localization: we first sample SRP on a coarse volume and then we refine the search in proximity of the global maximum of the coarse SRP.

III. Self-calibration from volumetric acoustic maps

Let us consider the presence of two cross-shaped microphone arrays, as depicted in Figure 2. The arrays are internally calibrated, i.e. the position of each microphone in the array is known. For each array a reference microphone is defined, and the local coordinate system is referred to that position, as depicted in Figure 2, where the reference microphone is the central one. The arrays are not mutually calibrated, i.e the second array is not informed of the pose of the first one and viceversa. Furthermore, we assume that the two arrays are internally synchronized but mutually asynchronous, i.e. the clock is not shared. The source is moved at L different positions in space. Each array provides localizes it at $\hat{\boldsymbol{X}}_l$ and $\hat{\boldsymbol{X}}'_l$, $l = 1, \ldots, L$ referred to the local coordinate systems. In order to write a relationship between

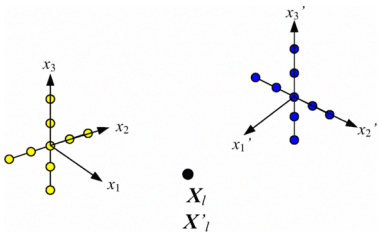


Fig. 2. An acoustic source is located in space and it is localized by the pair of arrays. The local coordinate systems are centered in the reference microphones of each array. The estimations of the source location in the local coordinate systems are $\boldsymbol{X}_l$ and $\boldsymbol{X}'_l$.

$\boldsymbol{X}_s$ and $\boldsymbol{X}'_s$ in a compact way, we convert $\boldsymbol{X}_l$ and $\boldsymbol{X}'_l$ into the homogeneous representations, given by

$$\mathbf{X}_l = [\boldsymbol{X}_l^T, 1]^T \quad \mathbf{X}'_l = [\boldsymbol{X}_l'^T, 1]^T \; . \tag{7}$$

With this notation at hand, under the assumption that $\mathbf{X}_l$ and $\mathbf{X}'_l$ are correctly localized, the homography that relates $\mathbf{X}_l$ and $\mathbf{X}'_l$ is:

$$\mathbf{X}'_l = \mathbf{H}\mathbf{X}'_l \; , \tag{8}$$

where the matrix $\mathbf{H}$ accounts for the roto-translation between the local coordinate systems and it has the following internal structure

$$\mathbf{H} = \begin{bmatrix} \mathbf{R}(\theta,\phi) & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \; , \tag{9}$$

where $\mathbf{R}(\theta,\phi)$ is the three-dimensional rotation matrix and $\mathbf{t}$ is the translation vector that bring the coordinate system of the first array to the coordinate system of the second one. Our goal is to estimate this rigid motion given a set of corresponding points $\mathbf{X}_l \leftrightarrow \mathbf{X}'_l \;, l = 1, \ldots, L$. This problem is common in the literature of computer vision and it is known as *calibrated reconstruction* [12]. We observe that the matrix $\mathbf{H}$ in (9) has 5 d.o.f. as it is univocally determined when the angles θ and ϕ and the components of the translation vector are given.

Replacing (9) into (8) and considering the presence of an additive measurement error $\mathbf{V}_l$ we get

$$\mathbf{X}'_l = \mathbf{R}(\theta,\phi)\mathbf{X}_l + \mathbf{t} + \mathbf{V}_l \; ,$$

We proceed to the estimation of $\mathbf{R}(\theta,\phi)$ and $\mathbf{t}$ using a least squares approach. We define the cost function as

$$\Sigma^2 = \sum_{l=1}^{L} ||\mathbf{X}'_l - \hat{\mathbf{R}}\mathbf{X}_l - \mathbf{t}||^2 \; . \tag{10}$$

In order to find the rotation matrix and the translation vector that minimize the cost function in eq.(10), we define some auxiliary variables

- the centroids of the sets of corresponding points $\mathbf{X}_l$ and $\mathbf{X}'_l$ are

$$\overline{\mathbf{X}} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{X}_l \quad , \overline{\mathbf{X}}' = \frac{1}{L}\sum_{l=1}^{L}\mathbf{X}'_l$$

- the average is subtracted to the corresponding points to give $\mathbf{X}_{cl} = \mathbf{X}_l - \overline{\mathbf{X}} \leftrightarrow \mathbf{X}'_{cl} - \overline{\mathbf{X}}'$
- the correlation matrix for the corresponding points $\mathbf{X}_{cl} \leftrightarrow \mathbf{X}'_{cl}$ is

$$\mathbf{C} = \sum_{l=1}^{L} \mathbf{X}'_{cl}\mathbf{X}_{cl}^T \; ;$$

- the singular value decomposition of the matrix $\mathbf{C}$ is given by $\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$.

With these definitions at hand, the estimations of the rotation matrix and translation vector are [13]

$$\hat{\mathbf{R}}(\theta,\phi) = \mathbf{V}\mathbf{U}^T \; ,$$

and

$$\hat{\mathbf{t}} = \overline{\mathbf{X}}' - \hat{\mathbf{R}}(\theta,\phi)\overline{\mathbf{X}} \; .$$

Rigid motion estimation turns out to be very sensitive to the presence of outliers in the measurement set. In order to remove them, we have implemented the well-known RANSAC algorithm [14], [12], which finds a robust set of inliers among the correspondences $\mathbf{X}_l \leftrightarrow \mathbf{X}'_l, l = 1, \ldots, L$, which fits the rigid motion model.

IV. Experiments

The accuracy of the rigid motion estimation is tested by comparing the estimated rotation matrix and translation vector with the actual ones, using the metrics defined in [15]:

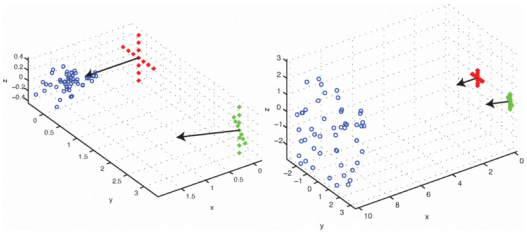
$$\epsilon_R = \arccos\left(\frac{\mathrm{tr}(\hat{\mathbf{R}}(\theta,\phi)^T\mathbf{R}(\theta,\phi)) - 1}{2}\right), \tag{11}$$

$$\epsilon_t = \arccos(\hat{\mathbf{t}}^T\mathbf{t}) \; . \tag{12}$$

In (12), translation vectors are supposed to be scaled in order to have unitary norm.

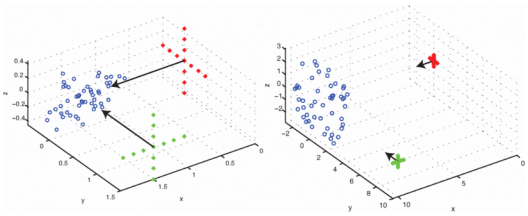
The first simulation setup is shown in Figure 3: two cross-shaped arrays (each consisting of 13 microphones). The second array is displaced by rotating the first one by $-\pi/8$ around the vertical axis and translated by 3m. Circles denote the positions of the sources for each repetition of the experiment. In Figure 3(a) sources are located in the near-field, while in Figure 3(b) they are located in the far-field. In particular, a source is considered to be in the near-field when its distance to the microphone array is smaller than the Fraunhofer distance. For each simulation, a variable number of sources (ranging from 8 to 50) have been considered. Each simulation has been repeated ten times in order to average over several realizations. The same setup has been repeated for the second experiment, however this time the second camera is rotated by $-\pi/2$, as shown in Figure 4.

Figure 5(a) shows the calibration error as a function of the number of correspondences for the nearfield configuration. The results for the methodology presented in this paper and in [5] are denoted with the subscripts "DSB" and "SRP", respectively. We notice that the 3D rigid motion estimation enables a more accurate self-calibration with respect to [5]. The same experiment has been repeated in the far-field



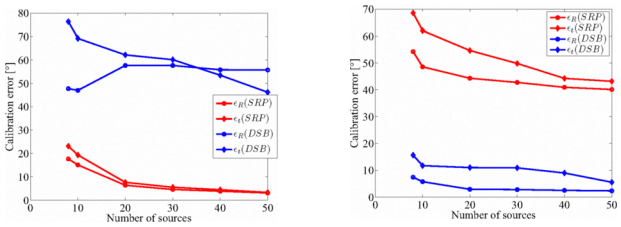
(a) (b)

Fig. 3. Geometry of the first simulation setup: the second camera location is obtained by rotation of $-\pi/8$ around the vertical axis and translated by 3m. Circles denote the positions of the sources for each repetition of the experiment. In (a) sources are located in the near-field, while in (b) are located in the far-field.



(a) (b)

Fig. 4. Geometry of the second simulation setup: the second camera location is obtained by rotation of $-\pi/2$ around the vertical axis and translated by 3m.



(a) (b)

Fig. 5. Calibration error for nearfield (a) and farfield (b) sources as a function of the number of correspondences for the configuration in Figure 3. The results for the methodology presented in this paper and in [5] are denoted with the subscripts "DSB" and "SRP", respectively.

configuration depicted in Figure 3(b). Results are shown in Figure 5(b). We notice that in this situation the accuracy of self-calibration based on projective and volumetric images is inverted: Delay and Sum Beamformer behaves better than SRP.

Figure 6 shows the self-calibration results for the configuration in Figure 4(a). The results confirm that self-calibration based on volumetric acoustic images behaves better than the methodology in [5]. Similar considerations to Figure 5(b) apply when we move the sources in the far-field, as in Figure 4(b).

V. Conclusions

In this paper we presented a novel algorithm for the self-calibration of microphone arrays in dry enclosures using volumetric acoustic maps. Simulation results demonstrate that the algorithm can efficiently estimate in the near-field the mutual position of the arrays even using a limited number of corresponding points, thus filling the gap of the methodology in [5].

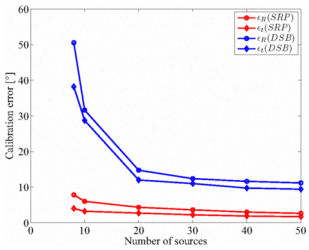


Fig. 6. Comparison between self-calibration based on projective and volumetric acoustic images for the geometry depicted in Figure 4(a)

References

[1] A. J. Weiss and B. Friedlander, "Array shape calibration using sources in unknown locations-a maximum likelihood approach," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'88)*, Apr. 1988, pp. 70 – 73 vol.1.

[2] D. K. R. Moses and R. Patterson, "A self-localization method for wireless sensor networks," in *Eurasip J. Appl. Signal Process. Special Issue on Sensor Networks*, Mar. 2003, p. 348358.

[3] V. C. Raykar and R. Duraiswami, "Automatic position calibration of multiple microphones," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, May 2004, pp. iv–69 – iv–72 vol.4.

[4] I. K. V. Raykar and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE transactions on speech and audio processing*, vol. 13, pp. 70–83, Jan. 2005.

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

[6] R. A. O'Donovan and J. Neumann, "Microphone arrays as generalized cameras for integrated audio visual processing," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, Jun. 2007, pp. 1–8.

[7] H. Trees, *Optimum Array Processing: Detection, Estimation and Modulation Theory*. Wiley, 2002, vol. IV.

[8] M. Hennecke, T. Plotz, G. Fink, J. Schmalenstroer, and R. Hab-Umbach, "A hierarchical approach to unsupervised shape calibration of microphone array networks," in *IEEE/SP 15th Workshop on Statistical Signal Processing, 2009, SSP '09*, Sept. 2009, pp. 257–260.

[9] I. McCowan, M. Lincoln, and I. Himawan, "Microphone array shape calibration in diffuse noise fields," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 666–670, Mar. 2008.

[10] J. Dmochowski, J. Benesty., and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2510–2526, Nov. 2007.

[11] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 288–292, 1993.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in computer vision*, 2nd ed. Cambridge Univ. Press, 2003.

[13] D. Eggert, A. Lorusso, and R. Fisher, "Estimating 3-d rigid body transformations: a comparison of four major algorithms," *Machine Vision and Applications*, vol. 9, no. 5-6, pp. 272–290, 1997.

[14] 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*, vol. 24, pp. 381–395, June 1981.

[15] G. Chesi, "Camera displacement via constrained minimization of the algebraic error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 370–375, Feb. 2009.