Geometric calibration of distributed microphone arrays

Alessandro Redondi, Marco Tagliasacchi, Fabio Antonacci, Augusto Sarti

Dipartimento di Elettronica e Informazione, Politecnico di Milano
P.zza Leonardo da Vinci, 32 - Milano, Italy
{redondi, tagliasa, antonacc, sarti}@elet.polimi.it

***Abstract*—Computational auditory scene analysis exploits signals acquired by means of microphone arrays. In some circumstances, more than one array is deployed in the same environment. In order to effectively fuse the information gathered by each array, the relative location and pose of the arrays needs to be obtained solving a problem of geometric inter-array calibration. We consider the case where the arrays do not share a synchronous clock, which impairs the use of time-difference of arrival measures across arrays. Conversely, each array produces an acoustic image, which describes the energy of acoustic signals received from different directions. We jointly consider acoustic images acquired by the different arrays and adapt computer vision techniques to solve the calibration problem, thus estimating the location and pose of microphone arrays sensing the same auditory scene. We evaluate the robustness of the calibration process in a simulated environment and we investigate the effect of the various system parameters, namely the number of probing signal locations, the resolution of the acoustic images, the non-ideal intra-array calibration.**

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. This requirement also holds for those techniques that attempt to exploit the knowledge of the geometry of the environment, as it has being explored in the ongoing EU-funded SCENIC project[1].

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, since commercially available hardware platforms limit the number of channels that can be synchronously acquired by the same device. Professional equipment able to acquire simultaneously more than 8-16 channels can be extremely 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 simple and accurate procedures for calibrating the geometry of the arrays, i.e. determining the location and pose of each array with respect to a selected coordinate reference system.

Microphone array calibration has been studied in the recent literature. Most of the works aim at solving the problem of intra-array calibration, i.e. determining the positions of each microphone within the array. This problem is formulated in a maximum likelihood framework in several papers [1][2][3]. The solution turns out to be the minimum of a nonlinear cost function. The work in [4] proposes a technique based on multi-dimensional scaling (MDS), which requires as input tape measures of the distances between the microphones. In [5] the authors consider a system setup with distributed laptop devices, each with a loudspeaker and a microphone, which are used to determine the relative positioning. The algorithm also accounts for non-ideal synchronization among the devices.

On the other hand, the problem of inter-array calibration is rather unexplored. Our work has been partially inspired to [6], where the authors consider combined audio and video processing. They propose a calibration process to correctly superimpose an acoustic map, obtained by means of a spherical microphone array, onto a visual image captured by a conventional camera. Nevertheless no details about the accuracy of the system is provided. The same system is used in [7] to study the acoustic response of an auditorium.

In this paper we consider the problem of calibrating the location and pose of microphone arrays by means of acoustic probing signals emitted by a loudspeaker moved at different locations, thus avoiding the cumbersome and inaccurate task of manual calibration with rulers. We focus on a scenario with two microphone arrays, where one of them serves as reference and we seek for the positioning of the second with respect to the reference one. Each microphone array works as an acoustic camera, producing a set of acoustic images, each representing the energy measured along different directions. At each time instant, a probing signal is emitted by a loudspeaker in a given location. The analysis of the acoustic images enables to obtain correspondences between the images. It is possible to gather a set of correspondences by moving the loudspeaker, thus enabling inter-array calibration. While in this paper we re-use conventional tools from computer vision, we notice that the characteristics of the acoustic images are rather different from those of conventional optical images. We can observe the following facts: 1) it is possible to extract a large number of correspondences from a pair of optical images, but a potentially significant fraction of them might represent outliers. In our setting, the number of correspondences is smaller, since it relates to the number of probing signal locations, but more reliable, as we do not have the problem of mismatches; 2) the resolution of acoustic images is typically very low, and it is determined by the array geometry; 3) intra-calibration of a microphone array might be inaccurate. Starting from these observations, in this paper we study the robustness of the calibration algorithm with respect to several factors that might affect the result: the number of acoustic

[1]The project SCENIC 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

MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.
978-1-4244-4464-9/09/$25.00

probes; the resolution of the acoustic images; the non-ideal intra-array calibration.

II. Background

An acoustic camera can be interpreted as an array of microphones properly combined with a beamforming algorithm that enables the generation of an acoustic image. In fact, under certain assumptions, a similar device behaves like a central projective camera, enabling the visualization of the acoustic response of a scene along different directions. In this section we introduce the basics of beamforming, the camera model and the relative geometry between two cameras.

A. Delay and Sum Beamforming (DSB)

With reference to Figure 1, we consider a sound source located in position $\mathbf{X} = [X, Y, Z]^T$ corresponding to the direction (θ_s, ϕ_s), emitting a monochromatic plane wave. The plane wave assumption is verified if we observe the sound source in its far-field. The sound field is acquired by an array of microphones of M elements, located in positions $\mathbf{p}_m : m = 0, 1..., M-1$. If the sensors are omnidirectional and with equal gains, the signal arriving at the m-th microphone is:

$$s_m(n) = s(n - \tau_m(\theta_s, \phi_s)) \tag{1}$$

where $s(n)$ is the signal arriving at the origin of the coordinate system (usually coincident with one of the elements of the array, called the reference microphone) and $\tau_m(\theta_s, \phi_s)$ is the time delay (in samples) from the origin to the m-th microphone. Under the far-field assumption, the delays corresponding to a generic direction of arrival (DOA) (θ, ϕ) can be calculated as

$$\tau_m(\theta, \phi) = \frac{\mathbf{a^T p}_m}{c} f_s \tag{2}$$

where c is the sound propagation speed, f_s is sampling frequency and $\mathbf{a}$ is a unit vector with direction (θ, ϕ) and pointing towards the reference microphone, which can be expressed as

$$\mathbf{a} = \begin{bmatrix} -\sin\theta \sin\phi \\ -\cos\theta \\ -\sin\theta \cos\phi \end{bmatrix} \tag{3}$$

The goal of beamforming [8] is to steer a beam in a specific direction of arrival (θ, ϕ) and to measure the corresponding acoustic signal. This steering can be achieved by compensating the delay of the m-th microphone with respect to the reference one and then summing up all contributions for $m = 1 \cdots M$:

$$y(n; \theta, \phi) = \sum_{m=0}^{M-1} x_m(n + \tau_m(\theta, \phi)) \tag{4}$$

B. Camera models

A camera is a mapping between the 3D world and a 2D image. With reference to Figure 1, let the center of projection $\mathbf{C}$ be the origin of a Euclidean coordinate system, and consider the plane $Z = f$, which is called the *image plane*. Under the general projective camera model, a point in space $\mathbf{X} = [X, Y, Z]^T$ is mapped to the point $\mathbf{x} = [fX/Z, fY/Z, f]^T$ on the image plane.

If we represent world and image points with homogeneous coordinates, such a mapping may be easily written in matrix form as

$$\mathbf{x} = \mathbf{PX}, \tag{5}$$

where $\mathbf{P}$ is a 3x4 matrix defined as

$$\mathbf{P} = \begin{bmatrix} f & & & 0 \\ & f & & 0 \\ & & 1 & 0 \end{bmatrix}. \tag{6}$$

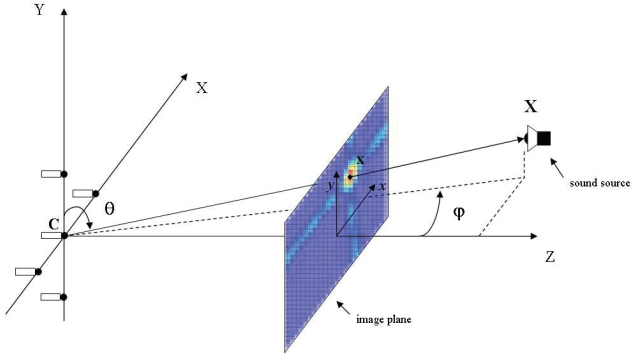


Fig. 1. A cross shaped microphones array on the XY plane is used as an acoustic camera of center $\mathbf{C}$. The sound source located in $\mathbf{X}$ is projected onto the image plane in the point $\mathbf{x}$.

In general, a camera is not necessarily centered in the world reference system. The location and pose of a camera can be expressed in terms of, respectively, the coordinate of the camera center $\widetilde{\mathbf{C}}$ and a rotation matrix $\mathbf{R}$. In this case the projective mapping of the point $\mathbf{X}$ on the image plane is defined by

$$\mathbf{x} = \mathbf{PX} = \mathbf{KR}[\mathbf{I}| - \widetilde{\mathbf{C}}]\mathbf{X} = \mathbf{K}[\mathbf{R}|\mathbf{t}]\mathbf{X}, \tag{7}$$

where $\mathbf{t} = -\mathbf{R}\widetilde{\mathbf{C}}$ and $\mathbf{K}$ is a 3×3 matrix:

$$\mathbf{K} = \begin{bmatrix} f & & p_x \\ & f & p_y \\ & & 1 \end{bmatrix}. \tag{8}$$

The matrix $\mathbf{K}$ is called *calibration matrix* and takes into account the internal camera parameters (focal length f and origin of image coordinates (p_x, p_y)). In Section IV we summarize a calibration method [9] that can be adopted to estimate the pose (related to $\mathbf{R}$) and location (related to $\mathbf{t}$), when the internal parameters are known.

C. Epipolar Geometry

The epipolar geometry represents the intrinsic relationship between two projective cameras and depends only on the parameters of the cameras. Suppose to have two cameras, imaging the same point object $\mathbf{X}$ from different locations. This results in an image point $\mathbf{x}$ in the first camera and $\mathbf{x}'$ in the second camera. The epipolar geometry shared by the two cameras is algebraically represented by a 3x3 matrix $\mathbf{F}$ called the fundamental matrix, such that for any pair of corresponding points $\mathbf{x} \leftrightarrow \mathbf{x}'$ in the two images this condition is satisfied:

$$\mathbf{x'^T F x} = 0 \tag{9}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are the 2D image points expressed in homogeneous coordinates, i.e $\mathbf{x} = [x, y, 1]^T$. An estimate of the fundamental matrix $\mathbf{F}$, needed to achieve the calibration, can be obtained from a number of corresponding image points.

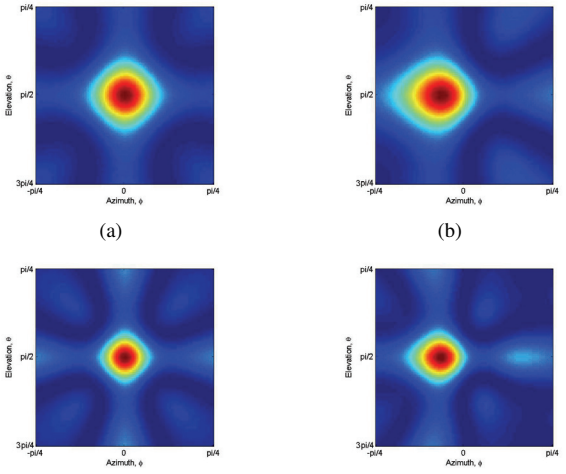
III. Acoustic Camera

The previous sections introduced all the elements needed to define an acoustic camera. Given a microphone array of arbitrary geometry (e.g. planar) placed in the 3D space, we select a point $\mathbf{C}$ that will be used as the reference for beamformer delays as well as the acoustic camera center. For example, in our system setup, we consider a cross-shaped array composed of M elements, with $\mathbf{C}$ coincident with the central microphone. We position the image plane at distance f from $\mathbf{C}$. In order to generate an acoustic image, one has to decide the

image resolution, i.e. the number of pixels per unit length along both the horizontal and vertical directions. Once the resolution is fixed, the image can be generated: the pixel having coordinates (i, j) in the image plane has a position in 3D space that corresponds, in spherical coordinates, to a pair of angles $(\theta_{i,j}, \phi_{i,j})$. Using these angles in (3) enables to steer the beamformer. To obtain the intensity value for the pixel (i, j) we consider the acoustic energy, which is defined as

$$E(i,j) = \frac{1}{L}\sum_{n=0}^{L-1} |y(n; \theta_{i,j}, \phi_{i,j})|^2, \tag{10}$$

where $y(n; \theta_{i,j}, \phi_{i,j})$ is the beamformer output and L is the length of the temporal acquisition frame. The acoustic camera internal parameters are calibrated by definition. In fact, once f is chosen, one can build a matrix K as in equation (8), with the proper principal point offset. As an example, Figure 2(a) shows the acoustic image of a source located in $[0, 0, 3]^T$, which corresponds to the direction $(\phi_s, \theta_s) = (0, \pi/2)$, generated by a cross-shaped microphone array centered in the origin with $M = 9$ sensors spaced 15cm apart. Figure 2(b) shows the same source imaged with a similar array centered in $[-1, 0, 0]^T$ and rotated by $\pi/8$ around the y axis. Figure 2(c) and 2(d) show the same source acquired by increasing the number of microphones of each array to $M = 13$.



(a) (b) (c) (d)

In order to perform calibration, we need to extract the coordinates of the source on the image plane. We notice in Figure III that the x image of a point source spreads to a large number of pixels. The characteristics of such point spread function ultimately depend on the microphone array aperture, which is constrained by the number of available sensors M and by the frequency of the probing signal ω. In fact, the value of ω imposes the maximum inter-element distance d to prevent spatial aliasing, i.e. $d \leq \pi c/\omega$. For cross shaped arrays, the aperture length is equal to $d(M-1)/2$.

IV. Calibrated Reconstruction

Given two microphone arrays in arbitrary positions, we want to determine the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$ that describe the pose and location of the second microphone array with respect to the first one. The inputs of the calibration algorithm consist of N acoustic image pairs, each pair corresponding to a different position of the probing source, and two calibration matrices $\mathbf{K}$ and $\mathbf{K}'$ relative to the two acoustic cameras. This problem is known

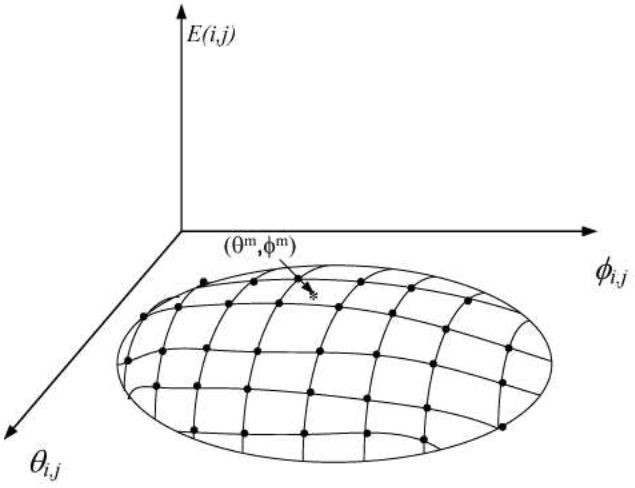


Fig. 2. Estimation of the maximum of the paraboloid from the knowledge of the corrupted samples $E(i, j)$.

in computer vision as *calibrated reconstruction*. Below we briefly summarize the steps, emphasizing the differences with respect to the case of optical images.

A. Estimating point correspondences

The first step consists of obtaining N points correspondences $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$, where $\mathbf{x}_i$ ($\mathbf{x}'_i$) denotes the coordinates of the i-th probing signal on the image plane of the first (second) acoustic camera. A simple strategy, consisting of peaking the locations of the maxima in the two acoustic images does not work. In fact, such locations are affected by measurement errors, due to finite image resolution, acquisition noise and non-ideal point spread function. These errors can severely affect the estimation of the epipolar geometry, also when robust algorithms are used to estimate the fundamental matrix [9]. Therefore, in order to obtain robust estimates of the point correspondences, we perform a parabolic fitting of the pixel values around the global maxima by means of least squares and we select the coordinates of the maxima of such paraboloids, as illustrated by Figure 2.

B. Estimating epipolar geometry

Given N reliable correspondences, the fundamental matrix $\mathbf{F}$ can be estimated when $N \geq 8$. To this end, we use the normalized 8-points algorithm with non linear minimization following [9]. Moreover, since our acoustic cameras are internally calibrated, we can estimate a lighter version of the fundamental matrix, i.e. the essential matrix $\mathbf{E}$, which is related to $\mathbf{F}$ by the following equation:

$$\mathbf{E} = \mathbf{K'^T F K}, \tag{11}$$

We notice that in our setting $\mathbf{K}$ and $\mathbf{K}'$ are known exactly, thus they are not affected by errors. The essential matrix can be decomposed in a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$, corresponding to the second camera, with the reference system centered in the first camera. In this way, calibration is obtained up to a scale ambiguity on the translation vector, due to the intrinsic homogeneous nature of the projective geometry. This ambiguity can be solved in practice exploiting the fact that, unlike optical cameras, each acoustic camera can also provide the ranges of the imaged probing sources. This enables to set up a simple triangulation problem and solve the scale ambiguity on the translation vector.

V. Results

This section illustrates the simulations that we carried out in order to assess the robustness of the calibration. We consider two acoustic cameras, each consisting of a cross-shaped microphone array with $M = 9$ elements, and inter-element distance d = 15 cm. The probing signal is a sinusoid at frequency 1 kHz corrupted with additive Gaussian noise to achieve a SNR equal to 20dB. The sampling rate is 48 kHz. To show the effect of the microphone array aperture on the system performance, we repeated all the simulations using $M = 13$.

We evaluate the calibration error by using the following metrics:

$$\epsilon_{\mathbf{R}} = \arccos\left(\frac{\mathrm{tr}(\mathbf{R}^T\mathbf{R}_{true}) - 1}{2}\right), \tag{12}$$

and

$$\epsilon_{\mathbf{t}} = \arccos\left(\mathbf{t}^T\mathbf{t}_{true}\right), \tag{13}$$

where $\mathbf{R}_{true}$ and $\mathbf{t}_{true}$ are the true external parameters and $\mathbf{R}$ and $\mathbf{t}$ are the estimated ones. It is possible to consider $\epsilon_{\mathbf{R}}$ as the angle in the exponential coordinates representation of the rotation matrix $\mathbf{R}^T\mathbf{R}_{true}$, while $\epsilon_{\mathbf{t}}$ is the angle between the translation vectors $\mathbf{t}$ and $\mathbf{t}_{true}$. Both metrics range from 0 (perfect estimation) to π (worst-case estimation) [10].

Every experiment consists of these steps:

- Two acoustic cameras are positioned in the 3D space. One is selected as the reference, and it defines the center of the world coordinate system. The other is rotated by an angle $\pi/8$ around the y axis and translated to $[-1, 0, 0]^T$.
- $N \geq 8$ probing sources are positioned at random locations in front of the two cameras in such a way that they are visible from both cameras, and only one source is active at each time instant. N acoustic image pairs are computed and point correspondences are determined.
- The essential matrix is estimated and decomposed in a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$. Calibration errors are computed with (12) and (13).
- System parameters are changed according to the specific case study as detailed below. We study the impact of different factors on the calibration error, including: the number of probing sources; the non-ideal intra-array calibration; the acoustic image resolution.

In order to provide reliable results that do not depend on the specific position of the probing source, we repeat the experiment 100 times for each experiment configuration. The results are averaged over all the repetitions of the experiment.

Figure 3 shows the impact of the number of probing sources N. Here we assume that the two microphone arrays are intra-calibrated, thus there is no error in the positioning of the microphones within each array. The field of view of both acoustic cameras is $\pi/2$. We select a resolution of 20×20 pixels, which is equivalent to a maximum scanning angle α of about $\pi/36$ radians (angular sampling is not uniform, but it gets smaller as the beam is steered from the center of the image to the edges). We observe from Figure 3 that a calibration error of less then $\pi/20$ radians is achieved by using at least 20-30 probing source locations. We emphasize that the parabolic fitting described in Section IV is essential to achieve an accurate calibration. In fact, simply picking the maxima of acoustic images to generate correspondences yields a much higher calibration error, e.g. $\epsilon_{\mathbf{R}} \simeq \pi/10 - \pi/4$.

In order to simulate intra-array calibration error, we perturb the positions of the microphones with additive Gaussian noise. Figure 4 shows the resulting calibration error when the standard deviation of the Gaussian noise is changed in the range $[0 - 0.01]$ meters, which is typically achievable with state-of-the-art intra-array calibration methods described in the literature. We use $N = 30$ probing source locations, and the same resolution as in Figure 3.

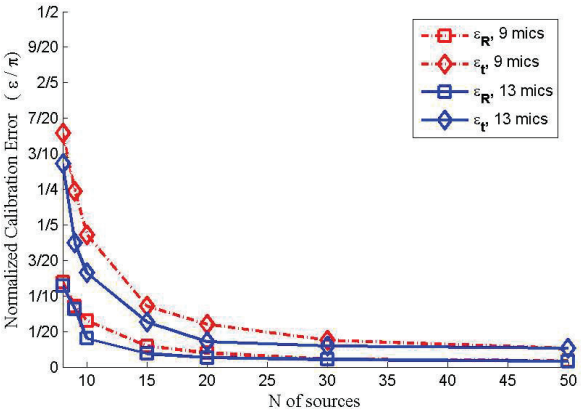


Fig. 3. Impact of the number of probing source locations on the calibration errors

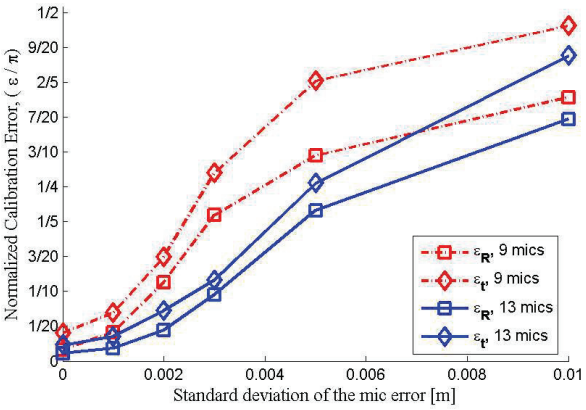


Fig. 4. Inter calibration error with respect to different microphone position error

We observe that accurate intra-array calibration is mandatory in order to achieve good inter-array calibration results. Indeed, an average error of more than few millimeters in microphone positioning can severely affect the calibration process.

Finally, we study the impact of different resolutions on the calibration error, when $N = 30$ and the standard deviation of the intra-array calibration error is set equal to 1mm. We measure the acoustic image resolution in terms of the maximum width of the scan angle. Among all parameters, resolution has the most significant impact on computational resources because it determines the number of pixels in each image, hence the number of beams to be formed. We notice from Figure 5 that good results are obtained even with acoustic images at low resolutions (e.g. 20×20 pixels), enabling the conditions for real-time generation of the acoustic image and for acoustic video.

VI. Conclusion

We have presented an approach to obtain geometric calibration of multiple microphone arrays. Using computer vision techniques applied to acoustic images, the relative location and pose between

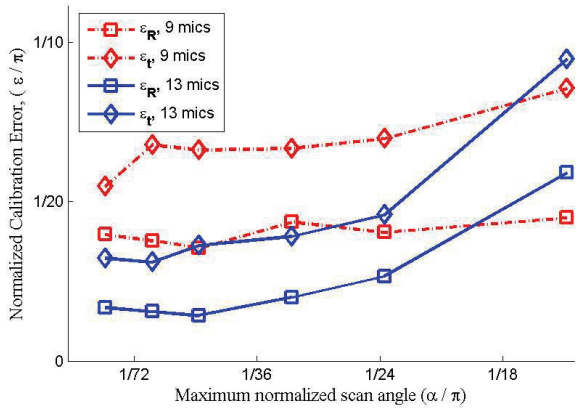


Fig. 5. Calibration error with different image resolutions

distinct and asynchronous microphone arrays can be obtained only with the use of probing sources. Our simulations show the impact of different factors on the robustness of the system. We emphasize that the problem of inter-array calibration can also be solved with methods that do not require the computation of acoustic images. In fact, processing the data acquired by each microphone array enables to obtain information about both the direction and range of the probing sources. Range information can indeed be exploited to solve the calibration problem directly. Comparing such methods with the one presented in this paper is subject of ongoing work.

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] S. Birchfield and A. Subramanya, "Microphone array position calibration by basis-point classical multidimensional scaling," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 1025–1034, Sep. 2005.

[5] I. K. V. Raykar and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 70–83, Jan. 2005.

[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] R. A. O'Donovan and D.Zotkin, "Imaging concert hall acoustics using visual and audio cameras," in *IEEE International Conference on Acoustics, Speech and Signal Processing. (ICASSP'08)*, Apr. 2008, pp. 5284 – 5287.

[8] H. V. Trees, Ed., *Optimum Array Processing*, ser. Detection, Estimation and Modulation Theory. New York, NY: Wiley, 2002, vol. IV.

[9] R.Hartley and A.Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, England: Cambridge University Press, 2003.

[10] G. Chesi, "Camera displacement via constrained minimization of the algebraic error," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, pp. 370–375, Feb. 2009.