Design and Calibration of Large Microphone Arrays for Robotic Applications

Design and Calibration of Large Microphone Arrays for Robotic Applications

Florian Perrodin*, Janosch Nikolic*, Joël Busset* and Roland Siegwart*

***Abstract*—Hearing is amongst the most important senses a modern robot must exhibit. Perceiving the acoustic world enables capabilities such as natural interaction with humans, interpreting spoken commands or the localization of victims during search and rescue tasks. Real-world robotic operations often take place in noisy, reverberant environments while requiring features such as source separation, accurate direction of arrival estimation or high performance noise suppression. This work presents a methodology to design, calibrate and operate large microphone arrays that enable such features.**

Recent micro electro-mechanical microphones in conjunction with reconfigurable logic tackle the weight, size, power consumption and cost constraints of robotic systems. A novel, automatic array shape calibration algorithm is developed for 2D and 3D arrays to face common experimental problems such as reverberation and poor signal-to-noise ratio when calibrating the array. The special case of a 2D array calibrated using sources moving in 3D is addressed. No prior information on array geometry is required, the process is fully automated and does not require any specific calibration equipment.

The example application of an acoustic camera is presented as a proof of concept. High-quality acoustic images are computed in real-time by generalized inverse beamforming. This demonstrates the effectiveness of the proposed design and illustrates the usefulness of such sensing capabilities for various robotic applications.

I. Introduction

This work presents the design, calibration and operation of large microphone arrays for robotic applications. Size, weight, cost and power consumption are often crucial factors for any sensor in a robotic system. The proposed design is therefore based on a new generation of digital micro electro-mechanical (MEMS) microphones. A field-programmable gate array (FPGA) interfaces up to 128 such microphones and employs a cascade of filters to obtain samples representing the actual sound pressure and reducing the amount of data that is transmitted to a host system. Section III outlines the design of the array and the pre-processing stages performed on the sensor itself.

Microphone array shape calibration is often a prerequisite for array processing algorithms such as source localization or beamforming. Inaccuracies in the relative position of the elements severely degrades the performance of such methods [1].

Often, accurate knowledge about array geometry is not available and needs to be estimated. Section IV addresses this problem of recovering the microphone positions without any a priori knowledge on the array geometry or sound source locations. Calibration is fully automated, can be performed in reverberant environments and does not rely on specialized equipment. Using a continuously moving noise source, such as e.g. a hand-held speaker, time difference of arrival (TDOA) is estimated between each pair of microphones in a robust manner at low SNR and in reverberant conditions.

To give a final proof of concept and to illustrate the benefit of having such hardware and sensing capabilities at the robots disposal, Section VIII demonstrates real-time operation of a state-of-the-art method to compute acoustic images of the scene. Such and similar information can be invaluable in e.g. search and rescue scenarios.

II. Previous Work

Microphone arrays have been successfully employed on robotic systems for human-machine interaction [2], speaker detection and sound source localization [3], amongst many other applications.

In [4], an uncalibrated eight element microphone array is used in conjunction with a blind speech separation algorithm for robot audition. Individual speakers are successfully separated without the need for array calibration as blind speech separation algorithms do not require knowledge about the shape of the array. However, the locations of the individuals are not recovered.

In [5], a 64-channel microphone array is mounted on a wall, while an eight channel microphone array is embedded in the head of a robot. The system is able to detect, localize and track sources using a particle filter. While good performances are shown, the robot's audition relies on external devices that are hardly movable and need to be installed beforehand.

In [6], a 32-channel circular array is used to localize sound sources in 2D using a classical delay-and-sum beamformer. The sampling rate is limited (11kHz) and the algorithms are kept simple to allow real time sound source localization.

Shape calibration information for small arrays is often extracted from computer aided design (CAD) models or by manually measuring inter-element distances followed by e.g. a dimension reduction method [7]. Numerous methods for automatic microphone array shape calibration have been proposed. Most of these methods require special infrastructure [8] and tedious calibration processes. Others require a good initial estimate of the microphone positions [9]. One notable exception is the method recently proposed in [10], which is able to recover the geometry of the array without any loudspeaker nor a priori on the array geometry. However,

*F. Perrodin, J. Nikolic, J. Busset and R. Siegwart are with the Autonomous Systems Lab, Swiss Federal Institute of Technology {first-name.name}@mavt.ethz.ch

the diffuse noise field assumption is not easy to satisfy and the final precision may remain limited.

Most of the existing work on microphone arrays related to robotics focuses on arrays with few microphone elements. While adapted for simple DOA estimation, these systems are unable to provide acoustic images and are less sensitive. Moreover, as under certain assumptions the signal to noise ratio (SNR) of beamformers scales linearly with the number of microphone elements, these systems remain limited in their ability to reduce the background noise. Thus, large microphone arrays are of primary interest to perform acoustic imaging as well as to amplify distant sound sources.

The algorithm proposed in this work aims at a calibration process that is fully automated, requires no specialized equipment and can be performed in the normal operating (reverberant) environment of the robot.

III. Microphone Array Design

One of the most important factors that prohibits the employment of current large microphone arrays for real-world robotic applications is their size and the complexity of their (often analog) electronic front-ends. Recently, digital MEMS microphones have been introduced and are nowadays found in most cell-phones on the market. Their quality is continuously improving, and pre-amplifier, signal conditioning and analog-to-digital conversion are often integrated on a single chip. This results in very compact and lightweight designs.

However, the problem of managing a high number of input/outputs (IOs) as well as real-time signal decoding remains: standard microcontrollers and digital signal processors (DSPs) have a limited number of IOs and converting the digital output from the microphones (often encoded using pulse density modulation (PDM)) to a more practical format is computationally expensive: a standard implementation [11] includes a four-stage process (a cascaded integrator-comb (or CIC) filter followed by two half-band filters and an FIR low-pass filter) that requires at least 616 processor cycles per sample (with 8x vectorization), resulting in a 30MHz requirement only to process one single microphone at 48kHz and 16-bit width. This limits the use of low-power DSPs for large microphone arrays (40 to 100+ elements).

These two problems are solved by using programmable logic: with an FPGA, numerous IOs are available and massive parallelization makes the filtering process easy: with the proposed design it is possible to handle up to 128 microphone elements with a fraction of the total resources available on the low-cost XILINX Spartan-6 LX45 that was employed.

At last the microphone array must be able to transmit 94Mbit/s for 128 microphones at 48kHz, 16bits. This is achieved through a USB2 port which enables a maximum throughput of up to 480Mbit/s while being an interface commonly available on embedded computers employed in robotic applications. The whole process is depicted in Fig. 1. An image of the arrays and one of a dual microphone element are shown in Fig. 2.

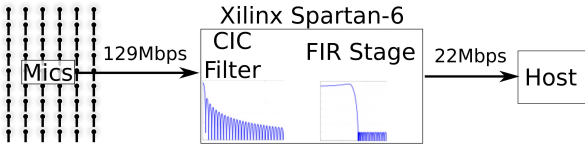
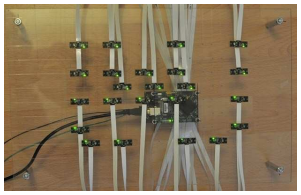
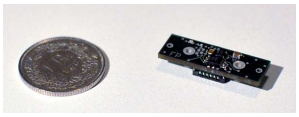


Figure 1: MEMS microphone data acquisition, preprocessing and transmission to a host system.



(a) Microphone array and FPGA processing board.



(b) Dual microphone element outputting a digital differential signal requiring no subsequent signal conditioning or analog to digital conversion.

Figure 2: Flexible 44 element microphone array prototype for robotic applications.

IV. Shape calibration

The calibration process for small microphone arrays is often simplified: all microphones are soldered on a single PCB resulting in sub-millimeter accurate knowledge of the geometric shape calibration parameters beforehand. If required these calibration parameters can be refined by employing a batch smoother. For large microphone arrays however, calibration is more difficult. The microphones are typically not mounted on a single planar PCB but follow more complex shapes to optimize performance. Also, thermal stresses can lead to bending of larger PCBs, and available a-priori knowledge of shape parameters is less accurate. Thus automatic shape calibration is necessary to achieve acceptable performance of e.g. beamforming algorithms.

We propose a shape calibration pipeline that can be applied in almost any environment: we slowly move a sound source from a distance of a few meters from the array (that is in the far-field) assuming that we do not know the position of the sounding source. Then we estimate both the geometry of the array and the position of the sources. The starting point is to find the time-difference of arrival (TDOA) between each pair of microphones and for each source position.

Shape calibration algorithms based on TDOA measurements can usually be divided into two main parts. The first part consists of estimating the TDOA between each pair of microphones, using for example generalized cross-correlation (see [12] for a comprehensive study) or adaptive eigenvalue decomposition [13]. These TDOA measurements serve as an input to the second part, where the positions of the microphones are estimated by fitting a model on the observed measurements. A classical approach consists of finding by some means all inter-element distances and to then apply the Multi-Dimensional Scaling (MDS) method to get the best p-dimensional explanation for the microphone

array geometry in a least square sense [7]. However, MDS is very sensitive to erroneous data (outliers) and often fails to recover the microphone positions in practice.

A. Affine structure from sound (ASFS)

An algebraic method to find the 2D microphone positions knowing the TDOA was presented by Thrun [14]. It has the advantage to fully exploit the structure of the problem while being computationally tractable. Here, this algorithm is presented and generalized to the p dimensional case where $p = 2$ or 3.

M incoming sounds are emitted from unknown locations $\forall j \in \{1, 2, \ldots, M\}$, $\boldsymbol{s}_j \in \mathbb{R}^p$ and N microphones are located in $\forall i \in \{1, 2, \ldots, N\}$, $\boldsymbol{m}_i \in \mathbb{R}^p$. The origin is set to $\boldsymbol{m}_1$, that is $\boldsymbol{m}_1 = \mathbf{0}$. An incoming sound j is recorded at time $\Delta_{i,j}$ by the microphone i. Note that for a given source, only the relative arrival time between microphones is used. Thus, one can arbitrarily set the time origin and $\forall j$, $\Delta_{1,j} = 0$.

If $\boldsymbol{u}_{i,j} \in \mathbb{R}^p$ is the unitary vector that uniquely defines the direction of the incoming sound source j with respect to the microphone i, the algebraic distance (denoted by $|\cdot|$) between this source and that microphone is:

$$\forall i \in [2, N], \forall j \in [1, M], |\boldsymbol{m}_i - \boldsymbol{s}_j| = \boldsymbol{u}_{i,j} \cdot (\boldsymbol{m}_i - \boldsymbol{s}_j)$$

In the far field approximation, $\forall i$, $\forall j$, $\boldsymbol{u}_{i,j} \approx \boldsymbol{u}_j$ and

$$\forall (i, j), \Delta_{i,j} = c^{-1} \boldsymbol{u}_j \cdot \boldsymbol{m}_i$$

where c is the sound velocity in air. By defining

$$\begin{aligned} \Gamma &= (\boldsymbol{u}_1 \ldots \boldsymbol{u}_M) \in \mathcal{M}_{p,M}(\mathbb{R}) \\ X &= (\boldsymbol{m}_2 \ldots \boldsymbol{m}_N)^\top \in \mathcal{M}_{N-1,p}(\mathbb{R}) \\ \Delta &= (\Delta_{i,j})_{2 \le i \le N, 1 \le j \le M} \in \mathcal{M}_{N-1,M}(\mathbb{R}) \end{aligned}$$

the previous equalities can be written in a matrix form:

$$c\Delta = X\Gamma$$

Knowing Δ, the problem of finding X and Γ can be summarized by the optimization program

$$\langle X^\star, \Gamma^\star \rangle = \underset{X,\Gamma}{\operatorname{argmin}} \|X\Gamma - c\Delta\|^2 \quad (1)$$

$$\text{s.t } \operatorname{diag}\left(\Gamma^\top \cdot \Gamma\right) = \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix} \quad (2)$$

where $\forall Y \in \mathcal{M}(\mathbb{R})$, $\|Y\|^2 = \sum_{i,j} |y_{i,j}|^2$. The constraint enforces that $\forall j$, $\|\boldsymbol{u}_j\|_2^2 = 1$. This problem is solved by a two-step minimization method. First, Δ is decomposed using Singular Value Decomposition (SVD):

$$\Delta = U \Sigma V^\top$$

where $U \in \mathcal{M}_{N-1}(\mathbb{R})$, $\Sigma \in \mathcal{M}_{N-1,M}(\mathbb{R})$ and $V \in \mathcal{M}_M(\mathbb{R})$, and one reduced this equation to its p largest components: $\Delta_r = U_r \Sigma_r V_r$ where $\Sigma_r \in \mathcal{M}_p(\mathbb{R})$, $U_r \in \mathcal{M}_{N-1,p}(\mathbb{R})$ and $V_r \in \mathcal{M}_{p,M}(\mathbb{R})$. We cannot yet set $X^\star = U_r \Sigma_r$ and $\Gamma^\star = V_r^\top$ since there is no reason that the constraint (2) is satisfied. Thus, in a second step, one introduces an invertible matrix $C \in \mathrm{GL}_p(\mathbb{R})$ and one can write

$$\forall C \in \mathrm{GL}_p(\mathbb{R}), \Delta_r = U_r \Sigma_r C^{-1} C V_r^T$$

where C can be arbitrarily chosen (provided it is invertible). Thus, using a non-linear optimization procedure, one can find:

$$C^\star = \underset{C \in \mathrm{GL}_p(\mathbb{R})}{\operatorname{argmin}} \left\| \operatorname{diag}\left(\left(CV_r^T\right)^\top \cdot \left(CV_r^T\right) \right) - \mathbf{1} \right\|_2^2$$

The C matrix that has to be found in the previous equation is a full homography. Since we cannot recover the real system coordinate, we can remove the rotations from the set of candidate matrices by imposing that C is upper triangular. Thus, the optimization procedure involves only $\frac{p(p+1)}{2}$ parameters. Then, the minimization can be performed for $C \in \mathcal{T}_p(\mathbb{R})$. One has to check afterwards that the solution is invertible or not close to singular. In the noise-free case, $C^\star$ is invertible by construction.

One finally obtains

$$X^\star = U_r \Sigma_r C^{\star -1} \text{ and } \Gamma^\star = C^\star V_r^\top$$

B. Extension for planar arrays with 3D source distribution

Two dimensional microphone arrays are very common because they are easy to build, and they avoid to deal with occlusion issues: a source emitting a sound is either heard by all the microphones (if it lies in the front of the array) or by none of them (if it lies in the back) whereas in a 3-dimensional array, the structure, if not properly designed, can prevent a source from being heard by all the microphones.

If the array is planar but the calibrating sources are moved in 3D, the rank of Δ_r will be 2 and not 3 leading to degeneracy. $\Sigma_r = \operatorname{diag}(\lambda_1, \lambda_2, 0)$ has only two non-zero values. Thus, the 3$^{\text{rd}}$ line of V_r is arbitrary which leads to an indetermination in the sources and microphones positions. To overcome this issue, we propose an extension of the previous algorithm by replacing the procedure to determine $C^\star$.

First, this problem cannot be solved without additional assumptions on the measurements: for example, a far field source lying at position $S(\alpha, \beta)$, where α is the azimuth and β is the altitude, produces exactly the same TDOAs on the array as a source lying in position $S'(\alpha', \beta')$ where $\alpha' = \alpha$ and $\beta' = \arccos(\frac{\cos(\beta)}{k})$ on the same array but scaled by a factor k. Thus, if the source distribution does not include positions with altitude under $\beta_{\min}$, there is an under-determined scale factor in the range $[\cos(\beta_{\min}), +\infty]$. In other words, one can always explain the same measured TDOAs with a smaller array sliding the estimations of the source positions toward the horizon (ie. the plan of the array). More generally, the calibration of the array can be solved up to a linear transform represented by a 2D matrix : two scale factors along two orthogonal axis, a shearing factor and a rotation. The recovery of the rotation is unnecessary in the case we are not searching for the orientation of the array. Thus three parameters are missing.

To overcome these ambiguities, one has to add some hypothesis on the placement of the sources. Let $V_{2D} \in \mathcal{M}_{2,M}(\mathbb{R})$ be the first two rows of V_r. If Γ was known, then the 2D column vectors of V_{2D} would lie on the projection of the unit sphere onto the xOy plane, that is a disc centered in 0 and of radius 1. In the case where Γ is known up to a matrix $C_{3D} \in GL_3(\mathbb{R})$, the column vectors of V_{2D} lie in an ellipse centered in zero. Finding this ellipse allows to find the 2D transformation $C^\star_{2D}$ that is needed to recover the 2D position of the microphones. Without further assumptions, this ellipse cannot be recovered, because the column vectors of V_{2D} are not enforced to touch the border of the ellipse.

A simple hypothesis to add is that some sources where positioned on the plane of the array at least in two different directions (ie. there are sources with zero-altitude for different azimuths). This ensures that at least two column vectors of V_{2D} touch the border of the ellipse at two different points. Since the ellipse is centered, this is sufficient to fit it.

This fit can be done by searching the ellipse centered at the origin, with minimal area such that all the 2D column vectors of V_{2D} lie inside it. This can be done by classical minimization techniques.

Once the ellipse is found, one wants to find the matrix $C^\star_{2D}$ – that one constrains to be upper triangular $C^\star_{2D} \in \mathcal{T}_2(\mathbb{R})$ since the rotational part is not needed – that transforms the ellipse $\mathcal{E}$ into a unit circle $\mathcal{C}$. $\mathcal{E}$ can be represented as a matrix $E \in \mathcal{M}_2(\mathbb{R})$ such that $\forall \mathbf{x} \in \mathcal{E}, \mathbf{x}^\top E \mathbf{x} = 1$, and each point $\mathbf{y}$ on $\mathcal{C}$ verifies $\mathbf{y}^\top \mathbf{y} = 1$. Thus $C^\star_{2D}$ is such that $\forall \mathbf{x} \in \mathcal{E}, (C^\star_{2D}\mathbf{x})^\top (C^\star_{2D}\mathbf{x}) = 1$, that is $\forall \mathbf{x} \in \mathcal{E}, \mathbf{x}^\top C^{\star\top}_{2D} C^\star_{2D}\, \mathbf{x} = 1$. If $\mathcal{E}$ is not degenerated, this means $E = C^{\star\top}_{2D} C^\star_{2D}$. One can show that if E is of the following form, $C^\star$ can be computed as follows.

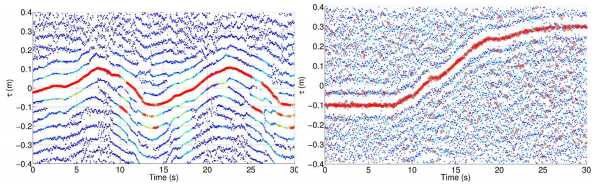
$$E = \begin{pmatrix} \alpha & \frac{\beta}{2} \\ \frac{\beta}{2} & \delta \end{pmatrix} \qquad C^\star_{2D} = \begin{pmatrix} \varepsilon_1\sqrt{\alpha} & \varepsilon_2 \frac{\beta}{2\sqrt{\alpha}} \\ 0 & \varepsilon_1\sqrt{\delta - \frac{\beta^2}{4\alpha}} \end{pmatrix}$$

where $\varepsilon_{1,2} = \pm 1$ corresponding to the two reflections across Ox and Oy.

C. TDOA estimation in reverberant environments

The method described in the previous section assumes that all TDOAs are known. In outdoor or in low reverberant conditions, this can be achieved relatively easy using maxima detection in generalized cross-correlation (GCC). However, in highly reverberant rooms this estimation can be plagued by false peaks in GCC that have a higher amplitude than the one due to the direct path. This is illustrated in an experiment (Fig. 3) where a pair of microphones (sampling frequency: 16kHz) is rotated slowly while a speaker emits white noise. Cross-correlation between a pair of microphones is computed for each window of 1024 samples and the maxima of this function are represented as a function of time. The color of the points corresponds to the maxima height. The red circles indicate the global maximum at each time, which is commonly used to compute the TDOA.

In a highly reverberant environment (Fig. 3a), the global maximum is not always the direct path TDOA even for large



(a) In a highly reverberant environment. (b) At low SNR.

Figure 3: Maxima of the cross-correlation with respect to time. The vertical axis represents the position of the peaks in meters (time multiplied by c, the sound velocity). The color shows how high a maxima is (red is the highest). The red circles highlight the highest peak at a given time.

periods of time (see $t = 10$s). To explain the other peaks, one can write at the first order:

$$\forall t,\ y_i(t) = s(t - \tau_{i,1}) + \alpha s(t - \tau_{i,2})$$

where y_i is the signal received by the microphone i, s is the signal output by the speaker, αs is the signal due to the main reflection ($0 \leq \alpha < 1$), $\tau_{i,1}$ and $\tau_{i,2}$ are the propagation time between emission and reception ($\tau_{i,2} > \tau_{i,1}$ since $\tau_{i,1}$ is the direct path). When computing the cross-correlation of a pair of microphones (i,j) (denoted by $\star$), one obtains:

$$\begin{aligned} y_i \star y_j = {} & f(\tau_{i,1} - \tau_{j,1}) + \alpha f(\tau_{i,1} - \tau_{j,2}) \\ & + \alpha f(\tau_{i,2} - \tau_{j,1}) + \alpha^2 f(\tau_{i,2} - \tau_{j,2}) \end{aligned} \tag{3}$$

where $\forall t,\ f(\tau_a - \tau_b)(t) = s(t - \tau_a) \star s(t - \tau_b)$.

Equation 3 consists of two types of terms: terms one and four that correlate two signals coming from the same spatial source (the reflection can be replaced by a coherent source symmetrical to the original source with respect to the reverberating plane). When turning the pair of microphones, the TDOA $\tau_a - \tau_b$ corresponding to this type of terms should cross zero (when the vector formed by the pair of microphones is orthogonal to the direction of the source). In particular TDOA of term one and four can cross each other.

The second type of terms (terms two and three) correlate signals coming from different coherent spatial sources. One term comes from the direct path while the other comes from the indirect path. We have $\forall j,\ \tau_{j,2} > \tau_{j,1}$, which implies $\forall (i,j),\ \tau_{i,1} - \tau_{j,1} < \tau_{i,1} - \tau_{j,2}$. Therefore the peak coming from term two is always strictly above the one coming from term one. Similarly, the one coming from term three is always strictly below the one coming from term one. Thus, this peaks coming from this type of terms can never cross the one coming from term one.

The explanation of Fig 3a is the following: the main curve of peaks corresponds to term one of 3 because it has overall the highest amplitude and it is continuous. The curve apart from the main one correspond to term 2 and 3: they never cross the main curve.

If the SNR is smaller as in Fig. 3b, the true maxima becomes smaller and results in false TDOA estimation.

D. Using continuity for TDOA estimation

To overcome the problems of the previous Section, a framework where a sound source is moving slowly is proposed. Thus, the continuity of the TDOA with respect to time is exploited to obtain robust estimates.

A sound source is considered slowly moving if its movement is not discernible within a same window of the cross-correlation. If T_S is the sampling frequency and N_W the number of samples per window $T_W = N_W T_S$. Given v_a the angular speed of the sound source, $v_a T_W \ll 1$ is desired. For $T_S = 1/16000$s, $N_W = 1024$ and a moving source at a distance of $D = 3$m from the array, the linear speed v of the source should be negligible compared to $46\text{m} \cdot \text{s}^{-1}$, for example $|v| < 50\text{cm} \cdot \text{s}^{-1}$ which is not restrictive. Equivalently, one can use a fixed source while moving array slowly ($|v_a| \ll 16\text{rad} \cdot \text{s}^{-1}$).

Additionally, the sound source is supposed to be moved all around the array. This is especially important in degenerate cases (e.g. a planar array in a 3D space).

The method is the following: first, cross-correlations are computed and local peaks extracted. Then Dijkstra's algorithm ([15]) is applied and TDOAs are estimated. An optional dimension reduction step can be performed. Finally, the geometry of the microphone array is recovered using the ASFS algorithm.

V. Continuity enforcement

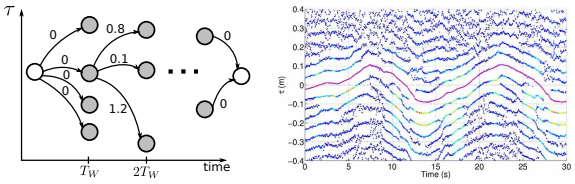
The idea of using continuity comes from the fact that peaks due to reverberation do not lead to a continuous line because the reverberating material has a limited spatial extent: to receive the reverberated signal, the array should be in a cone whose vertex is the virtual source and whose base is the surface of the material.

The shortest path algorithm proposed by Dijkstra is described in [15]. It finds the shortest path between two given vertices of a connected graph. Here, the vertices are the maxima of the cross-correlation and the edges are defined as shown in Fig. 4a: For each time-step, a directed edge is added between each peak of the current time and each peak of the next time-step. The weight of this edge depends on the difference between the two TDOAs $\tau_{t,i} - \tau_{t+T_W,j}$. The squared value of this difference can for example be used. Edges can be added to allow to skip one time-step: in very low SNR conditions, some peaks due to the direct path could disappear and the algorithm would fail. Note that the amplitude of the maxima is not used here, but could be taken into account. Figure 4b shows the result of Dijkstra's algorithm in the reverberant case.

VI. Dimension reduction

For each pair of microphones (i, j) a time-sequence $\left(\tilde{\Delta}_{i,j,t}\right)_{1 \leq t \leq M}$ of TDOAs is recovered. Thus, more information is available than required by the ASFS method which requires a two dimensional array of measurements. However, the information in $\tilde{\Delta}$ is redundant. At each time-step t, the ASFS algorithm needs $N-1$ TDOAs relative to the first microphone while the full computation provides the $N-1$ required TDOAs and the linear combination of these TDOAs. It is either possible to compute only $N-1$ TDOAs (e.g for all microphone pairs$(1, j)$) to save computational power, or one can process all pairs and apply a dimension reduction algorithm.

For a finite cross-correlation window length T_W, all measurements $\tilde{\Delta}_{i,j}$ have a limited precision, thus it can be worth to combine them to reduce the error. MDS is a classical method to infer relative positions given a distance matrix like $\tilde{\Delta}_{i,j}$. Since we are looking for the time sequence, the final dimension will be one.



(a) Input graph of the algorithm. (b) Result of the Dijkstra algorithm.

Figure 4: The Dijkstra's algorithm can be used to enforce continuity on the TDOAs.

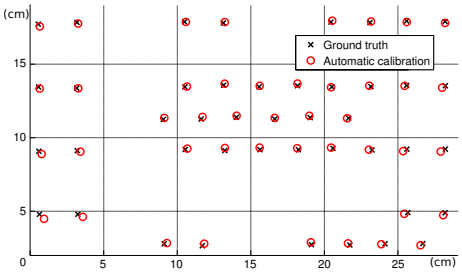


Figure 5: Microphone element positions after automatic calibration (ground truth and output of the algorithm). The mean error is 1.6mm

VII. Calibration experiments

Experiments were conducted on a 44 element planar microphone array sampled at 16kHz in a room where $\text{RT}_{60} = 0.5$s. The number of samples per cross-correlation window is $N_W = 1024$ with an overlap of 512 samples. The sound source is fixed in the far-field while the array is rotated slowly in all directions. It is ensured that the sound source crosses the plane of the array twice in two approximately orthogonal directions. Results are shown in Fig. 5. The mean error is under 2mm.

VIII. Example: real-time acoustic imaging

Sound can be a very rich source of information and may aid robots to accomplish their goals in a wide variety of tasks. Victims in a search and rescue operation might for instance be invisible to normal cameras, while acoustic sensor arrays are able to detect and accurately localize individuals. To demonstrate the effectiveness of the proposed design we implemented a real-time acoustic camera based on generalized inverse beamforming [16] as an example.

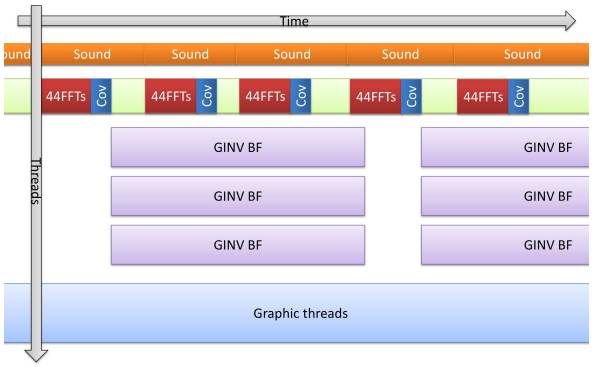


Figure 6: Thread overview of the real-time acoustic camera application: whenever a block of sound samples is received, an FFT is computed for each microphone and the spatial covariance matrix is computed. Then generalized inverse beamformers are computed at the desired set of frequencies. Acoustic images are generated with rates exceeding 60fps.

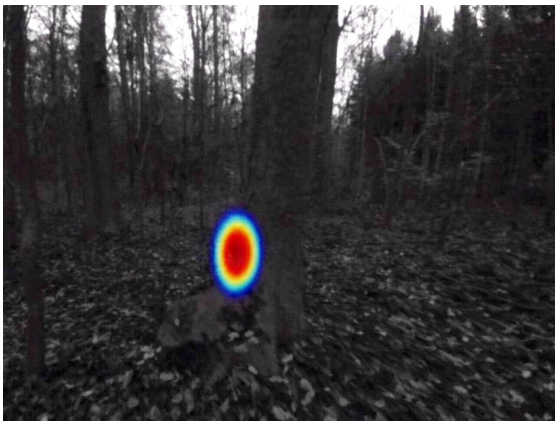


Figure 7: Acoustic image superimposed on camera image. Color indicates sound pressure level (red is maximum, transparent indicates pressure is below a fixed threshold). The person shouting behind the tree can be easily detected.

Using $\mathcal{L}_2$ norm minimization, our implementation of the generalized inverse beamformer reaches 60fps on a standard laptop computer. The program is multi-threaded to enable simultaneous imaging of multiple acoustic frequency bands. An overview of the threading is shown in Fig. 6.

The setup was tested under real conditions to search for a person hidden behind a tree in a forest shouting for help. A video camera was mounted in the center of the microphone array and its pose with respect to the output of the generalized inverse beamformer calibrated. Figure 7 as well as the included supplementary video show an acoustic image superimposed on the camera image. Color indicates sound pressure level and clearly marks the person shouting for help.

IX. Conclusion

The microphone array design presented in this work fulfills the specific requirements that arise for large arrays in robotic applications. Lightweight and low-cost MEMS digital microphones in conjunction with pre-processing performed on an FPGA enables real-time operation of different array processing algorithms.

An automatic shape calibration method was presented that allows for quick array calibration on-site and without the need for special equipment. This gives the robot designer more flexibility in microphone placement and allows to seamlessly integrate the array. The calibration algorithm was tested on a 44 element array and gives good results.

A real-time acoustic camera algorithm was implemented as a case study of the large microphone array on robotic platforms and demonstrated the feasibility of running complex sound processing algorithms on embedded platforms.

References

[1] P. M. Schultheiss, "Optimum range and bearing estimation with randomly perturbed arrays," *The Journal of the Acoustical Society of America*, vol. 68, no. 1, p. 167, 1980.

[2] K. Nakadai, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition HARK and its evaluation," *Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots*, pp. 561–566, 2008.

[3] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, pp. 216–228, Mar. 2007.

[4] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2123–2128, 2004.

[5] K. Nakadai, H. Nakajima, and M. Murase, "Real-Time Tracking of Multiple Sound Sources by Integration of In-Room and Robot-Embedded Microphone Arrays," *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.

[6] Y. Tamai, S. Kagami, Y. Amemiya, Y. Sasaki, H. Mizoguchi, and T. Takano, "Circular microphone array for robot's audition," in *Sensors, 2004. Proceedings of IEEE*, pp. 565–570, IEEE, 2004.

[7] S. Birchfield, "Geometric microphone array calibration by multidimensional scaling," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, vol. 5, pp. 157–160, 2003.

[8] V. Raykar, I. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 70–83, Jan. 2005.

[9] Y. Rockah and P. Schultheiss, "Array shape calibration using sources in unknown locations–Part II: Near-field sources and estimator implementation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, pp. 724–735, June 1987.

[10] I. McCowan, M. Lincoln, and I. Himawan, "Microphone Array Shape Calibration in Diffuse Noise Fields," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 666–670, Mar. 2008.

[11] N. Hegde, "Seamlessly Interfacing MEMS Microphones with Blackfin Processors," *EE-350 Engineer-to-Engineer Note*, 2010.

[12] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, 1976.

[13] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, pp. 384–91, Jan. 2000.

[14] S. Thrun, "Affine structure from sound," *Advances in Neural Information Processing Systems*, vol. 18, p. 1353, 2006.

[15] E. Dijkstra, *A short introduction to the art of programming*. Technische Hogeschool Eindhoven, 1971.

[16] T. Suzuki, "L1 generalized inverse beam-forming algorithm resolving coherent/incoherent, distributed and multipole sources," *Journal of Sound and Vibration*, vol. 330, pp. 5835–5851, Nov. 2011.