2019

Object Position Labelling in Video Using PRBS Audio Multilateration

Padraic McEvoy
Technological University Dublin, padraic.mcevoy@tudublin.ie

Paul Leamy
Technological University Dublin, paul.leamy@tudublin.ie

Damon Berry
Technological University Dublin, damon.berry@tudublin.ie

See next page for additional authors

Recommended Citation

McEvoy, P., Leamy, P., Berry, D., Dorran, D. & Burke, T. (2019). Object position labelling in video using PRBS audio multilteration. *IMVIP 2019: Irish Machine Vision & Image Processing*, Technological University Dublin, Dublin, Ireland, August 28-30. doi: 10.21427/h8ve-j137

Object Position Labelling in Video using PRBS Audio Multilateration

Pádraic McEvoy, Paul Leamy, Damon Berry, David Dorran, Ted Burke

Biomedical Research Group, Technological University Dublin

Abstract

Supervised machine learning approaches for tracking objects' positions in video typically require a large set of images in which the positions are labelled. Human labelling is time-consuming and automatic position labelling using visual markers is generally not possible because visible markers would corrupt the data. Here, we present an approach in which an object is tracked using a hidden tag that emits a PRBS audio signal. Four microphones arranged in a planar cross formation capture parallel recordings of the PRBS signal. Multilateration, using the time difference of arrival (TDoA) of the PRBS at each microphone, is used to estimate the position of the emitter. Here, we describe and evaluate the method by which the TDoAs are obtained and the emitter position is calculated. When evaluated, the approach yielded three-dimensional position estimates with a mean error of 18.56 cm. In its present form, the method is suitable for applications in which precision is not a priority, but three-dimensional object coordinates are required rather than two-dimensional camera view coordinates.

Keywords: Multilateration, PRBS, TDoA, Position Tracking, Machine Vision

1 Introduction

Manual annotation of video is a time-consuming task which nevertheless is required for many machine vision applications [Ayache and Quénot, 2008], including generation of training and validation datasets for machine learning [Torabi et al., 2015]. Video annotation can focus on many aspects, including speech recognition, human activity, object location and trajectories [Katsaggelos et al., 2015]. As a result of its wide applicability, object localisation in video and audio has been a subject of research for many years [Strobel et al., 2001]. We propose a method for indirect audio source localisation [Strobel et al., 2001], which can be used to automatically label video for the purpose of object tracking and position annotation within video data sets. The method is intended to provide automatic "ground truth" estimates of the positions of significant objects so that the video can be labelled to support supervised machine learning [Kwak et al., 2015] using video datasets.

2 Related work

Various approaches to localisation in video rely on image-only localisation, with emphasis on mechanisms such as object colour matching, spatio-temporal action tubes and bound boxes [Kwak et al., 2015], to aid the localisation process. The increased availability of smart speakers with their miniaturised far field microphone arrays alongside the increasing quality of smartphone camera systems, offers potential for inexpensive and ubiquitous multi-modal audio and video localisation.

A number of approaches for audio-visual fusion of multi-modal data are found in the literature [Ayache and Quénot, 2008], including mixing of audio and video localisation data. Where audio localisation is employed, the use of microphone arrays and some form of multilateration is common [Gustafsson and Gunnarsson, 2003], [Nishida et al., 2003], [Raykar et al., 2003]. For localisation of an audio source in a three-dimensional space, an arrangement of multiple microphones in a plane can be employed to receive and localise the sound. The

microphone array detects the incoming sound and the sensor system typically then uses one of two approaches. The first approach, time of arrival (ToA) [Raykar et al., 2003], involves synchronisation and accurate measurement and communication of the time of transmission (ToT) of the sound. The second approach, time difference of arrival (TDoA) [Gustafsson and Gunnarsson, 2003], is an asynchronous approach which relies only on the delays between the arrival of transmitted sound at the various microphones in the receiving array. When using TDoA, the delays in sound arrival times between receivers can be used alongside knowledge of the geometry and spacing of the receiver array, to establish the approximate location of an object from which the sound was emitted [Gustafsson and Gunnarsson, 2003]. Speakers and microphones in the audible [Gustafsson and Gunnarsson, 2003], or ultrasonic range [Nishida et al., 2003] can be used for this purpose.

A pseudorandom binary sequence (PRBS) is a binary sequence with properties similar to white noise. The PRBS used in our system is a so-called *maximum length sequence* (MLS). Such a sequence can be efficiently generated, for example using a microcontroller to mimic a linear feedback shift register (LFSR). The autocorrelation of a PRBS is a delta function at lag 0, with a small offset at all other points [MacWilliams and Sloane, 1976]. The properties of a PRBS makes it a reproducible and frequency-rich source signal for localisation.

3 Theory

In this section, we describe how the position of a PRBS audio emitter can be identified using audio signals recorded by four microphones arranged in a planar cross formation. The speed of sound, c, is $343\,\mathrm{m\,s^{-1}}$. The sampling frequency, f_s, is 48 kHz. The sampling period, $T_s = 1/f_s = 20.833\,\mu\mathrm{s}$. The positions of the PRBS emitter and the four microphones are specified in a right-handed Cartesian coordinate system. The four microphones are arranged in a cross formation, with two on the x-axis and two on the y-axis. As compared with four arbitrary locations, this configuration simplifies the geometric analysis considerably. The microphone positions used in our present configuration are

$$\vec{\mu}_1 = \begin{bmatrix} x_{\mu 1} \\ y_{\mu 1} \\ z_{\mu 1} \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.0 \\ 0.0 \end{bmatrix}, \quad \vec{\mu}_2 = \begin{bmatrix} x_{\mu 2} \\ y_{\mu 2} \\ z_{\mu 2} \end{bmatrix} = \begin{bmatrix} -1.0 \\ 0.0 \\ 0.0 \end{bmatrix}, \quad \vec{\mu}_3 = \begin{bmatrix} x_{\mu 3} \\ y_{\mu 3} \\ z_{\mu 3} \end{bmatrix} = \begin{bmatrix} 0.0 \\ 0.5 \\ 0.0 \end{bmatrix}, \quad \vec{\mu}_4 = \begin{bmatrix} x_{\mu 4} \\ y_{\mu 4} \\ z_{\mu 4} \end{bmatrix} = \begin{bmatrix} 0.0 \\ -0.5 \\ 0.0 \end{bmatrix} \tag{1}$$

where all coordinates are specified in metres. The loudspeaker emitting the PRBS signal is located at point $\vec{p}$.

$$\vec{p} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{2}$$

where it is assumed that $z > 0$ (i.e. that the PRBS emitter remains on one side of the microphone array at all times). The distance of each microphone from the emitter is

$$R_1 = \left\| \vec{p} - \vec{\mu}_1 \right\|, \qquad R_2 = \left\| \vec{p} - \vec{\mu}_2 \right\|, \qquad R_3 = \left\| \vec{p} - \vec{\mu}_3 \right\|, \qquad R_4 = \left\| \vec{p} - \vec{\mu}_4 \right\| \tag{3}$$

Initially, the microphone positions, $\vec{\mu}_1$, $\vec{\mu}_2$, $\vec{\mu}_3$ and $\vec{\mu}_4$, are known but $\vec{p}$ is unknown. Hence, R_1, R_2, R_3 and R_4 are also unknown initially. We define the propagation time from the emitter to each microphone as

$$\tau_1 = \frac{R_1}{c}, \qquad \tau_2 = \frac{R_2}{c}, \qquad \tau_3 = \frac{R_3}{c}, \qquad \tau_4 = \frac{R_4}{c} \tag{4}$$

The type of PRBS signal emitted by the loudspeaker is a maximum length sequence (MLS) generated using an 8-bit linear feedback shift register (LFSR). The size in bits of the LFSR determines the length of the generated MLS. An m-bit LFSR yields an MLS of length $2^m - 1$, the sequence $MLS_m(n)$. In our system, an 8-bit LFSR is used, resulting in a sequence $MLS_8(n)$ that is 255 bits long. The LFSR clock frequency ($f_{LFSR} = 16\,\mathrm{kHz}$) is one third that of the main system clock ($f_s = 48\,\mathrm{kHz}$). This is equivalent to assigning a

bit-repeat factor $\Delta = 3$ to the emitted PRBS, increasing its length by a factor of three. This extended PRBS is $b(n)$.

$$b(n) = MLS_8\left(\left\lfloor \frac{n}{3} \right\rfloor\right) \tag{5}$$

Therefore, the period in bits of the transmitted PRBS (at the main system clock frequency) is

$$N = \frac{f_s}{f_{LFSR}} \times (2^n - 1) = 765 \text{ bits} \tag{6}$$

The period in milliseconds of the same transmitted PRBS is

$$T_{prbs} = \frac{2^n - 1}{f_{LFSR}} = NT_s = 15.938\,\text{ms} \tag{7}$$

We define the propagation time from the emitter to the nearest microphone as

$$\tau_0 = \min\{\tau_1, \tau_2, \tau_3, \tau_4\} \tag{8}$$

The time difference of arrival (TDoA) for each microphone is defined in discrete-time units as

$$l'_1 = \tfrac{\tau_1 - \tau_0}{T_s} \qquad l'_2 = \tfrac{\tau_2 - \tau_0}{T_s} \qquad l'_3 = \tfrac{\tau_3 - \tau_0}{T_s} \qquad l'_4 = \tfrac{\tau_4 - \tau_0}{T_s} \tag{9}$$

The letter l is used for each TDoA because, in the course of the calculation, it is obtained in the form of a *lag* value of a peak in a cross-correlation function.

The audio signals from the microphones are digitised using a four-channel analog-to-digital converter (ADC), yielding four discrete-time 16-bit audio signals $m_1(n)$, $m_2(n)$, $m_3(n)$ and $m_4(n)$. Each of these signals contains a time-shifted and attenuated version of the emitted PRBS, combined with other sounds that occurred during recording. To calculate the TDoAs, l'_1, l'_2, l'_3 and l'_4, a window of length N is extracted from each of the four discrete-time audio signals, $m_1(n)$, $m_2(n)$, $m_3(n)$ and $m_4(n)$. For simplicity, it is assumed here that the window begins at $n = 0$ and ends at $n = N - 1$. To estimate the relative difference in propagation time for each microphone, each signal $m_i(n)$ is circularly cross-correlated with a pristine model of the PRBS, $b(n)$. The four resulting circular cross-correlation signals, $g_{m_1 b}(l)$, $g_{m_2 b}(l)$, $g_{m_3 b}(l)$ and $g_{m_4 b}(l)$, are defined as follows.

$$g_{m_i b}(l) = \sum_{n=0}^{N} m_i(k)\, b(n) \tag{10}$$

where $k = (n + l) \bmod N$ and $i \in \{1, 2, 3, 4\}$. Provided that the amplitude of the time-delayed PRBS in the audio signal is sufficient, we will observe a prominent peak in $g_{m_i b}(l)$ at a lag value l'_i that depends on τ_i (the propagation time for mic i) as well as on the arbitrary phase difference between the emitter's PRBS cycle and that of the model signal. The emitted and model PRBS signals have the same bit rate and period and are essentially scaled and time-shifted versions of each other, but the time difference between them is not known. Little, if anything, can therefore be gleaned about the distance from emitter to microphone using the lag value of the peak in one cross-correlation function in isolation. However, by identifying the lag value of the dominant peak in *all four* cross-correlation signals, the TDoA for each microphone can be recovered. First, we find the lag value (in samples) of the maximum magnitude value in each cross-correlation signal, as follows.

$$l_i = \arg\max_l g_{m_i b}(l) \tag{11}$$

By circularly shifting (modulo N) the four lags (l_1, l_2, l_3 and l_4) by the correct offset we can preserve the time differences between them, but arrange them so that one is equal to zero and the others are as close as possible to zero. Each of these shifted lags is basically the desired discrete-time TDoA for the corresponding microphone. The offset by which the four lags should be circularly shifted (modulo N) is

$$l_0 = \arg\min_l \left(\sum_{i=1}^{4} ((l_i + l) \bmod N) \right) \tag{12}$$

The four shifted lags are then defined as follows.

$$l'_i = (l_i + l_0) \bmod N \quad \text{for } i \in \{1,2,3,4\} \tag{13}$$

Each microphone's shifted lag is the difference in propagation time (measured in samples) from the emitter to that microphone, as compared with the microphone nearest to the emitter which has zero shifted lag. Multiplying a shifted lag by the sampling period, T_s, and the speed of sound, c converts it to a difference in *distance*.

$$r_i = l'_i \times T_s \times c \quad \text{for } i \in \{1,2,3,4\} \tag{14}$$

Suppose that the distance from the emitter to the nearest of the four microphones is r_0. The distance from the emitter to any one of the four microphones can be written as

$$R_i = r_i + r_0 \quad , \quad i \in \{1,2,3,4\} \tag{15}$$

If the true distances R_1, R_2, R_3 and R_4 were known, then four spheres S_1, S_2, S_3 and S_4 could be constructed, each sphere S_i having radius $R_i = r_i + r_0$ and centred on $\vec{\mu}_i$, which would all intersect at only one point in the $z > 0$ region: $\vec{p}$, the location of the emitter.

The key to locating $\vec{p}$ is to find the value of r_0, which we achieve using an iterative approach. Initially, we choose an upper and lower bound for r_0,as described at the end of this section, which we call r_{high} and r_{low} respectively. Over the course of the iterative process, these bounds will move closer and closer to each other until the difference between them is less than a specified tolerance, at which point iteration ceases and the estimate of $\vec{p}$ is obtained.

We divide the spheres into two pairs: S_1 and S_2, which are centred on the x-axis; and S_3 and S_4, which are centred on the y-axis. The intersection of S_1 and S_2 is a circle of radius r_{12} which lies in the plane $x = x_{12}$. Consider the point $\vec{q}$ (shown in Figure 1), which lies on the circle of intersection of S_1 and S_2 and is the only point that satisfies the following requirements.

$$\vec{q} = \begin{bmatrix} x_q \\ y_q \\ z_q \end{bmatrix} \quad \text{where} \quad \|\vec{q} - \vec{\mu}_1\| = R_1 \;\; , \;\; \|\vec{q} - \vec{\mu}_2\| = R_2 \;\; , \;\; y_q = 0 \;\; \text{and} \;\; z_q > 0 \tag{16}$$

Picturing the x-y plane as horizontal, with the z-axis pointing vertically upwards, $\vec{q}$ is the uppermost point on the circle of intersection between S_1 and S_2. Radius r_{12} is found by considering the triangle formed between points $\vec{\mu}_1$, $\vec{\mu}_2$ and $\vec{q}$, shown in Figure 1. The lengths of two sides of the triangle are R_1 and R_2. The third side of the triangle lies on the x-axis. Its length is

$$d_{12} = \|\vec{\mu}_1 - \vec{\mu}_2\| \tag{17}$$

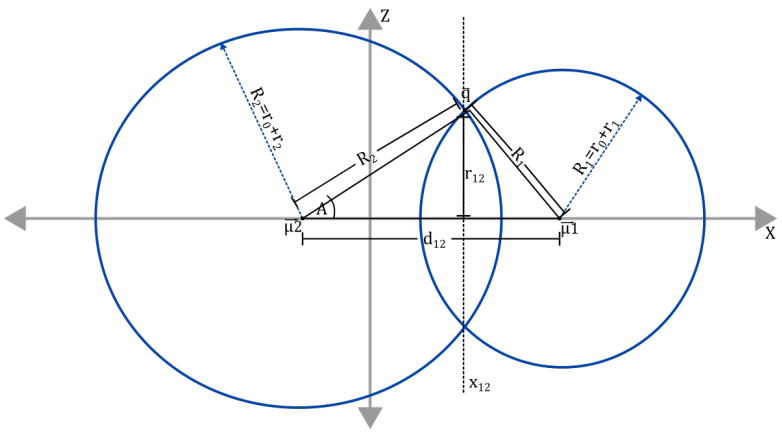


Figure 1: Intersection of spheres S_1 and S_2.

Applying the cosine rule to this triangle,

$$R_1^2 = R_2^2 + d_{12}^2 - 2R_2 d_{12}\cos A \tag{18}$$

$$\begin{aligned} \cos A &= \frac{d_{12}^2 + R_2^2 - R_1^2}{2R_2 d_{12}} & \sin A &= \sqrt{1-\cos^2 A} \\ x_{12} &= x_{\mu 2} + R_2 \cos A & r_{12} &= R_2 \sin A \end{aligned} \tag{19}$$

where $x_{\mu 2}$ is as defined in Equation 1. Similarly, the intersection of spheres S_3 and S_4 is a circle of radius r_{34} which lies in the plane $y = y_{34}$. Applying a similar analysis to that described for S_1 and S_2 yields the following.

$$d_{34} = \left\| \vec{\mu}_3 - \vec{\mu}_4 \right\| \tag{20}$$

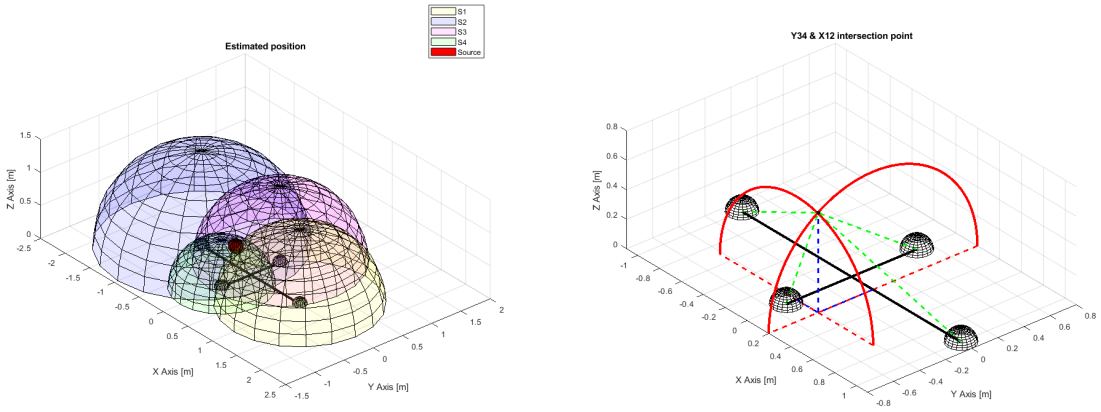
$$\begin{aligned} \cos B &= \frac{d_{34}^2 + R_4^2 - R_3^2}{2R_4 d_{34}} & \sin B &= \sqrt{1-\cos^2 B} \\ y_{34} &= y_{\mu 4} + R_4 \cos B & r_{34} &= R_4 \sin B \end{aligned} \tag{21}$$

If the correct values of R_1, R_2, R_3 and R_4 were used, the resulting values of x_{12} and y_{34} would be the true x and y coordinates of the emitter. Initially however, only r_1, r_2, r_3 and r_4 are known. Given r_0', a candidate value for r_0, estimates for the sphere radii can be calculated as

$$R_i' = r_0' + r_i \quad \text{for } i \in \{1,2,3,4\} \tag{22}$$

Using these values, estimates for r_{12}, r_{34}, x_{12} and y_{34} can be calculated using the above formulae. The circle of intersection between S_1 and S_2 intersects the plane $y = y_{34}$ at two points, of which only one has a positive z coordinate. We define the z-coordinate of this point as z_{12}. Similarly, the circle of intersection between S_3 and S_4 intersects the plane $x = x_{12}$ at two points, of which only one has a positive z coordinate. We define the z-coordinate of this point as z_{34}. Applying Pythagoras,

$$z_{12} = \sqrt{r_{12}^2 - y_{34}^2} \qquad \text{and} \qquad z_{34} = \sqrt{r_{34}^2 - x_{12}^2} \tag{23}$$



(a) S_1,S_2,S_3 and S_4 intersection point estimating the position of the sound source.

(b) x_{12} and y_{34} circles intersection point

Figure 2: Sound source location estimation using spherical surface intersections

For a given value of r_0', if z_{12} and z_{34} are calculated and found to be equal, then $r_0' = r_0$. Given upper and lower bounds for r_0 and treating z_{12} and z_{34} as functions of the candidate radius (i.e. $z_{12}(r)$ and $z_{34}(r)$), the following iterative process is used to refine the upper and lower bounds and home in on the true value of r_0.

1. Select upper and lower bounds, r_{high} and r_{low}, ensuring that $r_{high} > r_0$ and $r_{low} < r_0$.
2. Set $r'_0 = \frac{r_{high}+r_{low}}{2}$.
3. If $\text{sgn}\,(z_{12}(r'_0) - z_{34}(r'_0)) = \text{sgn}\,(z_{12}(r_{high}) - z_{34}(r_{high}))$ then set $r_{high} = r'_0$.
4. Otherwise, set $r_{low} = r'_0$.
5. If $|r_{high} - r_{low}| \geq tolerance$ then goto step 2. Otherwise, cease iteration.

When the iteration ceases, the x, y and z coordinates of $\vec{p}$ can be obtained as follows.

$$x = x_{12}, \quad y = y_{34}, \quad z = z_{12} \tag{24}$$

A good initial value for r_{high} is the greatest distance from any microphone to any point in the search area. A good initial value for r_{low} is the minimum value of r'_0 that guarantees that S_1 intersects S_2 and S_3 intersects S_4.

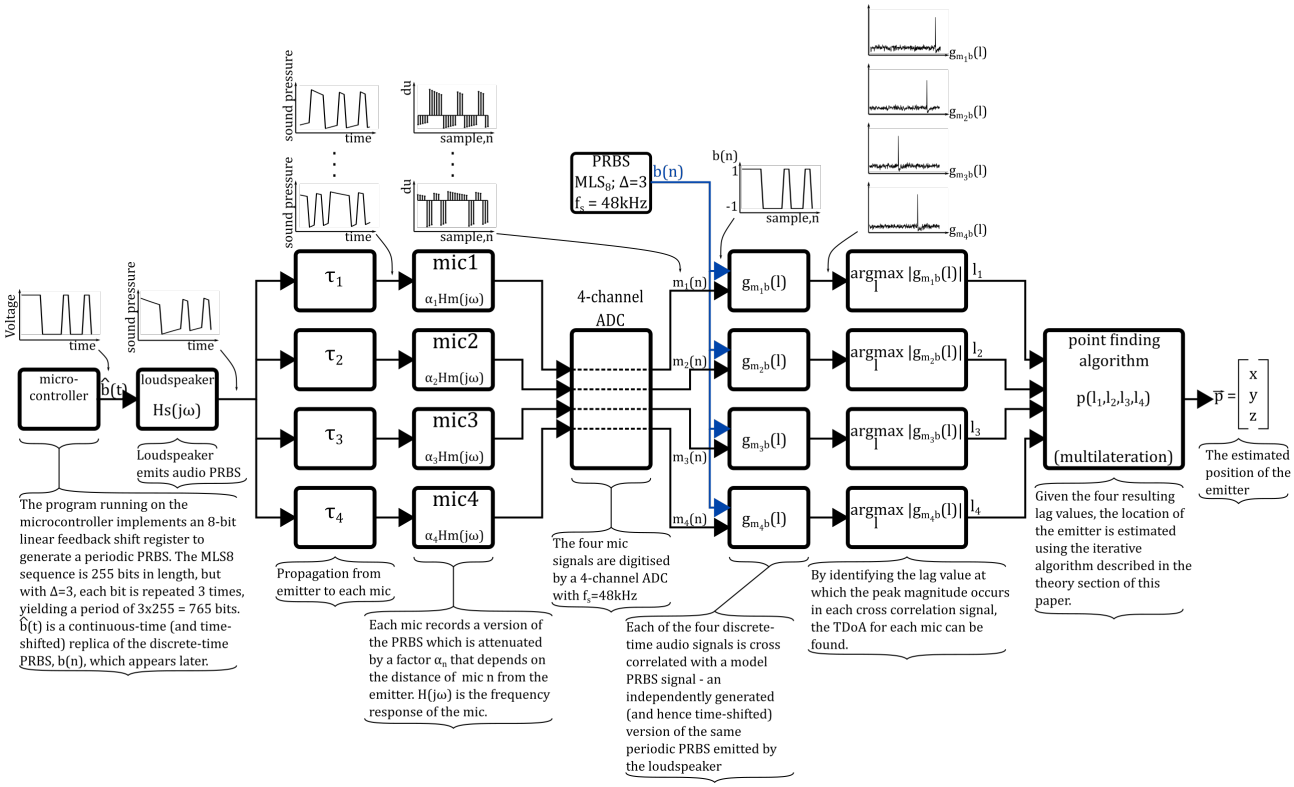
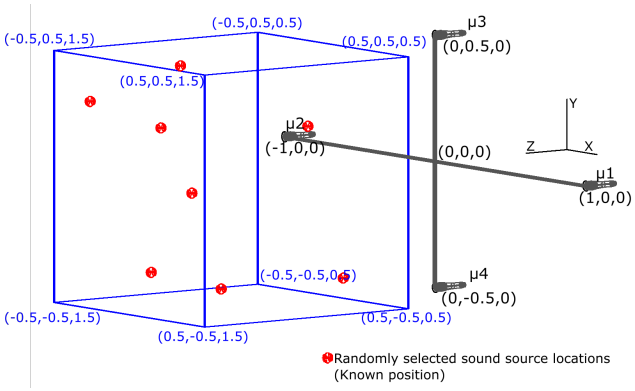


Figure 3: Block diagram of position measurement system

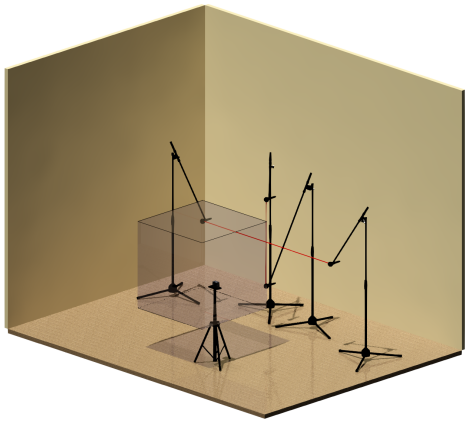
4 Experimental Setup

The microphone array and associated data acquisition system used in this investigation comprises a *ZOOM H5* four-track portable audio recorder with four condenser microphones attached. The microphones were positioned in a planar array in a cross pattern as described in Equation 1. An *Arduino Nano* micro-controller, with a timer interrupt was used to output the PRBS signal explained in Equation 5 to a small speaker. Sixteen sound source (speaker) positions were selected to gather audio data. A hypothetical cube of 1x1x1 meters was imagined as the position constraints, placed 0.5 meters from the planar array on the z-axis and centred with the origin point on the other two axis. A tripod was used to position the speaker at each of the eight vertices of the cube and at another eight randomly selected known locations. Figure 4a illustrates the concept of using the

cross planar microphone array and the sound source position constraints, while Figure 4b represents how the experiment's equipment was set up for data capture.



(a) Data capture concept



(b) Positioning of microphone array and speaker

Figure 4: Planar microphone array setup

The recorder acquired four synchronised audio files at a sampling frequency (f_s) of 48kHz for each speaker position. For each of the sound source positions, circular cross correlation was carried out between each microphone recording, and a generated PRBS within a window length N. The correlation peaks were used to acquire the TDoA of the PRBS signal to each of the microphones. The first ToA was considered to be zero, and the TDoA or lags were shifted as described in Equation 12. Figure 3 outlines this method of data capture and processing. The determined TDoA for each microphone was then used in the iterative algorithm dicussed in Equations 15 - 24, to estimate the position (x, y, z coordinate) relative to the origin point.

5 Results and Analysis

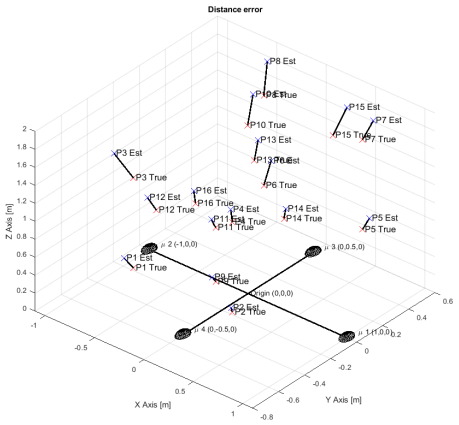


Figure 5: Error for known positions

Point number	Known distance [m]	Estimated distance [m]
1	0.8660	0.9952
2	0.8660	0.9590
3	1.6583	2.0242
4	1.6583	1.8971
5	0.8660	0.9530
6	0.8660	1.0441
7	1.6583	1.8577
8	1.6583	1.9628
9	0.7483	0.8228
10	1.4866	1.7620
11	1.2806	1.4095
12	1.2369	1.4135
13	1.2369	1.4236
14	0.7348	0.8376
15	1.500	1.7753
16	0.9899	1.1435
	Mean Error [%]	14.92

Table 1: Distance error

Figure 5 shows the position estimations compared to true values. In each case the estimated distance between the origin point and the speaker is greater than that of the true value. Table 1 presents the estimated and true value distance for each selected point to the origin, with a mean error of 14.92%. Inaccuracies arising from the calculation of the z coordinate for each point were significantly larger than the other two axis. The

use of the planar microphone array was selected to reduce the complexity of the source position estimation calculations compared to an array of microphones in arbitrary, known locations. Further investigation is needed, to determine more appropriate microphone array positioning, with the goal of reducing these inaccuracies.

6 Conclusion

We propose a multilateration approach to object position estimation, using a cross planar microphone array. An iterative technique is used to estimate the intersection point between four spherical surfaces, each centred on one of the four microphones in the array. The experimental results presented show inaccuracies, but the pattern identified warrants further investigation into the positioning of the microphones within the array. Upon determining the optimal positions within an array accompanied with the availability of commercial products such as smart speakers and phones, a system applying the described technique could be applied to a wide variety of annotated data acquisition use-cases as well as numerous other applications.

References

[Ayache and Quénot, 2008] Ayache, S. and Quénot, G. (2008). Video corpus annotation using active learning. In *European Conference on Information Retrieval*, pages 187–198. Springer.

[Gustafsson and Gunnarsson, 2003] Gustafsson, F. and Gunnarsson, F. (2003). Positioning using time-difference of arrival measurements. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 6, pages VI–553. IEEE.

[Katsaggelos et al., 2015] Katsaggelos, A. K., Bahaadini, S., and Molina, R. (2015). Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE*, 103(9):1635–1653.

[Kwak et al., 2015] Kwak, S., Cho, M., Laptev, I., Ponce, J., and Schmid, C. (2015). Unsupervised object discovery and tracking in video collections. In *Proc of IEEE intl conf on computer vision*, pages 3173–3181.

[MacWilliams and Sloane, 1976] MacWilliams, F. J. and Sloane, N. J. (1976). Pseudo-random sequences and arrays. *Proceedings of the IEEE*, 64(12):1715–1729.

[Nishida et al., 2003] Nishida, Y., Aizawa, H., Hori, T., Hoffman, N. H., Kanade, T., and Kakikura, M. (2003). 3d ultrasonic tagging system for observing human activity. In *Proceedings 2003 IEEE/RSJ Intl. Conf. on Intelligent Robots Systems (IROS 2003)(Cat. No. 03CH37453)*, volume 1, pages 785–791. IEEE.

[Raykar et al., 2003] Raykar, V. C., Kozintsev, I., and Lienhart, R. (2003). Position calibration of audio sensors and actuators in a distributed computing platform. In *Proc ACM intl conf Multimedia*, pages 572–581. ACM.

[Strobel et al., 2001] Strobel, N., Spors, S., and Rabenstein, R. (2001). Joint audio-video object localization and tracking. *IEEE signal processing magazine*, 18(1):22–31.

[Torabi et al., 2015] Torabi, A., Pal, C., Larochelle, H., and Courville, A. (2015). Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*.