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

Mathew Legg and Stuart Bradley

***Abstract*— We present a calibration technique for an acoustic imaging microphone array, combined with a digital camera. Computer vision and acoustic time of arrival data are used to obtain microphone coordinates in the camera reference frame. Our new method allows acoustic maps to be plotted onto the camera images without the need for additional camera alignment or calibration. Microphones and cameras may be placed in an ad-hoc arrangement and, after calibration, the coordinates of the microphones are known in the reference frame of a camera in the array. No prior knowledge of microphone positions, inter-microphone spacings, or air temperature is required. This technique is applied to a spherical microphone array and a mean difference of 3 mm was obtained between the coordinates obtained with this calibration technique and those measured using a precision mechanical method.**

***Index Terms*— Camera, microphone, array, array shape, calibration, acoustic.**

I. Introduction

OPTICAL cameras are commonly attached to microphone arrays for applications such as generating acoustic images of sound sources on an object, for conference room applications, and for underwater sonar imaging. These applications generally require the coordinates of the microphones, the position and orientation of the cameras, and the speed of sound to be known. For example, acoustic maps may be generated from microphone array data using delay and sum beamforming and image sharpening (deconvolution) algorithms [1]–[5]. Errors in the acoustic map occur if the microphone coordinates and speed of sound are inaccurate. The required accuracy of the microphone coordinates scales with wavelength and so increases with audio frequency. Commonly, an eigenvalue calibration technique, developed by Dougherty [1], is used to correct for errors in the speed of sound, microphone phase, and microphone position. However, this calibration algorithm requires prior knowledge of the microphone coordinates. Underbrink [6] states that, for higher frequencies, this calibration technique requires the microphone position to be known with an accuracy better than 2.5mm.

Manuscript received August 1, 2012; revised January 22, 2013 and May 20, 2013; accepted May 26, 2013. Date of publication June 17, 2013; date of current version September 5, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Marios S. Pattichis.

M. Legg is with the Brunel Innovation Centre, Brunel University, Uxbridge UB8 3PH, U.K. (e-mail: mathew.legg@brunel.ac.uk).

S. Bradley is with the Physics Department, University of Auckland, Auckland 1142, New Zealand (e-mail: s.bradley@auckland.ac.nz).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2013.2268974

The microphone coordinates may be known with some accuracy if, for example, the array structure is built using a milling machine or laser cutter. If this is not the case, the coordinates of the microphones will need to be measured. Microphone coordinates may be obtained using manual methods. Birchfield et al. obtained the distance between microphones in an array with a tape measure and used multidimensional scaling to obtain the coordinates of the microphones [7], [8]. This method becomes very difficult as the number of microphones in an array increases. An array having M microphones would require $M(M-1)/2$ measurements of the distance between microphones. An alternative method is to use measurement tools such as a Faro-arm, a laser scanner, survey equipment [9], or a sonic digitiser [6]. Such equipment, however, is costly and often is not a practical option.

Even if the microphone positions are known accurately, there may be some phase differences between microphones. This will give the same effect as microphone position errors. Rather than finding the geometric coordinates of the microphones, more accurate results will generally be achieved if "acoustic microphone coordinates" are obtained, which allow for these phase errors. An automatic method for obtaining microphone positions using acoustic methods is, therefore, desirable.

Calibration techniques that obtain the coordinates of microphones in an array are commonly referred to as array shape calibration. Array shape calibration methods have been developed using sound sources where the source positions are unknown [10]–[23]. This type of calibration method is referred to as self calibration. A problem with self calibration techniques is that the microphone position errors are generally too large for high frequency beamforming measurements. To address this, Dobler et al. [24] developed a technique which increases the number of sources to 50 - 100 using a taser (spark generator). Time difference of arrival measurements were obtained using a reference microphone and an iterative method was used to obtain both microphone and sound source locations. After calibration, an accuracy of better than 2mm for a 2D array and 4mm for a 3D array was achieved. However, as with other self calibration methods, the microphone positions are obtained relative to each other, so are in an arbitrary reference frame.

Array shape calibration techniques, which use sound sources at known locations, have also been developed [25]–[28]. These techniques do not result in arbitrary reference frames. However, they require bulky calibration rigs and the number of sound sources, and hence the accuracy, is limited to the

number of sources in the calibration rig. Lauterbach et al. [29] used 8 tweeters at known locations in a panel. To increase accuracy of the measured source locations, the sound from the tweeters is funnelled through conduits inserted into the panel. Microphone position errors of the order of 1mm could be achieved. The error in microphone positions was shown to decrease as the number of sources was increased.

Acoustic imaging microphone arrays often contain a camera which is usually located at the centre of the array. None of the previous array shape calibration methods have included camera alignment or calibration. An acoustic map, generated using the microphone data, is overlayed as a transparency over the camera image. These are sometimes referred to as "acoustic cameras". Almost no literature was found relating to the methods used to align the camera with the microphones for such systems. Dougherty describes in Optinav's Array 24 operation manual [30] a method of aligning a camera with a beamformed map using manual rotation and translation of coordinates in software. Camera distortion correction can also allowed for. Bing et al. [31], [32] use two stereo cameras on either side of a cross microphone array for acoustic holography imaging. However, the method used to obtain the rigid body transformation between the cameras and the microphone array is unclear.

Alignment of the camera with the array is also used in conference room literature. Cameras may be used to identify the location of a person's head. This information is then used to "steer" the microphone array towards the person for speech enhancement. Most conference room array literature found assumed that the alignment between the camera and the array was known. Some found the alignment but required that the camera could see the array [33]–[36]. Ettinger et al. [37] present a calibration technique for a system that automatically steers a pan and tilt (PTZ) camera using an ad-hoc microphone array. Rather than finding microphone coordinates it obtains the relationship between the time differences of arrivals of microphones and the pan and tilt angles of the PTZ camera.

Calibration methods have also been developed for "opti-acoustic sensors" which merge an ultrasound sonar device with cameras for underwater object imaging. The parameters relating the cameras to the sonar devices were determined using computer vision techniques [38]–[43]. A sonar scan made over a checkerboard has also been used [44], [45]. These devices differ from microphone phased arrays, which do not transmit sound but only listen to the sound emitted by objects and have very low spatial resolution in the direction along the array axis.

A. Motivation for New Calibration Technique

The calibration technique presented in this work was developed at the University of Auckland as part of research comparing 2D and 3D acoustic imaging methods [46], [47]. This work combined computer vision and microphone array techniques and, therefore, required accurate knowledge of microphone coordinates and the position and orientation of any cameras in the array. The initial calibration of the microphone arrays and cameras was done using a manual process. The microphone coordinates for a 2D array had been obtained using the CAD model used to build the array panel, while those for a 3D spherical array, see Figure 1, were obtained using a *Faro-Arm*[1]. In both cases, however, the uncertainty in these microphone coordinates was considered to be too large. Careful labelling of each microphone and holder ensured that microphone coordinates were always associated with the correct microphone. The main camera in an array was initially roughly positioned in the array structure. A speaker was then setup approximately a meter in front of the array. The microphone coordinates were then incrementally rotated and translated in software until the main peak in the beamformed/deconvolution map overlaid the speaker in the camera image for a range of speaker positions. This calibration process was time-consuming, however. It was also very difficult to determine if a position error in an acoustic map was due to errors in the speed of sound, the coordinates or phase of microphones, the alignment of the cameras with the array, or due to scanning the array over specified 2D or 3D scanning surfaces. An automatic calibration technique was, therefore, desirable which combined microphone and camera calibration and which ideally did not require any prior knowledge of microphone or camera positions or the speed of sound. No existing calibration technique was found in the literature which fulfilled these requirements.

Fig. 1. Photo of a 72 element spherical microphone array containing three cameras which was used to test the calibration technique.

In this paper, we will present a calibration technique that combines camera calibration with microphone position calibration. Microphones and cameras may be placed in arbitrary positions and orientations. No *a priori* information about microphone or camera positions is required. A small calibration rig is used which consists of sound sources surrounding a checkerboard pattern attached to a piece of plexiglass. The 3D coordinates of the sound sources are obtained, in the reference frame of a camera, using information obtained during camera calibration. These sound source locations, combined with Time of Flight *(TOF)* data, are used to obtain the microphone positions in the camera's reference frame. We have applied

[1] www.faro.com/gage/

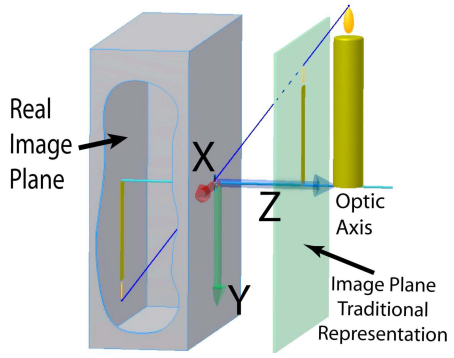


Fig. 2. Diagram illustrating the pinhole camera model. It is usual to ignore the inversion of the image and to instead assume that an non-inverted image is formed on a virtual image plane at a distance f_o in front of the camera.

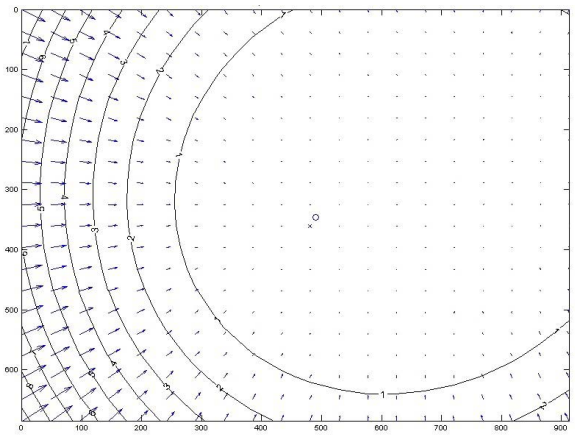


Fig. 3. Example of camera distortion. Both radial and tangential distortion can be seen.

this technique to a spherical array and have found a distance difference of 3mm compared to those obtained using a *Faro-Arm*.

II. Computer Vision Theory

The most basic model of a camera is the pinhole camera. Light passes through a pinhole and forms an inverted image on the screen a distance f_o behind. A right handed camera reference frame is defined such that the origin is at the pinhole, the X and Y axes are parallel to the image plane and the Z axis points outwards, see Figure 2. Consider light propagating from a point P having 3D coordinates

$$\vec{\boldsymbol{X}} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{1}$$

defined in this camera reference frame. This light would form an image at point P on the image plane having 2D coordinates

$$\vec{\boldsymbol{x}} = f_o \, \vec{\boldsymbol{x}}_N \tag{2}$$

where

$$\vec{\boldsymbol{x}}_N = \begin{bmatrix} X/Z \\ Y/Z \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}. \tag{3}$$

All depth information is lost in the projection.

Instead of a pinhole, real cameras contain lenses which introduces distortion, see Figure 3. One commonly used fifth order distortion model, which assumes radial and tangential distortion [48]–[50], may be described by

$$\vec{\boldsymbol{x_d}} = \left(1 + \boldsymbol{\zeta}_1 \, R^2 + \boldsymbol{\zeta}_2 \, R^4 + \boldsymbol{\zeta}_5 \, R^6\right) \vec{\boldsymbol{x}}_N + \vec{\delta}_x \tag{4}$$

where

$$\vec{\delta}_x = \begin{bmatrix} 2 \, \boldsymbol{\zeta}_3 \, x \, y + \boldsymbol{\zeta}_4 \, (R^2 + 2 \, x^2) \\ \boldsymbol{\zeta}_3 \, (R^2 + 2y^2) + 2 \, \boldsymbol{\zeta}_4 \, x \, y \end{bmatrix}, \tag{5}$$

$\boldsymbol{\zeta}$ is a (5×1) array of radial and tangential distortion coefficients and $R = \sqrt{x^2 + y^2}$. For many cameras, a first order distortion

$$\vec{\boldsymbol{x_d}} = \left(1 + \boldsymbol{\zeta}_1 \, R^2\right) \vec{\boldsymbol{x}}_N, \tag{6}$$

is often sufficient.

Modern cameras obtain a digital image using pixel sensors. Each pixel may have a different size in the X and Y axes directions. The alignment of the sensors in the direction of these axes may not be at right angles but be skewed by an angle α_c. Also, the convention is to define the top left pixel as the origin of the pixel coordinate system. Therefore, the light from point P will be projected to a point P having pixel coordinates

$$\vec{\boldsymbol{p}} = \begin{bmatrix} \boldsymbol{f}_{c\,1} \; (\vec{\boldsymbol{x_d}}1 + \alpha_c \; \vec{\boldsymbol{x_d}}2) + \boldsymbol{c}_{c\,1} \\ \boldsymbol{f}_{c\,2} \; \vec{\boldsymbol{x_d}}2 + \boldsymbol{c}_{c\,2} \end{bmatrix}. \tag{7}$$

where $\boldsymbol{f}_c$ is the camera's focal length in pixels and $\boldsymbol{c}_c$ is the pixel coordinates of the center of projection. The parameters $\boldsymbol{f}_c$, $\boldsymbol{c}_c$, α_c, and $\boldsymbol{\zeta}$ are referred to as the intrinsic parameters of a camera and describe the projection of 3D coordinates, in the camera reference frame, into 2D pixel coordinates.

A. Extrinsic Parameters

In the previous section, the 3D coordinates for point P were defined in the camera reference frame. In practice, it is usually more convenient to define the coordinates in some 'real world' reference frame. In order to convert the coordinates from a 'real world' reference into the camera reference frame, the rigid body rotation and translation

$$\vec{\boldsymbol{X}}_{\{\mathrm{Cam},\mathcal{F}\}} = \mathbf{R} \, \vec{\boldsymbol{X}}_{\{\mathrm{RW},\mathcal{F}\}} + \vec{\boldsymbol{T}}. \tag{8}$$

may be used where $\mathbf{R}$ is a (3×3) rotation matrix and $\vec{\boldsymbol{T}}$ is a (3×1) translation vector. Together $\mathbf{R}$ and $\vec{\boldsymbol{T}}$ define the transformation from the 'real world' reference frame into the camera reference frame and are referred to as the extrinsic parameters of the camera.

B. Camera Calibration

Camera calibration seeks to obtain the intrinsic parameters ($\boldsymbol{f}_c$, $\boldsymbol{c}_c$, α_c, and $\boldsymbol{\zeta}$) and extrinsic parameters ($\mathbf{R}$ and $\vec{\boldsymbol{T}}$) of a camera. A commonly used calibration technique uses images obtained of a checkerboard image from different positions and orientations [49]–[51]. Between 10 and 20 images of a checkerboard pattern are often used. For each image, the pixels corresponding to the checker pattern corners are obtained. These combined with the dimensions of each checker square

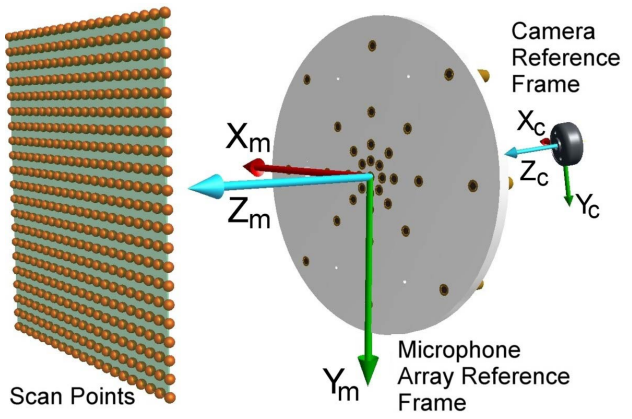


Fig. 4. Diagram illustrating the coordinates of acoustic imaging scan point coordinates $\vec{\xi}$ and microphone coordinates $\vec{X}_m$ defined in a "microphone array reference frame". These coordinates may be converted into the camera reference frame using a rigid body transformation.

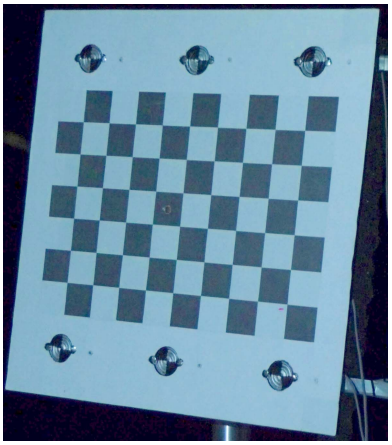


Fig. 5. Photo of calibration rig used to calibrate cameras and the coordinates of microphones. It consists of six speakers surrounding a checkerboard pattern attached to a sheet of plexiglass.

are used to obtain the calibration parameters. For each image, the extrinsic parameters $\mathbf{R}$ and $\vec{T}$ of a camera are obtained. These describe the camera's position and orientation relative to the checkerboard 'real world' reference frame. This reference frame is right handed and is usually defined such that one extreme inner corner is the origin, X and Y axes lie in the plane of the board and are aligned with the checker pattern such that Z points out of the board, see Figure 7. This calibration method may also be used for stereo camera calibration where the extrinsic parameters required to convert between two or more cameras may be calculated. This is required for stereo triangulation techniques [52], [53]. Refer to works by [54], [55], and [56] for more details.

III. Projection of an Acoustic Map onto a Camera Image

Algorithms such as delay and sum beamforming may be used to generate an acoustic map from microphone data. These algorithms generally use time delays which may be calculated, in the near-field, using

$$\Gamma_{mn} = \frac{\|\vec{X}_m - \vec{\xi}_n\|}{c} \tag{9}$$

where $\vec{X}_m$ is the microphone coordinates, $\vec{\xi}_n$ is the n^{th} scan point coordinates, and c is the speed of sound. Usually the coordinates $\vec{X}_m$ and $\vec{\xi}$ will be defined in a traditional "microphone array" reference frame. For a planar microphone array, this reference frame is generally defined to have its origin at the centre of the array, the X and Y axes in the plane of the array, and has the Z axis pointing outwards from the array, see Figure 4. The acoustic map may be plotted as a transparency over the camera's image. This requires that the acoustic map coordinates are converted into the camera reference frame. This could be achieved by placing the camera at the center of the array and physically rotating the camera until the camera and microphone reference frames are the same. This process can be time consuming. A more general method is to convert the microphone coordinates into the camera's reference frame using

$$\vec{\xi}_{\{\mathrm{Cam}\mathcal{F}\}} = \mathbf{R}_m \, \vec{\xi}_{\{\mathrm{Mic}\mathcal{F}\}} + \vec{T}_m,$$

where $\mathbf{R}_m$ and $\vec{T}_m$ are respectively a rotation matrix and a translation vector. The parameters $\mathbf{R}_m$ and $\vec{T}_m$ are the extrinsic parameters of the camera relative to the microphone reference frame.

The camera intrinsic parameter $\boldsymbol{\zeta}$, which describes lens distortion, may be used to undistort a camera image. The coordinates $\vec{\xi}_{\{\mathrm{Cam}\mathcal{F}\}}$ may then be projected onto undistorted camera images using Equation (7) and the intrinsic parameters $\boldsymbol{f}_c$, α_c, and $\boldsymbol{c}_c$, see Section II. Therefore, the calibration parameters required to generate an acoustic map and accurately project it onto a camera image are

- Microphone coordinates $\vec{X}_m$.
- Extrinsic parameters $\mathbf{R}_m$ and $\vec{T}_m$ describing transformation between the microphone array and camera reference frames.
- Camera intrinsic parameters ($\boldsymbol{f}_c$, $\boldsymbol{c}_c$, α_c, and $\boldsymbol{\zeta}$).
- Speed of sound c.

IV. Materials and Methods

A. Microphone Array

The microphone array, that was used to test the calibration technique, consisted of 72 microphones in a spherical array structure, see Figure 1. The microphone data are simultaneously sampled at 96 $kSPS$ using three Data Translation DT9836 and six DT9816 boards. Three cameras were attached to the array: a small one at the front and one at either side.

B. Calibration Rig

A calibration rig was laser cut from plexiglass plastic and six 26mm diameter speakers inserted into the surface. A printed checkerboard pattern was glued over the top, see Figure 5. The coordinates of the speakers, relative to the checkerboard speaker pattern, are given in Table I. Each speaker was connected through cables to one of the analog outputs of Data Translation DT9836 boards. These analog

TABLE I
SPEAKER COORDINATES IN CHECKERBOARD REFERENCE FRAME

Speaker	X [mm]	Y [mm]	Z [mm]
1	-60	0	-2.5
2	210	0	-2.5
3	-60	240	-2.5
4	210	240	-2.5
5	-60	120	-2.5
6	210	120	-2.5

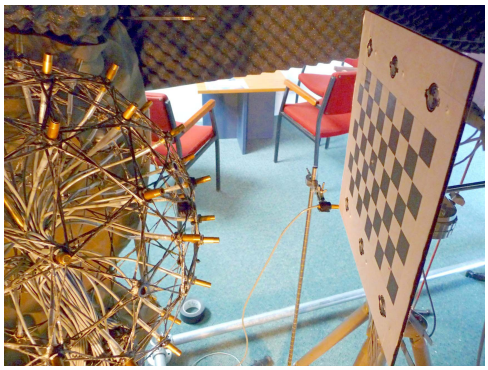


Fig. 6. Setup for obtaining the 3D coordinates of the reference microphone in the camera's reference frame. This was used for automatic speed of sound measurements.

outputs were synchronised to the analog inputs using an external clock and trigger. This ensured that a repeatable signal was able to be played and recorded. This calibration rig was attached to a tripod.

C. Data Acquired for Calibration

The first step in data acquisition was to obtain a reference recording for each speaker. This was obtained by placing a reference microphone about 2mm from the surface of the speaker and simultaneously playing and recording a short burst of maximum length sequence white noise. The reference signal was to be used later for time of flight calculations.

If automatic speed of sound measurements were to be made, the reference microphone was then placed on a stand in front of the array just outside the view of the camera. The corner of the checkerboard calibration rig was placed at the centre of the microphone, see Figure 6, and a camera image obtained of the checkerboard pattern. This image was later used to enable the 3D coordinates of a reference microphone to be obtained in the camera's reference frame.

The checkerboard rig was then removed and setup in view of the camera. A camera image of the checkerboard was then taken. For each speaker in turn, a burst of white noise, identical to that used for the reference recording, was played and simultaneously recorded on all microphone channels. A separate recording file was created for each sound source. This process was repeated for a range of positions and orientations of the calibration rig.

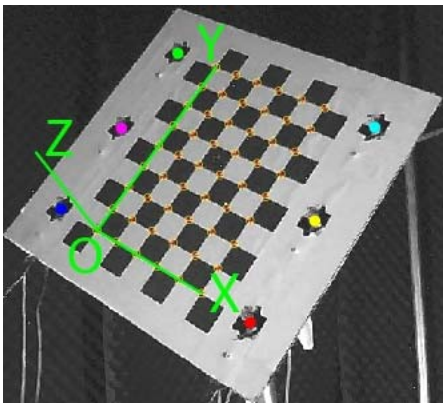


Fig. 7. Plot showing how the camera's extrinsic parameters may be obtained from an image of a checkerboard pattern. These may then be used to convert the 3D coordinates of the speakers from the checkerboard reference frame into the camera reference frame. The camera's intrinsic parameters may then be used to project the speaker positions onto the camera image (coloured dots).

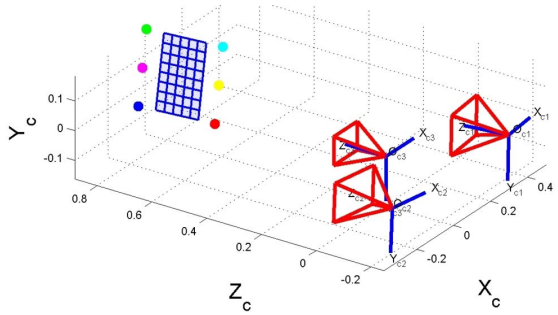


Fig. 8. Plot showing the 3D position and orientation of the checkerboard (blue grid) on the calibration rig and the coordinates of the six speakers (coloured dots) relative to the three cameras in the spherical microphone array. Note that the reference frame used in this plot is that of the front camera.

D. Camera Calibration

The camera calibration intrinsic parameters ($\boldsymbol{f}_c$, $\boldsymbol{c}_c$, α_c, and $\boldsymbol{\zeta}$) were obtained using the images of the printed checkerboard pattern and Bouguet's [50] *MATLAB Camera Calibration Toolbox*, see Section II-B. For each camera image of the checkerboard, the position and orientation of the camera relative to the checkerboard were obtained. These positions and orientations are referred to as the camera's extrinsic parameters ($\mathbf{R}$, $\vec{\boldsymbol{T}}$). If more than one camera partially share the same field of view, stereo calibration was then performed. Other software such as *openCV* [49] could have been used instead of Bouguet's toolbox.

E. Obtaining 3D Coordinates of Sound Sources and Reference Microphone Using a Camera's Extrinsic Parameters

The 3D coordinates of the speakers in the calibration rig may be obtained in a camera's reference frame using the extrinsic parameters obtained during camera calibration, see Figures 7 and 8. For each camera image, the extrinsic parameters $\mathbf{R}$ and $\vec{\boldsymbol{T}}$ will be obtained. If the coordinates of a speaker in the checkerboard 'real world' reference frame are $\vec{\boldsymbol{X}}_{s\{\mathrm{RW}\mathcal{F}\}}$, the coordinates of the speaker in the camera

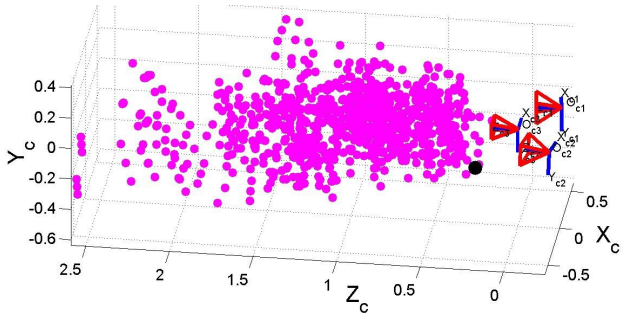


Fig. 9. Coordinates of sound sources (magenta dots) and reference microphone (black dot) relative to the three cameras in the microphone array.

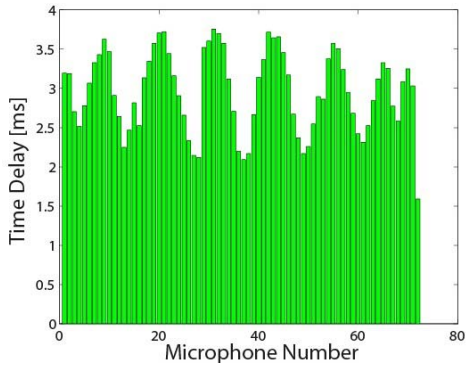


Fig. 10. Example of time delays for a single sound source location obtained for microphone position calibration.

reference frame will be

$$\vec{X}_{s\{\mathrm{CAM}\mathcal{F}\}} = \mathbf{R}\,\vec{X}_{s\{\mathrm{RW}\mathcal{F}\}} + \vec{T}. \tag{10}$$

This is performed for each speaker and each camera image to generate a point cloud of speaker coordinates, see Figure 9. The same technique was used to obtain the 3D coordinates of the reference microphone.

F. Time of Flight

A time of flight measurement was then obtained for each sound source location and each microphone in the array. The time difference of arrivals of signals between two microphones may be measured from the cross correlation of the microphone signals. The General Cross Correlation (GCC) is given by

$$\mathcal{R}_{m,m'}(\tau) = \int_{-\infty}^{+\infty} \boldsymbol{\Psi}_{m,m'}(\omega)\, \boldsymbol{U}_m(\omega)\, \boldsymbol{U}^{\dagger}_{m'}(\omega)\, \exp(i\omega\tau)\, d\omega \tag{11}$$

where $\boldsymbol{\Psi}$ is a function that allows filtering prior to peak detection [57]. If $\boldsymbol{\Psi} = 1$, then the standard cross correlation function is obtained. The time difference of flight is obtained by obtaining the peak in the cross correlation

$$\Delta_{m,m'} = arg\ \max_{\tau}\ \left[\ \mathcal{R}_{m,m'}(\tau)\ \right]. \tag{12}$$

Reference microphone recordings of the speakers had been made so that time of flight, rather than time difference of arrival, could be measured. For a given microphone recording, the delay between the microphone recording signal and a reference microphone recording $\Delta_{m,m_{ref}}$, was obtained using sub-sample delay estimation code written by a colleague [58], [59]. This iterative method fits the peak of a standard cross correlation to achieve sub-sample time delay measurements. Figure 10 shows an example of time delays obtained using this method for a single sound source and 72 microphones.

The theoretical time of flight (TOF) from a sound source at $\vec{X}_s$ to a microphone at $\vec{X}_m$ may be given by

$$\Gamma_{m,s} = \frac{\|\vec{X}_m - \vec{X}_s\|}{c}. \tag{13}$$

In reality, an extra term $\delta(\vec{X}_m, \vec{X}_s)$ could be added to allow for any extra propagation time due to extra distance travelled by the sound wave due to the microphone facing away from the sound source or due to obstructions such as the array structure or cables. Since all the sound sources used in this work were located in front of the array, it was considered that this effect would have minimal detrimental effect for the acoustic imaging described in Section VI and was, therefore, not addressed in this work.

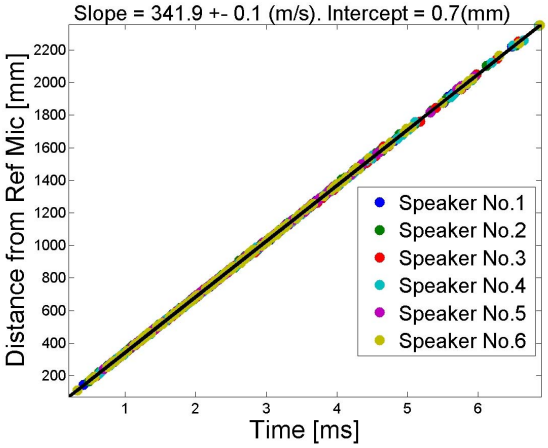


Fig. 11. Speed of sound calculation using time of flight measurements and the coordinates of sound sources and the reference microphone. These coordinates were obtained using computer vision techniques.

G. Calculating the Speed of Sound Using TOF and Computer Vision

The theoretical time of flight values given by Equation (13), rely on accurate speed of sound estimates. These may be obtained by measuring the temperature [28]. This, however, requires a thermometer with an accuracy of the order of $\pm 0.1^{\circ}$ for accurate measurements. Dobler et al. [24] suggest using measured distances between two or more microphones to enable solving for the speed of sound. Another factor that could lead to errors in Equation (13), is if the dimensions of the checker pattern used for camera calibrations were inaccurate. This could make the positions of the sources appear closer or further away than they really are.

To correct for this, a method was developed for measuring the speed of sound, which was automatically scaled by a correction factor, which allowed for any error in the dimension of the checker pattern. The position of a reference microphone

is obtained in the camera reference frame using the method described in Section IV-E. For each sound source, the Euclidean distance is calculated from the sound source coordinates to the reference microphone coordinates and the corresponding time of flight values obtained. A least squares fit is then performed to obtain the scaled valued of the speed of sound. In Figure 11, an example of speed of sound measurements is shown. The measured speed of sound obtained using this method was 341.9 m/s with standard deviation of 0.1 m/s.

H. Maximum Likelihood Estimator

The microphone coordinates may be obtained using the theoretical and measured time of flight values. A maximum likelihood estimator may be defined as

$$\Theta_{ML} = \arg\min_{\Theta} \left[J\left(\Theta, \Delta\right) \right] \tag{14}$$

where

$$J\left(\Theta, \Delta\right) = \sum_{m} \sum_{s} \frac{\left(\Gamma_{m,s}(\Theta) - \Delta_{m,s}\right)^2}{\varphi_{m,s}^2} \tag{15}$$

is the cost function to be minimised, Θ are the suite of parameters to be estimated and $\varphi_{m,s}$ is a weighting value related to the noise [15]. The parameters to be estimated are microphone coordinates and fine tuned sound source coordinates. If an initial estimate of the microphone coordinates is known, then the rotation and translation vector between the camera and microphone reference frame also need to be obtained. To allow for this, the microphone coordinates in the camera reference frame are defined as

$$\vec{X}_{m\,\{\mathrm{Cam}\mathcal{F}\}} = \mathbf{R}_m \, \vec{X}_{m\,\{\mathrm{Mic}\mathcal{F}\}} + \vec{T}_m.$$

Rather than using the rotation matrix, a rotation may be expressed in terms of a (3×1) rotation vector $\vec{\mathbf{\Omega}} = [\Omega_x, \Omega_y, \Omega_z]^{\mathsf{T}}$, where each component describes the amount of rotation around a given axis. $\mathbf{R}$ may be calculated from $\vec{\mathbf{\Omega}}$ using Rodrigues' formula

$$\mathbf{R} = \mathbf{I}_{3\times 3} \, \cos(\|\vec{\mathbf{\Omega}}\|) + [\Lambda] \frac{\sin(\|\vec{\mathbf{\Omega}}\|)}{\|\vec{\mathbf{\Omega}}\|} + \vec{\mathbf{\Omega}}\vec{\mathbf{\Omega}}^{\mathsf{T}} \, \frac{1 - \cos(\|\vec{\mathbf{\Omega}}\|)}{\|\vec{\mathbf{\Omega}}\|^2} \tag{16}$$

where

$$\Lambda = \begin{bmatrix} 0 & -\Omega_z & \Omega_y \\ \Omega_z & 0 & -\Omega_x \\ -\Omega_y & \Omega_x & 0 \end{bmatrix}. \tag{17}$$

and $\mathbf{I}_{3\times 3}$ is a 3×3 identity matrix [53], [55]. Therefore, the suite of parameters Θ to be solved for are $\vec{X}_m$, $\vec{T}_m$, $\vec{\mathbf{\Omega}}_m$ (or $\mathbf{R}_m$), and refined values of $\vec{X}_s$.

I. Algorithm Used to Find Microphone Coordinates

An iterative method is used to solve for the unknown parameters by the minimisation given in Equation (14). This method is based on the method described by Dobler et al. [24]. The entire process may be described by

1) Initialise $\mathbf{R}_m$ and $\vec{\mathbf{\Omega}}_m$ to [0;0;0];
2) Initialise microphone coordinates.
   a) If approximate microphone coordinates are known, these are used.
      i) Adjust $\vec{\mathbf{\Omega}}_m$ until a minimum of $J(\Theta)$ is reached.
      ii) Adjust $\vec{T}_m$ until a minimum of $J(\Theta)$ is reached.
      iii) Repeat (i) to (ii) with decreasing step sizes until an exit condition is met.
   b) Else, if no *a priori* knowledge of the microphone positions exist, initialise microphone coordinates using random numbers.
3) Adjust $\vec{X}_m$ for each microphone in turn until a minimum of $J(\Theta)$ is reached.
4) Repeat 3 with smaller step sizes until an exit condition is reached.
5) Adjust $\vec{X}_s$ for each sound source until a minimum of $J(\Theta)$ is reached.
6) Adjust weighting of $\varphi_{m,s}$ to decrease the effect of sound sources which have the highest contribution to $J(\Theta)$.
7) Repeat 3 to 6 with reducing step sized until exit condition is met.

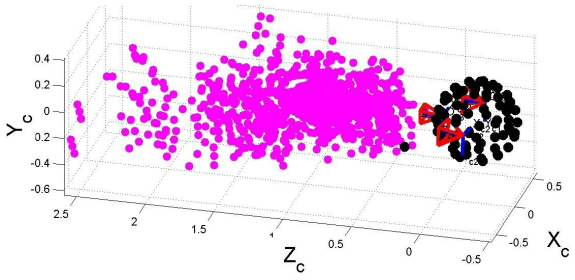


Fig. 12. Measured coordinates of microphones and sound sources relative to the three cameras in the array.

J. List of Parameters Obtained During Calibration

The parameters acquired during the entire calibration process are:

- Camera intrinsic ($\boldsymbol{f}_c$, $\boldsymbol{c}_c$, α_c, and $\boldsymbol{\zeta}$) and stereo extrinsic parameters.
- Speed of sound.
- Microphone coordinates $\vec{X}_m$ and extrinsic parameters $\vec{T}_m$ and $\mathbf{R}_m$ describing the rigid body transformation between the microphone and camera reference frames.
- Sound source coordinates $\vec{X}_s$ in the camera reference frame.

V. Results

Figure 12 shows the microphone and sound source coordinates relative to the cameras in the array obtained using this calibration method for the spherical array. No *a priori* knowledge of the microphone coordinates had been used. The microphone coordinates had been initialized to the coordinates [0;0;0]. For the calibration described here, 140 positions of the calibration rig were used giving 840 sound source locations. Figure 13 shows the coordinates of the microphones and cameras obtained by the calibration.

A. Difference Between the Microphone Coordinates and Those Obtained by Faro Arm

These microphone coordinates were compared with those obtained using the *Faro-Arm*. A mean Euclidean distance

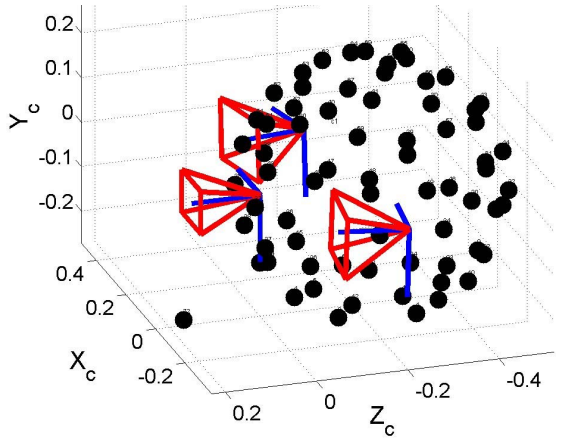


Fig. 13. Microphone coordinates relative to the cameras in the array.

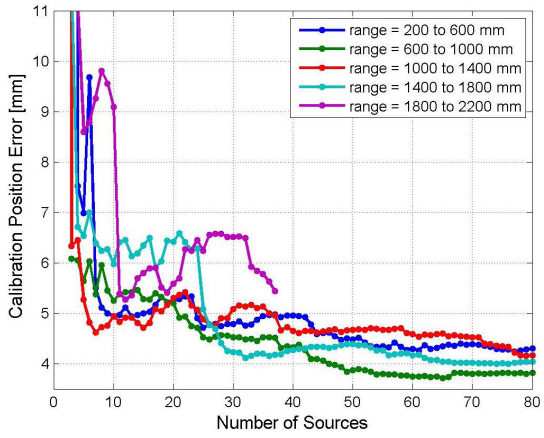


Fig. 14. Mean calibration microphone coordinates error as a function of number of sources for different ranges of distances of the sources from the array camera. In the calibration process used, only the microphone positions were adjusted. No optimisation of source locations was performed.

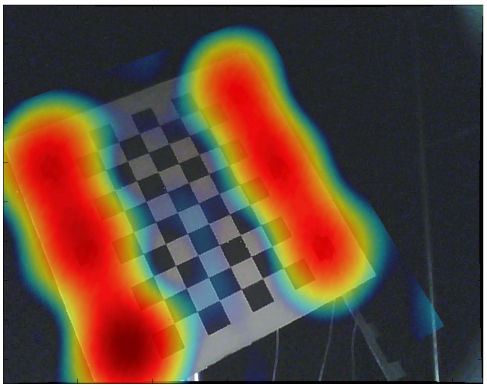
difference of 3mm (1/3 the diameter of the microphone including microphone case) was obtained relative to Faro-Arm measurements with a maximum difference of 11mm at the back of the array behind the camera. The difference between the two sets of measured coordinates could be partially due to the extra time of flight required for the acoustic signal to propagate to microphones facing away from the sound sources or to propagate around parts of the array structure or cables. If the microphone coordinates obtained using time of flight were indeed different from the true geometric coordinates, this does not necessarily mean an error for acoustic imaging/beamforming since it may be correcting for phased differences due to these extra travel paths.

Further processing of the raw calibration data was performed to investigate how the positions of the calibration rig and the total number of calibration sound sources affect the accuracy of the calibration results. Rather than making a single calibration using all the sound sources, as had been done previously, calibrations were instead repetitively performed using an incrementally increasing number of sound sources located within specific distance ranges from the array. For each distance range, a mean calibration microphone position error was calculated as a function of the number of sound sources

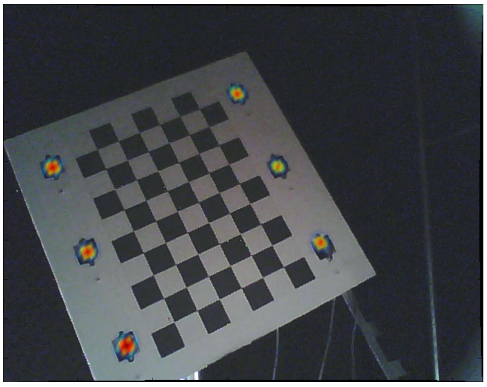
TABLE II

NUMBER OF SOURCES REQUIRED TO ACHIEVE A MICROPHONE POSITION ERROR OF LESS THAN σ FOR SOURCES IN FIVE DIFFERENT RANGES OF DISTANCES FROM THE ARRAY CAMERA. THIS TABLE USES DATA PLOTTED IN FIGURE 14

Source Range [m]	0.2-0.6	0.6-1.0	1.0-1.4	1.4-1.8	1.8-2.2
$\sigma = 10$ mm	3	3	3	3	4
$\sigma = 5.5$ mm	7	6	4	24	36
$\sigma = 5.0$ mm	23	18	35	25	
$\sigma = 4.5$ mm	51	35	71	27	



(a) Frequency domain beamforming acoustic map.



(b) Image sharpening of the beamformed map obtained using the *CLEAN-SC* deconvolution algorithm.

Fig. 15. Acoustic map overlayed over camera image using calibration data and a 3D scanning surface passing through the sound sources.

used. To reduce the processing time, the positions of the sound sources were not optimised during the calibrations. Figure 14 shows the mean difference in the microphone coordinates, compared to Faro-Arm measurements, as a function of the number of sound sources for five distance ranges. In general, the accuracy decreases as the sound sources are moved away from the array and as the number of sound sources is reduced. Table II summarises the results from this figure. While only three sound source locations were required to obtain a mean microphone position error of 10 mm, between 27 and 71 sound source locations were required to achieve an accuracy of better than 4.5 mm. The distance range of 0.6 to 1.0 m provided the lowest overall errors.

The estimates of the error in the calibration have been obtain using the Faro-Arm measurement of the microphone

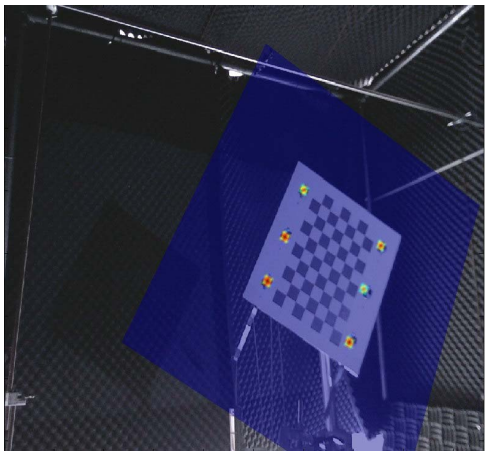


Fig. 16. Image from the camera to the left side of the microphone array with a 3D *CLEAN-SC* acoustic map overlayed. The *CLEAN-SC* map was generated using a 3D scanning surface passing through the sound sources. This enables cameras to be placed at arbitrary locations relative to the array and still achieve accurate registration of the acoustic map and camera image.

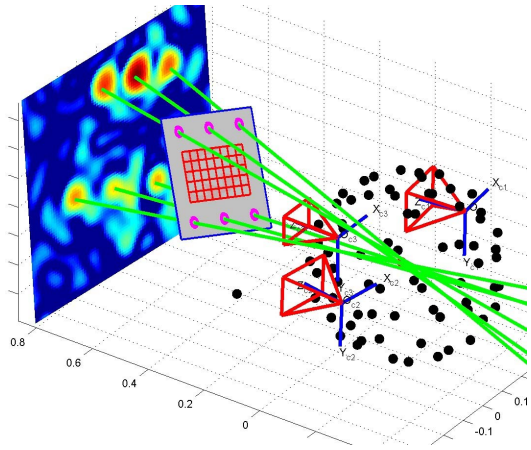


Fig. 17. Plot showing a beamformed map generated from experimental data using a 2D scanning surface located behind the sound sources. The green lines are rays passing from the location of the *CLEAN-SC* peaks in the acoustic map and passing through the corresponding speaker coordinates. The rays approximately cross at the center of the microphone coordinates. This shows that the microphone array may be modelled as a perspective projection camera where the center of projection is the center of the microphone coordinates.

coordinates. However, it should be noted that there will be some errors in these measurements. There will have been some parallax error associated with the angle of the 6mm diameter probe, at the end of the robotic arm, contacting the surface of the microphone. This parallax error might be expected to be negligible for some microphones but to increase to several mm for some microphones such as those at the back of the array. Also, during the Faro-Arm measurements, three of the microphones coordinates had been measured a second time and gave position errors of between 1 and 3.5mm compared to the previous measurements. Some movement of the array may have occurred during the measurements. The microphones were designed to be removable from the lattice array structure and were a firm slide fit into the microphone holders. Some additional movement of the microphones could, therefore, also have occurred during transport and experimental setup after these measurements were made.

Future work should include experiments with arrays with 2D grid or linear patterns of microphones, where a better estimation of the accuracy of the calibration method could be achieved. These arrays would not have the issues with extra propagation paths since all the microphones would be facing in the same direction and there would be no obstruction due to array structure or cabling. The microphone coordinates could also be more accurately measured since one can use a milling machine or a laser cutter. Also, if the array is made of a rigid material, it should maintain its shape better than a lattice type structure.

VI. Application for Acoustic Imaging

The accuracy of the calibration technique was next tested in relation to its ability to accurately locate sound sources using acoustic imaging [46]. For each position of the calibration rig, microphone recordings had been made where a burst of white noise had been played on one speaker at a time and then where uncorrelated white noise had been played on four and then on six speakers simultaneously. Using an automated process, frequency domain beamforming and *CLEAN-SC* deconvolution plots were then generated for each microphone array recording. First cross spectral matrices $\mathbf{G}$ were then generated using a data block length K of 4096 samples, an overlap of data blocks of fifty percent, and a Hanning window. The cross spectral matrices were then summed into twelfth octave frequency bands. These cross spectral matrices were then calibrated using Dougherty's eigenvalue calibration method [1]. A beamforming acoustic map $\boldsymbol{b}$ was then generated in each twelfth octave frequency band using

$$\boldsymbol{b}(\vec{\xi}_n) = \mathbf{w}^{\dagger}(\vec{\xi}_n)\ \mathbf{G}\ \mathbf{w}(\vec{\xi}_n), \quad \vec{\xi}_n = \vec{\xi}_1 \ldots \vec{\xi}_N \tag{18}$$

where $\vec{\xi}_n$ is the n^{th} scanning point coordinate and $\mathbf{w}$ is the array steering vector. The array steering vectors were calculated using either

$$\mathrm{w}_m(\vec{\xi}_n) = \frac{1}{M}\ \|\vec{X}_m - \vec{\xi}_n\|\ \exp(i\,\omega\,\|\vec{X}_m - \vec{\xi}_n\|/c), \tag{19}$$

or

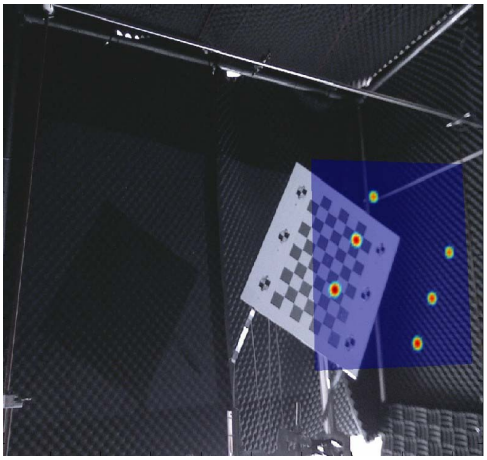
$$\mathrm{w}_m(\vec{\xi}_n) = \frac{\frac{\exp(i\,\omega\,\|\vec{X}_m - \vec{\xi}_n\|/c)}{\|\vec{X}_m - \vec{\xi}_n\|}}{\sqrt{\sum_{m=1}^{M} \frac{1}{(\|\vec{X}_m - \vec{\xi}_n\|)^2}}}, \tag{20}$$

where ω is the angular frequency [3].

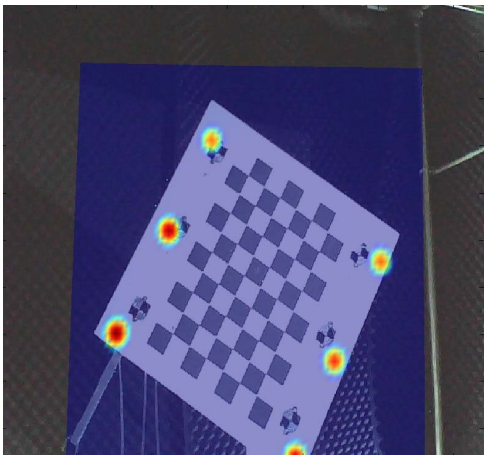
Deconvolution of these beamforming maps was performed using *CLEAN-SC* [5], however, other deconvolution algorithms such as *DAMAS*, *DAMAS*2, or *TIDY* [3], [4], [60] could have been used. Beamforming and *CLEAN-SC* were performed in an automatic process cycling through all recordings. Figure 15 shows an example of a beamforming and *CLEAN-SC* acoustic map. For each *CLEAN-SC* map, the peak magnitudes $\boldsymbol{b}_{max}$ and their coordinates $\vec{\xi}_{max}$ in the *CLEAN-SC* maps are obtained and used for error analysis.

A. 3D Scanning Surface Passing Through the Sound Sources

The 3D scan point coordinates $\vec{\xi}$ were generated which corresponded to the surface of the calibration rig. This was achieved by defining a 2D grid in the checkerboard reference



(a) Image from the camera to the left side of the microphone array with a 2D acoustic map overlayed.



(b) Image from the camera at the front of the microphone array with a 2D acoustic map overlayed.

Fig. 18. The *CLEAN-SC* map was generated using a 2D scanning surface located in front of the sound sources. Poor registration of the acoustic map and camera images is achieved. This illustrates the need for camera to be at the center of microphone coordinates if 2D acoustic maps, which do not pass through the sound source locations, are to be plotted over the camera's image.

frame and using the camera's extrinsic parameters and Equation (8) to convert it into the reference frame of the main array cameras in the array. Since the 3D scanning surface passed through the speakers, a position error $\boldsymbol{\varepsilon}$ for a source located at $\vec{\boldsymbol{X}}_s$ may be calculated as

$$\boldsymbol{\varepsilon} = \|\vec{\boldsymbol{X}}_s - \vec{\boldsymbol{\xi}}_{max}\|, \tag{21}$$

where $\vec{\boldsymbol{\xi}}_{max}$ is the coordinate of the corresponding peak in the deconvolution map. A mean position accuracy of 6.6mm was obtained for sound sources ranging from 0.5 to 3 m from the center of the array in the frequency range of 5 to 15 kHz. This is remarkably accurate considering that this is $1/4^{th}$ the diameter of the speaker and comparable to the scanning surface grid spacing of 6mm. This indicates that the calibration method provided accurate results. A benefit of using a 3D scanning surface which passes through the sound source locations is that the cameras may be placed at arbitrary locations relative to the array and still achieve accurate registration of the acoustic and camera images to be achieved, see Figure 16.

B. 2D Scanning Surfaces at an Arbitrary Location Relative to the Sound Sources

Beamforming and *CLEAN-SC* acoustic maps were also generated using traditional 2D scanning surfaces which were offset from the sound source locations. Such an offset can cause errors due to parallax/projection errors [4] or due to incorrect focus (time delays) being used for beamforming [46], [61]. Different types of projection were investigated to find a relationship between the coordinates of the *CLEAN-SC* peaks and the coordinates of the sound sources. The projection method that gave the most accurate results was perspective projection from the center of the microphone coordinates, see Figure 17. This indicates that a microphone array, which uses beamforming and deconvolution, may be modelled as a perspective projection camera with the center of projection being the mean of the microphone coordinates [34] and a depth of field which is related to the acoustic wavelength, the position of the sound source, and the offset of the scanning surface from the sound source location [46], [61]. The implication of this is that, since an optical camera also uses perspective projection, the camera should be located at the center of the microphone coordinates if accurate registration between the camera images and 2D acoustic maps are to be achieved in the near-field. This is illustrated in Figure 18 where poor registration can be seen in the projection of the 2D acoustic map onto images captured by two camera which are not located at the center of the array. Manufacturers of "acoustic cameras" place their camera at the center of their microphone arrays, apparently for this reason.

VII. Conclusion

We have presented a complete microphone/camera calibration technique. Microphones and cameras may be placed in an ad-hoc arrangement. No prior knowledge of microphone positions, inter-microphone spacings, or air temperature is required. The coordinates of the microphones are obtained in the reference frame of a camera in the array. This technique was applied to a spherical array. Microphone coordinates obtained had a mean Euclidean distance difference of 3mm compared to those obtained using a *Faro-Arm*. This difference was half the diameter of the microphone capsule. The calibration results were then applied to acoustic imaging using beamforming and *CLEAN-SC* algorithms. A total of 840 sound sources, located at distances ranging from 0.5 to 3m from the center of the array, were used to generate acoustic images. These were shown to provide accurate results. A mean position difference of 6.6mm of the sound source coordinates in the acoustic maps was obtained compared to the coordinates obtained using optical computer vision techniques. This difference was $1/4^{th}$ the diameter of the speakers and comparable to the acoustic map grid spacing of 6 mm.

Acknowledgment

The authors would like to thank S. Warrington, O. Caughley, and T. Waldmeyer of the *Faculty of Science Workshop* and B. Davis and M. Hollis of the *Physics Electronic Workshop*. They would also like to thank their colleagues A. Strehz,

P. Behrens, and especially B. Vallés for their encouragement. They would also especially like to thank Associate Prof. R. Dougherty of *OptiNav* and G. Heilmann and others at *GFaI* for their encouragement and advice.

REFERENCES

[1] R. P. Dougherty, "Beamforming in acoustic testing," in *Aeroacoustic Testing*, T. J. Mueller, Ed. Berlin, Germany: Springer-Verlag, 2002, pp. 63–97.

[2] W. M. Humphreys, T. F. Brooks, W. W. Hunter, and K. R. Meadows, "Design and use of microphones directional arrays for aeroacoustic measurements," in *Proc. 36st AIAA Aerosp. Sci. Meeting Exhibit.*, Reno, NV, USA, Jan. 1998, pp. 1–24.

[3] T. F. Brooks and W. M. J. Humphreys, "A deconvolution approach for the mapping of acoustic sources (DAMAS) determined from phased microphone arrays," in *Proc. 10th AIAA/CEAS Aeroacoust. Conf.*, May 2004.

[4] R. P. Dougherty, "Extensions of DAMAS and benefits and limitations of deconvolution in beamforming," in *Proc. 11th AIAA/CEAS Aeroacoust. Conf.*, Monterey, CA, USA, May 2005.

[5] P. Sijtsma, "CLEAN based on spatial source coherence," in *Proc. 13th AIAA/CEAS Aeroacoust. Conf.*, Rome, Italy, May 2007, pp. 357–374.

[6] J. R. Underbrink, "Aeroacoustic phased array testing in low speed wind tunnels in aeroacoustic testing," in *Aeroacoustic Testing*, T. J. Mueller, Ed. Berlin, Germany: Springer-Verlag, 2002, pp. 98–217.

[7] S. T. Birchfield, "Geometric microphone array calibration by multidimensional scaling," in *Proc. ICASSP*, vol. 5. Apr. 2003, pp. 157–160.

[8] S. Birchfield and A. Subramanya, "Microphone array position calibration by basis-point classical multidimensional scaling," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1025–1034, Sep. 2005.

[9] C. Manthe, A. Meyer, and F. Gielsdorf, "Geometric calibration of acoustic camera star48 array," in *Proc. Integrat. Generat., FIG Working Week*, Stockholm, Sweden, Jun. 2008, pp. 1–15.

[10] Y. Rockah and P. Schultheiss, "Array shape calibration using sources in unknown locations–part I: Far-field sources," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 35, no. 3, pp. 286–299, Mar. 1987.

[11] Y. Rockah and P. Schultheiss, "Array shape calibration using sources in unknown locations–part II: Near-field sources and estimator implementation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 35, no. 6, pp. 724–735, Jun. 1987.

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

[13] A. J. Weiss and B. Friedlander, "Array shape calibration using eiegenstructure methods," *J. Signal Process.*, vol. 22, no. 3, pp. 251–258, 1991.

[14] V. C. Raykar, I. Kozintsev, and R. Lienhart, *Position Calibration of Audio Sensors and Actuators in a Distributed Computing Platform*. New York, NY, USA: ACM, 2003, p. 572.

[15] V. C. Raykar and R. Duraiswami, "Automatic position calibration of multiple microphones," in *Proc. ICASSP*, vol. 4. Montreal, PQ, Canada, May 2004, pp. 69–72.

[16] V. Raykar, I. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 70–83, Jan. 2005.

[17] D. N. Zotkin, V. C. Raykar, R. Duraiswami, and L. S. Davis, "Multimodal tracking for smart videoconferencing and video surveillance," in *Multimodal Surveillance: Sensors, Algorithms, and Systems*. Norwood, MA, USA: Artech House, 2007.

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

[19] M. Hennecke, T. Plotz, G. A. Fink, J. Schmalenstroer, and R. Hab-Umbach, "A hierarchical approach to unsupervised shape calibration of microphone array networks," in *Proc. IEEE/SP 15th Workshop Stat. Signal Process.*, Aug./Sep. 2009, pp. 257–260.

[20] R. Moses and R. Patterson, "Self-calibration of sensor networks," *Proc. SPIE*, vol. 4743, pp. 108–119, Apr. 2002.

[21] S. Thrun, "Affine structure from sound," in *Advances in Neural Information Processing Systems*, vol. 18. Cambridge, MA, USA: MIT Press, 2006, pp. 1353–1360.

[22] M. Chen, Z. Liu, L.-W. He, P. Chou, and Z. Zhang, "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2007, pp. 22–25.

[23] P. D. Jager, M. Trinkle, and A. Hashemi-Sakhtsari, "Automatic microphone array position calibration using an acoustic sounding source," in *Proc. 4th IEEE Conf. Ind. Electron. Appl.*, May 2009, pp. 2110–2113.

[24] D. Döbler, G. Heilmann, and M. Ohm, "Automatic detection of microphone coordinates," in *Proc. 3rd Berlin Beamform. Conf.*, Berlin, Germany, Feb. 2010, pp. 1–11.

[25] L. Seymour, C. Cowan, and P. Grant, *Bearing Estimation in the Presence of Sensor Positioning Errors*, vol. 12. Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers, 1987, pp. 2264–2267.

[26] J. Lo and S. Marple, *Eigenstructure Methods for Array Sensor Localization*, vol. 12. Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers, 1987, pp. 2260–2263.

[27] J. M. Sachar, H. F. Silverman, and W. R. Patterson, "Position calibration of large-aperture microphone arrays," in *Proc. IEEE ICASSP*, vol. 2. Orlando, FL, USA, May 2002, pp. 1797–1800.

[28] J. Sachar, H. Silverman, and W. Patterson, "Microphone position and gain calibration for a large-aperture microphone array," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 42–52, Jan. 2005.

[29] A. Lauterbach, K. Ehrenfried, L. Koop, and S. Loose, "Procedure for the accurate phase calibration of a microphone array," in *Proc. 15th AIAA/CEAS Aeroacoust. Conf., 30th AIAA Aeroacoust. Conf.*, Miami, FL, USA, May 2009.

[30] B. Dougherty. (2012, Mar.). *Optinav Array 24 Operation* [Online]. Available: http://www.optinav.com/Array24Operation.pdf

[31] B. Li, D. Yang, L. Shao, and X. Lian, "Development of acoustic video camera system based on binocular vision and short-time beamforming," in *Proc. INTER-NOISE*, Oct. 2008.

[32] D. Yang, Z. Wang, B. Li, and X. Lian. (2011, May). Development and calibration of acoustic video camera system for moving vehicles. *J. Sound Vibrat.* [Online]. 330(11), pp. 2457–2469. Available: http://linkinghub.elsevier.com/retrieve/pii/S0022460X10008035

[33] D. Zotkin, R. Duraiswami, H. Nanda, and L. Davis, "Multimodal tracking for smart videoconferencing," in *Proc. IEEE Int. Conf. Multimedia Expo*, Aug. 2001, pp. 36–39.

[34] A. O'Donovan, R. Duraiswami, and J. Neumann, "Microphone arrays as generalized cameras for integrated audio visual processing," in *Proc. IEEE Conf. CVPR* , Minneapolis, MN, USA, Jun. 2007, pp. 1–8.

[35] A. O'Donovan, D. Ramani, and J. Neumann, "Sensing the world with arrays of microphones and cameras," in *Proc. 19th Int. Congr. Acoust. Madrid*, Sep. 2007, pp. 1–6.

[36] D. Schulz, G. Thompson, C. L. D. Lo, R. Goubran, and M. Nasr, "System and method of self-discovery and self-calibration in a video conferencing system," U.S. Patent 7 403 217, Jul. 22, 2008.

[37] 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.

[38] A. Fusiello and V. Murino, "Calibration of an optical-acoustic sensor," *Mach. Graph. Vis.*, vol. 9, nos. 1–2, pp. 207–214, 2000.

[39] A. Fusiello and V. Murino, "Augmented scene modeling and visualization by optical and acoustic sensor integration," *IEEE Trans. Visualizat. Comput. Graph.*, vol. 10, no. 6, pp. 625–636, Nov. 2004.

[40] S. Negahdaripour, "Calibration of DIDSON forward-scan acoustic video camera," in *Proc. MTS/IEEE OCEANS*, vol. 2. Washington, DC, USA, Sep. 2005, pp. 1287–1294.

[41] S. Negahdaripour, H. Sekkati, and H. Pirsiavash, "Opti-acoustic stereo imaging, system calibration and 3-D reconstruction," in *Proc. IEEE Conf. CVPR*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.

[42] S. Negahdaripour, H. Sekkati, and H. Pirsiavash, "Opti-acoustic stereo imaging: On system calibration and 3-d target reconstruction," *IEEE Trans. Image Process.*, vol. 18, no. 6, pp. 1203–1214, Jun. 2009.

[43] S. Negahdaripour, "A new method for calibration of an opti-acoustic stereo imaging system," in *Proc. MTS/IEEE OCEANS*, Sep. 2010, pp. 1–7.

[44] N. H. Vilarnau, "Integration of optical and acoustic sensor data for 3D underwater scene reconstruction," Ph.D. dissertation, Dept. Comput. Eng., Univ. Girona, Girona, Spain, 2009.

[45] N. Hurtós, X. Cufi, and J. Salvi, "Calibration of optical camera coupled to acoustic multibeam for underwater 3D scene reconstruction," in *Proc. IEEE OCEANS*, Sydney, Australia, May 2010, pp. 1–7.

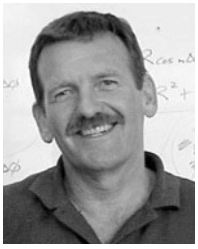
[46] M. Legg and S. Bradley, "Comparison of CLEAN-SC for 2D and 3D scanning surfaces," in *Proc. 4th Berlin Beamform. Conf.*, Berlin, Germany, Feb. 2012, pp. 1–11.

[47] M. Legg, "Microphone phased array 3D beamforming and deconvolution," Ph.D. dissertation, Dept. Phys., Univ. Auckland, Auckland, New Zealand, 2012.

[48] D. C. Brown, "Decentering distortion of lenses," *Photogram. Eng.*, vol. 32, no. 3, pp. 444–462, 1966.

[49] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. Sebastopol, CA, USA: O'Reilly Media, 2008.
[50] J.-Y. Bouguet. (2012, Mar.). *Camera Calibration Toolbox for MATLAB* [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/index.html
[51] Z. Zhang, "A flexible new technique for camera calibration," Microsoft Res., Microsoft Corporation, Tech. Rep. MSR-TR-98-71, Dec. 1998.
[52] J.-Y. Bouguet. (2012, Mar.). *Stereo Triangulation in MATLAB* [Online]. Available: http://www.multires.caltech.edu/teaching/courses/3DP/ftp/98/hw/1/triangulation.ps
[53] J.-Y. Bouguet, "Visual methods for three-dimensional modeling," Ph.D. dissertation, Dept. Electr. Eng., California Inst. Technol., Pasadena, CA, USA, May 1999.
[54] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. New York, NY, USA: Cambridge Univ. Press, 2004.
[55] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry, *An Invitation to 3D Vision: From Images to Geometric Models*. New York, NY, USA: Springer-Verlag, 2003.
[56] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*. Cambridge, MA, USA: MIT Press, 1993.
[57] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation fo time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 8, pp. 320–327, Aug. 1976.
[58] T. Wiens and S. Bradley. (2012, Mar.). *A Comparison of Time Delay Estimation Methods for Periodic Signals* [Online]. Available: http://www.nutaksas.com/papers/wiens_dsp_delay.pdf
[59] T. Wiens. (2012, Mar.). *Subsample Delay Estimation* [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/25210-subsample-delay-estimation
[60] R. P. Dougherty and G. G. Podboy, "Improved phased array imaging of a model jet," in *Proc. 15th AIAA/CEAS Aeroacoust. Conf., 30th AIAA Aeroacoust. Conf.*, Miami, FL, USA, May 2009.
[61] D. Döbler, G. Heilmann, and R. Schröder, "Investigation of the depth of field in acoustic maps and its relation between focal distance and array design," in *Proc. Inter Noise*, Shanghai, China, Oct. 2008, pp. 1–8.

Mathew Legg received the B.Sc., P.G.Dip.Sc., M.Sc., and Ph.D. degrees in physics from the University of Auckland, Auckland, New Zealand, in 2003, 2004, 2007, and 2012, respectively. His M.Sc. research related to investigating rain clutter rejection algorithms for acoustic radars. His Ph.D. research related to 3-D acoustic imaging using microphone phased array and computer vision techniques. This involves designing and building "acoustic camera" hardware and software using beam forming, deconvolution, and computer vision to generate acoustic maps on the 3-D surface of objects. He was a Teaching/Research Fellow with the Department of Physics, University of Auckland. He has recently become a Research Fellow with the Brunel Innovation Centre, Brunel University, London, U.K.



Stuart Bradley is a Professor with the Physics Department, University of Auckland, Auckland, New Zealand. He held the Chair of acoustics with the Acoustics Research Centre, University of Salford, Greater Manchester, U.K. He has been a Research Scientist with the New Zealand Meteorological Society, and CSIRO, Sydney, Australia. He designs acoustic wind profilers, which measure wind speed and direction up to a few 100 m above the ground. He has more than 35 years of experience in acoustics and remote sensing, and has been a participant and WP leader on a number of large EU projects, and on many other collaborative projects, as well as being a consultant to industry.