Ecole Polytechnique Fédérale de Lausanne

School of Computer and Communication Sciences (IC)

Laboratory of Audiovisual Communication (LCAV)

Semester project:

Calibration algorithms for audio-visual wearable devices

Wissem Allouchi

Supervisors:

Ivan Dokmanic, Juri Ranieri

Professor:

Martin Vetterli

1 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 such as source separation, source localization and tracking.

Systems that combine microphone arrays with a digital camera are getting more and more used for many applications.
For example, these systems are used for generating acoustic images of sound sources on an object, for under water sound imaging as well as for teleconference room application. In the latter case, the microphone array can for example be used to identify the person who is speaking and then order the camera to track it automatically. All the mentioned applications require the coordinates of the microphones, the position and the orientation of the camera and the speed of sound to be known accurately.
If the microphones coordinates and/or camera position and orientation and/or speed of sound are inaccurate then errors are introduced.
The required accuracy of the system depends on its related application but it is proportional to the wavelength of the acoustic signal. In other words, the required accuracy increases as soon as the frequency of the acoustic signals increases.

Many calibration techniques that aim at finding the parameters of the system (microphones positions, camera orientation, speed od sound) have been already developed. Manual methods have been presented where the microphone coordinates may be known with some accuracy if the array structure is built using a milling machine or a laser cutter. The coordinates of the microphones may also be estimated using a faro-arm or a laser scanner.
These manual techniques are in general not only costly and hard to use but they may introduce large errors. This case happens when there is some phase differences between microphones. The phase differences give the same effect as microphone position errors.
Other calibration techniques known as self-calibration techniques have also been developed. These methods rely on using sound sources with unknown positions. Although self-calibration techniques are useful for ad-hoc calibration settings, the errors in the microphones positions are considered to be too large for applications with high frequency acoustic signal such as high frequency beam forming.

Typically distances between sound sources and microphones in the microphone array or inter microphone distances are estimated. Then either a non-linear optimization can be soled for the coordinates or pair-wise distances can be estimated then the traditional multidimensional scaling is often used.
In addition to large errors, techniques using sound sources at unknown position suffer also from the problem that the microphone positions are obtained relative to each other. Hence, they are in an arbitrary reference frame.
In this work, we study calibration technique intended for acoustic imaging microphone arrays (microphone arrays with an attached digital camera) calibration. This technique combines camera calibration with array shape calibration and alignment of the camera with the array.
Alignment of the camera with the microphone array allows for a new range of applications for example [1] presents a calibration and alignment technique to build a

robust system that automatically steers a PTZ camera towards and track sound source localized using the microphone array.

The calibration technique developed during this work is based on the work of [2] it combines camera calibration with the microphone position calibration. The advantage of this method is that the microphones and the camera may be placed in ad-hoc positions and orientations. Additionally, no a priori information about microphone or camera positions is required.

2 System requirements

Before describing the calibration technique, it's useful to estimate the desired accuracy considering the respective microphones misplacement [3] considered different microphone positions offsets. For each range of the misplacement, the quality of the array was measured by computing the difference (percentage) between the actual maximum available map contrast (difference from the main lobe to the first side lobe) and the map contrast after introducing the microphone misplacement.

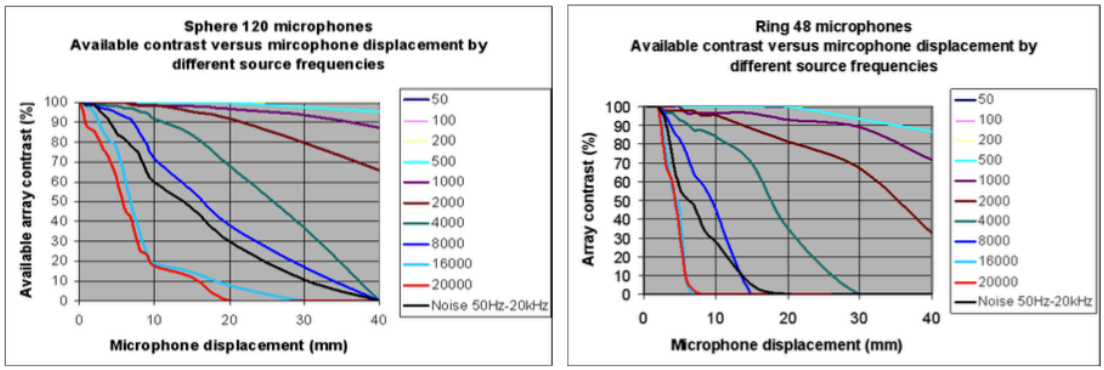

Figure 1: Performance of two different microphone arrays (spherical array to the left and a ring array to the right). The performance was computed by computing the percentage of the availabe map contrast after adding the microphone misplacement.

Figure 1 shows that even a microphone position offset of 2mm results in a reduction of the quality by 20% in the high frequency range.
We expect that the quality reduction is more severe when the microphone array is linear and has fewer elements. That's why we are willing to reach accuracy in the microphone positions less than 2mm.

3 Calibration method

A calibration rig is used. It consists of sound sources surrounding a checkerboard pattern attached to a Plexiglas support. The calibration process goes into 2 phases.

First step is the camera calibration. In this phase we use multiple images of the checkerboard pattern in different positions and orientations. After the camera calibration is done, one can find out the 3D coordinates of sound sources in the camera's coordinates system. The second step is the microphone array calibration. In this phase the 3D sound source positions combined with the time of flight of a tape signal from each sound source to each microphone are used in order to determine the microphone position in the camera coordinates system.

a) Camera Calibration

The camera calibration process aims at finding the internal and external parameters that affect the imaging process. The commonly used technique uses images from different positions of a checkerboard pattern with known dimensions. For each image, the pixels that correspond to the corners of the checkerboard pattern are obtained. The pattern corners combined with the square dimensions are used to obtain the intrinsic parameters of the camera. In addition, for each image, the extrinsic parameters of the camera R and T are obtained which will allow a mapping between the camera coordinates system and the real world coordinates system

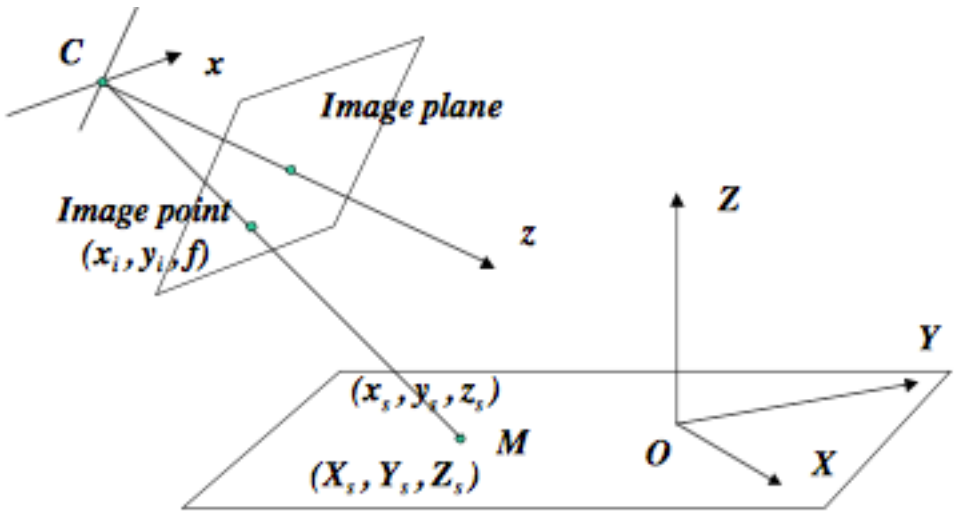

Figure 2 Mapping between real world reference frame and camera reference frames.

The real world coordinates system is right handed, it is usually defined such that one extreme inner corner is the origin, X and Y axes lie in the plane of the checkerboard pattern and such that the Z axes points out of the board.

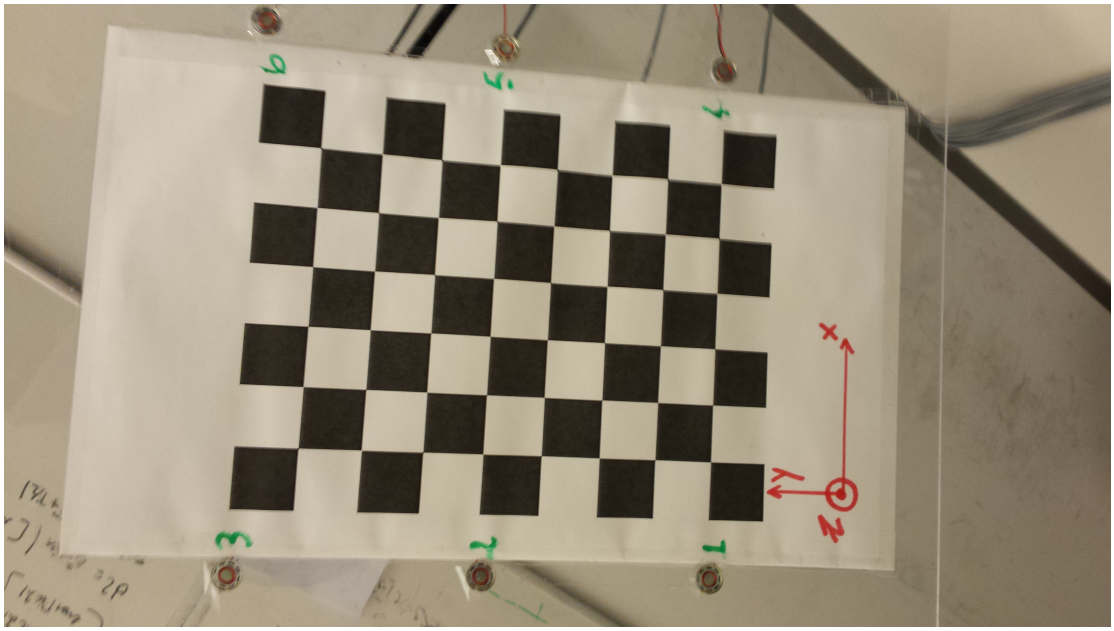

Figure 3: the calibration rig used. It contains 6 labeled speakers at known location

b) Microphone array calibration

As mentioned earlier, microphone array calibration seeks at finding microphones coordination within the array with respect to the camera coordinates system.
The technique described by [2] consists in finding the microphones coordinates using sound sources at known locations.

Mapping sound source coordinates

During the camera calibration phase, with each image taken, the speakers that are attached to the calibration rig emit a pilot signal. This signal is recorded simultaneously at each of the microphone array channels.
Using the camera extrinsic parameters, the speakers position in the calibration rig are mapped to the camera's reference frame.
In other words, if the speaker coordinates in the calibration rig coordinates system are $X_s^{RealWorld}$, then their corresponding coordinates in the camera reference frame are

$$X_s^{camFrame} = R^{image} X_s^{RealWorld} + T^{image}$$

Where R^{image} and T^{image} are the camera's extrinsic parameters for a particular image. Figure 2 shows such a mapping.

Measuring time of flight (TOF)

The next step is to compute the time of flight that the pilot signal spend to reach a given microphone from a given sound source location.
This time of flight corresponds to the delay between the pilot signal and the signal measured at the microphone. The TOF is usually measured using the Generalized Cross Correlation.
In fact the time delay between 2 signals corresponds to the peak of the GCC function of the two signals

$$R_{xy}(\tau) = \int_{-\infty}^{+\infty} \Psi_{xy} X(\omega) Y(\omega)^* e^{-i\omega\tau} d\omega$$

In a discrete time signal the best resolution one can achieve for the delay using this method is equal to the sampling period.
The resolution of the time delay estimation influences the accuracy of our calibration procedure. With a sampling frequency of 44.1KHz an error of ½ sample results in an error of ~4mm in the position estimation.
Thus, it is a critical to estimate the time delay accurately. Methods that interpolate the discrete GCC function between samples surrounding the peak are used in order to estimate the time delay in the subsample precision [4] describes an iterative method that proves its high performance especially in severe noise situations. [4] in its iterative method expresses the continuous time cross correlation function as:

$$R_{xy}(t) = S_{XY}(0) + \sum_{n=1}^{\left(\frac{N}{2}\right)-2} 2S_{XY}(n) e^{\frac{2\pi i n t}{Np}}$$

Then it is optimized to find the time delay such that:

$$\hat{d} = \underset{t}{\operatorname{argmax}}\, R_{xy}(t)$$

No closed form solution exists. So [4] uses successive parabolic interpolation in an iterative manner.

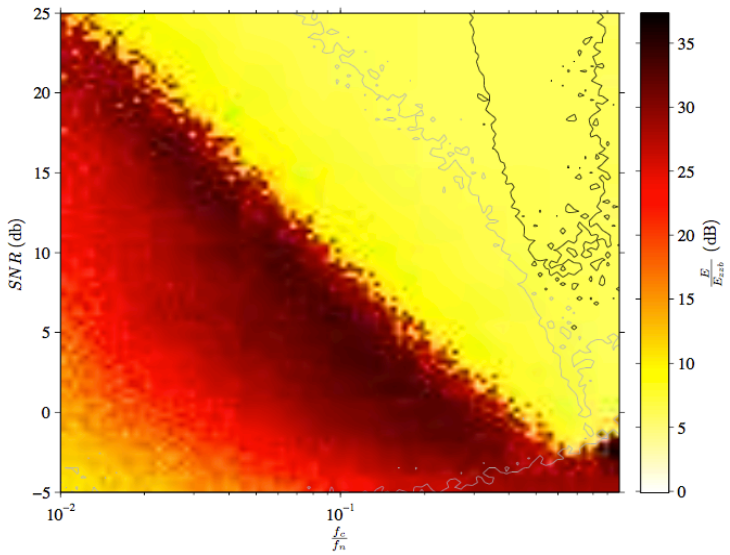

Figure 4: RMSE error for iterative subdelay estimation

Figure 4 shows the RMSE error for the iterative estimation method. The black lines show the area of the graph where this method has a better mean error than any other, with 95% confidence. The grey lines show the area where this method may be as

good as the best, with 95% confidence.

So far we have measured the time of flight between different sound sources and microphones, as well as the coordinates of the sound sources. Our goal is to estimate the coordinates of the microphones in the 3D camera's coordinates system.

c) Problem definition

Given a set if M microphones and S sound sources where their coordinates in the camera's 3D characterize the sound sources coordinates system

$$s_j = \left[x_s^j \; y_s^j \; z_s^j\right]$$

Which are known for all speakers j=1..S

Let $m_j = \left[x_s^j \; y_s^j \; z_s^j\right]$ be the coordinates of microphones that need to be estimated.

We denote by $TOF_{ij}^{estimated}$ the time of the flight that is measured using the GCC function between j^{th} source and i^{th} microphone respectively.
Let TOF_{ij}^{actual} be the time of flight of the acoustic signal

$$TOF_{ij}^{actual} = \frac{\left\|m_i - s_j\right\|}{c}$$

Where c is the speed of sound in the acoustic medium.

By assuming a Gaussian noise model for our observations, we can derive the maximum likelihood estimator as follow: Let Θ be a vector regrouping all the unknown parameters that need to be estimated.

Let $T(\Theta)$ be a vector of length M times S representing all the time of flights.
We can then express the model of the observations as

$$\Gamma = T(\Theta) + \eta$$

Where η is a zero-mean AWGN vector of length MS where each element has a variance σ_j^2 and of covariance matrix Ξ

Then the likelihood function of Γ can be written as

$$p(\Gamma|\Theta) = \aleph(T(\Theta), \Xi)$$

Hence,

$$p(\Gamma|\Theta) = (2\pi)^{-MS}(\det \Xi)^{-MS} e^{-\frac{1}{2}(\Gamma - T(\Theta))^t \Xi^{-1}(\Gamma - T(\Theta))}$$

Then we can express the simplified log-likelihood ratio as follow:

$$\text{loglikelihood ratio} = -\frac{1}{2}\left(\Gamma - T(\Theta)\right)^t \Xi^{-1}\left(\Gamma - T(\Theta)\right)$$

Maximizing the log likelihood function boils down to minimizing the distance between the observed, i.e.

$$\widehat{\Theta} = \arg\max_{\theta} -\frac{1}{2}\left(\Gamma - \mathrm{T}(\Theta)\right)^{\mathrm{t}}\Xi^{-1}\left(\Gamma - \mathrm{T}(\Theta)\right)$$

$$\widehat{\Theta} = \arg\min_{\theta} \frac{1}{2}\left(\Gamma - \mathrm{T}(\Theta)\right)^{\mathrm{t}}\Xi^{-1}\left(\Gamma - \mathrm{T}(\Theta)\right)$$

$$\widehat{\Theta} = \arg\min_{\theta} \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{S}\frac{(\mathrm{TOF}_{ij}^{\mathrm{actual}} - \mathrm{TOF}_{ij}^{\mathrm{estimated}})^2}{\sigma_{ij}^2}$$

We observe that the ML estimate is the solution of the non-linear least squares problem.
The Levenberg-Marquardt method can be used iteratively to find the microphones coordinates that minimize the objective function.
However, this minimization routine is very sensitive to initial conditions.
It may converge to another local minima if we do not provide a good enough starting condition.
[2] Uses an iterative algorithm to solve for the unknown microphones coordinates.

1) Initialization of the source coordinates with random numbers. Initialization of the microphones' coordinates with approximated values or random numbers. Consideration of boundary conditions due to the array geometry (e.g. two-dimensional array: all microphones are nearly arranged in-plane and all sources are in front of this plane).

2) Shifting of the first source with step size ss1 in x-, y-, z-direction by considering the boundary conditions until the sum of the error squares reaches its minimum

3) Repeat step 2 for all sources

4) Apply steps 2 and 3 to all microphones

5) Repeat steps 2-4 until the optimization of the errors squares ends

6) Reduce the step size ss1

7) Repeat steps 2 - 6 until termination condition has been reached

8) Verify the convergence criterion

9) If step 8 fails, change starting conditions and repeat the procedure from step 1

d) Practical issues

Severe errors may be introduced due to the uncertainties in the speed of sound. In fact the air temperature has a wide influence on the speed of sound.

$$c = 331.5\ ms^{-1}\sqrt{1 + \frac{\vartheta/°C}{273.15}}$$

According to the above equation a slight change in the air temperature (1°C) will lead to considerable change in the speed of sound ($0{,}6ms^{-1}$).
[2] Uses a reference microphone at a known position in order to enable for a speed of sound calculation.

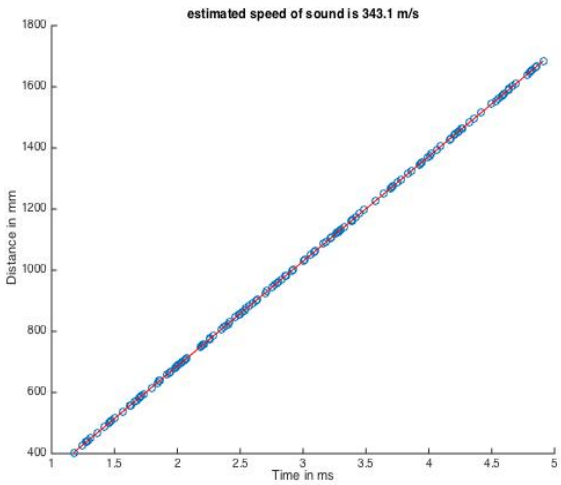

Figure 5: Least square fit for the speed of sound.

In other words, the TOF from all the sound sources to the reference microphone are collected, and then a least square linear fit is used to determine the speed of sound (slope of the regression vector)

e) Contribution and results

The contribution of this work compared to the work of [2] is the fact of introducing a smart initial guess to the non-linear least square optimization algorithm.

Closed form solution:

This solution was described by [5] it states that having S TOF measurement of the distance between a microphone and S sound source one can write

$$\left\|m_i - s_j\right\|^2 - \left\|m_i - s_k\right\|^2 = c^2 TOF_{ij}^2 - c^2 TOF_{ik}^2$$

In vector form this gives

$$(s_k - s_j)^T m_i = b_{jk}^i$$

Where,

$$b_{jk}^{i} = \frac{(c^2 TOF_{ij}^2 - c^2 TOF_{ik}^2 - \|s_j\|^2 + \|s_k\|^2)}{2} \quad (**)$$

Hence, each pair of sound sources give rise to an equation of 3 unknowns.
Having S sources, we will have S(S-1)/2 equation if we write **(**)** in matrix form we get

$$A = \begin{bmatrix} (s_1 - s_2)^T \\ .. \\ .. \\ (s_{S-1} - s_S)^T \end{bmatrix}, \qquad \boldsymbol{b}^i = \begin{bmatrix} b_{21}^i \\ .. \\ .. \\ b_{S(S-1)}^i \end{bmatrix}, \; A\,\boldsymbol{m}^i = \boldsymbol{b}^i$$

The closed form solution for the i^{th} microphone is then given by

$$m_i^{\text{closed form}} = (A^T A)^{-1} A^T b^i$$

Once we have this closed form solution we feed it to the non-linear least squares optimization routine as the initial guess to further refine it and give the final ML estimate.

Results:

Through out this project a software with a GUI has been developed in order to support the user during the measurement and the calibration phase.

The measurement process is as follow:
The user is first asked to provide the necessary parameters of the calibration such as the number of microphone in the array the number of speakers attached to the calibration rig, the calibration signal to be used. Then the user can start the calibration data acquisition.

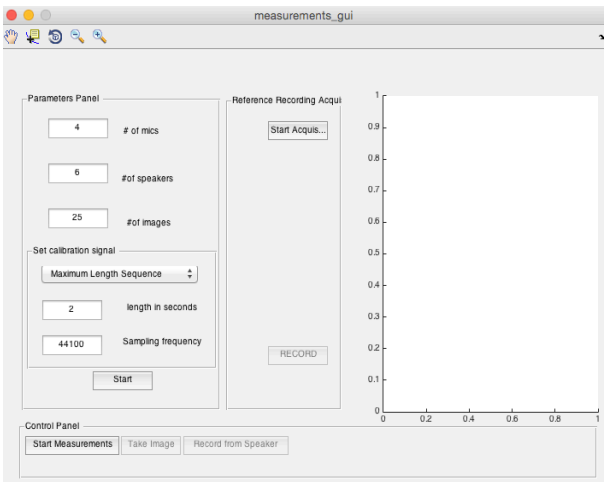

Figure 7: User interface for the measurement phase.

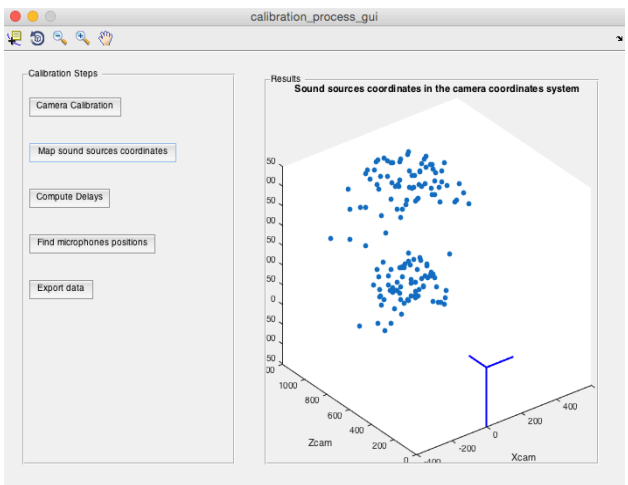

Figure 6b: User interface for the calibration phase.

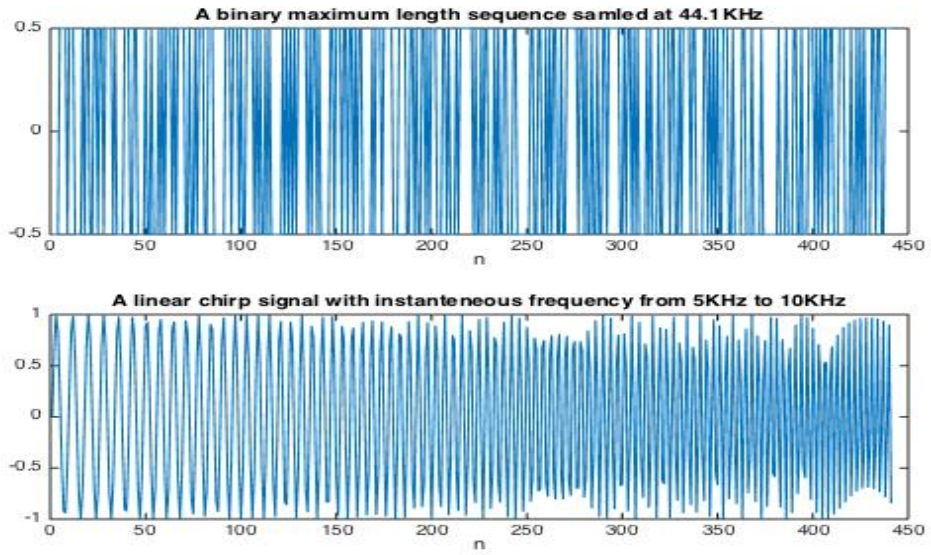

Figure 8: Different signals that can be used for the calibration

Once the measurement phase is finished, the user can proceed to the calibration phase, where he first needs to calibrate the camera, and then find the sound source location in the camera reference system.

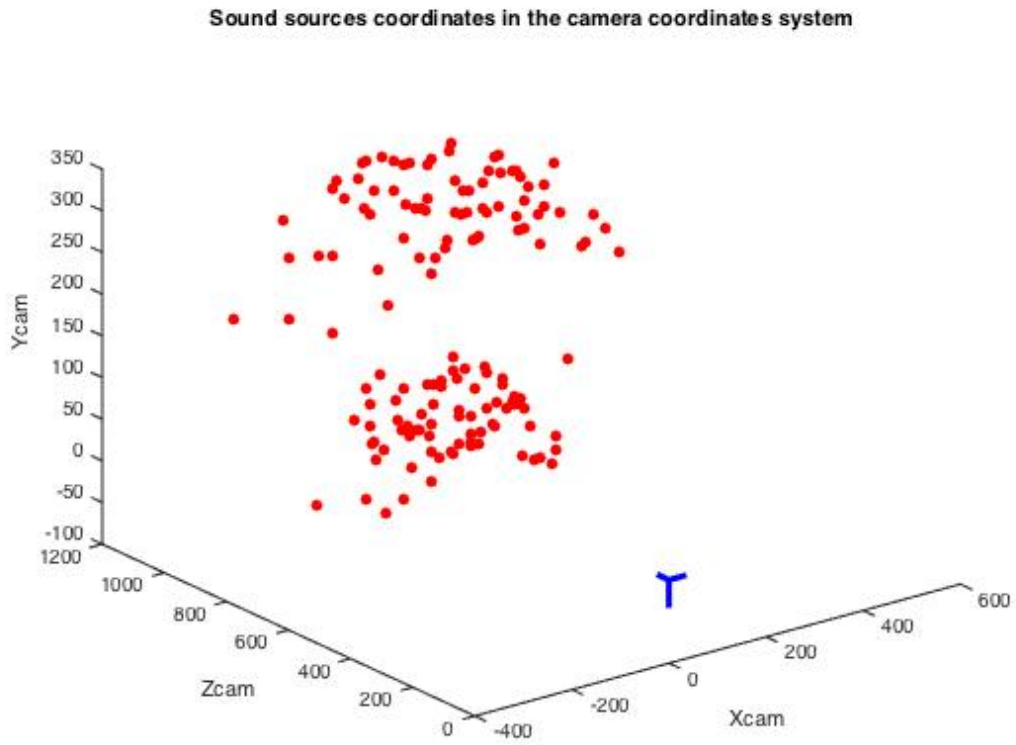

Figure 9: Sound sources mapped to the camera coordinates system

Then, a call to the main calibration algorithm will first compute the initial guess using the closed form solution then will call the ML refinement phase.

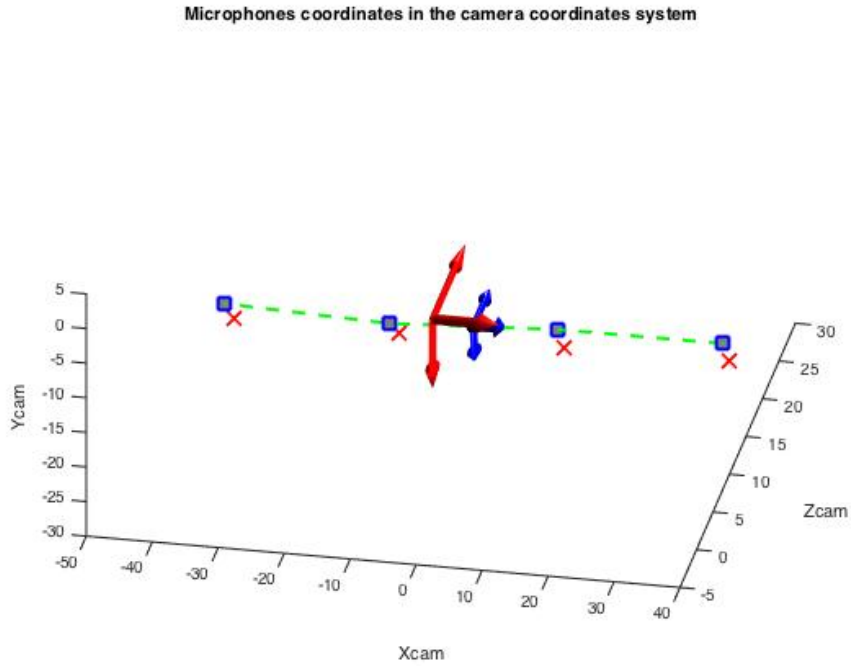

Figure 10: In bleu are the microphones coordinates estimated using the closed form solution. In red are the refined positions

This calibration technique was tested on a linear microphone array with 4 microphones. The camera was attached to the center of the microphone array (it could have been placed in an ad hoc position). 25 images were used for the camera calibration, which give rise to 150 sound sources (6 speakers were attached to the calibration rig).

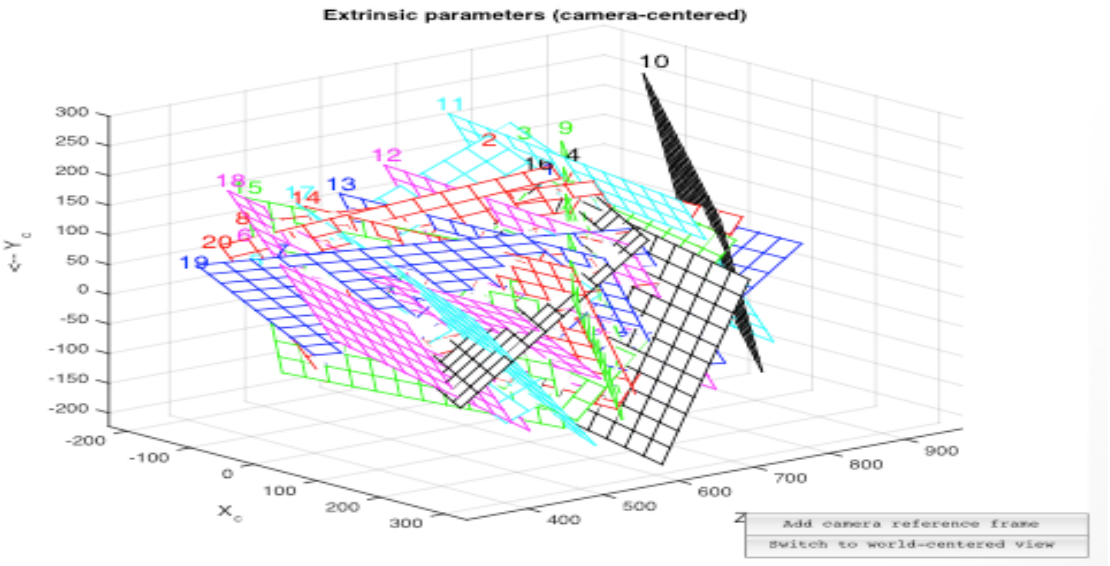

Figure 11: different positions of the calibration pattern.

Initially the microphones positions were carefully measured using manual process. Then coordinates of the microphone array were estimated and an error of less than 1.5

mm was achieved. The shape of the array was correctly estimated (inter microphone distances and alignment).
Figure 10 shows the estimated microphone positions using the closed form solution, then their positions after the maximum likelihood refinement phase.

4 References

[1] E. Ettinger and Y. Freund, “Coordinate-free calibration of an acoustically driven camera pointing system,” in Proc. 2nd ACM/IEEE ICDSC, Stanford, CA, USA, Sep. 2008, pp. 1–9.

[2] IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 22, NO. 10, OCTOBER 2013

Mathew Legg and Stuart Bradley

A Combined Microphone and Camera Calibration Technique With Application to Acoustic Imaging

[3] AUTOMATIC DETECTION OF MICROPHONE COORDINATES
Dirk Döbler, Gunnar Heilmann, Marcus Ohm

[4] A Comparison of Time Delay Estimation Methods for Periodic Signals
Travis Wiens, Stuart Bradley University of Auckland
Auckland, New Zealand

[5] AUTOMATIC POSITION CALIBRATION OF MULTIPLE MICROPHONES
Vikas C. Raykar and Ramani Duraiswami
Perceptual Interfaces and Realities Lab., University of Maryland, CollegePark