Simultaneous asynchronous microphone array calibration and sound source localisation

Daobilige Su, Teresa Vidal-Calleja and Jaime Valls Miro

***Abstract*— In this paper, an approach for sound source localisation together with the calibration of an asynchronous microphone array is proposed to be solved simultaneously. A graph-based Simultaneous Localisation and Mapping (SLAM) method is used for this purpose. Traditional sound source localisation using a microphone array has two main requirements. Firstly, geometrical information of microphone array is needed. Secondly, a multichannel analog-to-digital converter is necessary to obtain synchronous readings of the audio signal. Recent works aim at releasing these two requirements by estimating the time offset between each pair of microphones. However, it was assumed that the clock timing in each microphone sound card is exactly the same, which requires the clocks in the sound cards to be identically manufactured. A methodology is hereby proposed to calibrate an asynchronous microphone array using a graph-based optimisation method borrowed from the SLAM literature, effectively estimating the array geometry, time offset and clock difference/drift rate of each microphone together with the sound source locations. Simulation and experimental results are presented, which prove the effectiveness of the proposed methodology in achieving accurate estimates of the microphone array characteristics needed to be used on realistic settings with asynchronous sound devices.**

I. INTRODUCTION

Processing the signals from a microphone array has proven to be an effective approach to improve robot audition. Many robot audition systems based on microphone arrays have been proposed in the literature [1] [2] [3] [4]. By exploiting this technique, robots are able to localize and track different sound sources, separate speeches coming from several people simultaneously and automatically recognize each separated speech. Most of these studies, however, require hardware synchronisation of each independent microphone channel. Specifically, synchronisation needs a special sound capturing device such as a multi-channel analog-to-digital (ADC) converter. While several commercial products exist, they are either too expensive or too large in size to be integrated inside robotic platforms [5]. Moreover, given the constraints of embedding microphones in a robot, often alongside other perception devices, it is difficult to measured the exact location of the microphones accurately. Instead of the exact location, a transfer function between each microphone and a sound source is measured. These measurements, however, can be quite time-consuming since they need to be obtained at multiple intervals of sound source directions (*e.g.* every 5 degree) [5].

Recent methods have started to relax these assumptions, for instance self-localisation approaches for ad-hoc arrays have been proposed in [6] [7] [8] [9]. Most of these approaches can achieve high accuracy in microphone array self-localisation. Both [6] and [9] provide closed-form estimators, in contrast to the standard iterative solution. The method presented in [6] also considers an acoustically determined, orientation estimate of a device that contains a microphone array. This method has been used in localisation of an asynchronous source in [10]. Raykar et al.'s work [11] on self-localisation formulates a maximum likelihood estimation for unknown parameters of a microphone array (time offsets, microphone positions) and measurements (time difference of arrival (TDOA) or time of flight (TOF)) by utilising active emissions. Ono et al. [12] presents a TDOA-based cost function approach, which does not require controlled calibration signal for estimating self-localisation, source localisation and temporal offset estimation. An online approach utilising simultaneous localisation and mapping (SLAM) is presented by Miura et al. [5], which used extended Kalman filtering and delay-and-sum beamforming to calibrate the stationary array.

While these methods being capable of computing individual microphone locations and the time offsets between different microphone channels, all of them are based on the assumption that the clock interval, in each independent sound card dedicated to each channel, is identical to those of the others. This is a strong assumption that disregards errors from fractional differences in clock intervals, which will accumulate over time. Sound cards, especially those designed for general consumption, have indeed noticeable drifts. An example is shown in Fig. 1. A microcontroller, connected to the signal line of each of these three microphones, generates a simultaneous pulse after a fix time interval, remaining at high impedance until the next regular pulse. Fig. 1 shows a detail of the difference in arrival time for each microphone, whilst Fig. 2 represents the evolution in the differences between each pair of channels with respect to the first one (vertical axis indicates the offset as number of samples to normalise the comparison). From this simple setup, it can be easily observed how time-offsets between pairs of channels keep increasing over time due to clock drifts from the small variations in the clock intervals of each sound card.

The method proposed here overcomes this issue by calibrating the asynchronous microphone array. The approach uses a graph-based SLAM method to calibrate the array,

All authors are associated to Centre for Autonomous System (CAS), University of Technology, Sydney (UTS), Australia. daobilige.su@student.uts.edu.au, {teresa.vidalcalleja, jaime.vallsmiro}@uts.edu.au

Preprint submitted to 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems. Received March 4, 2015.

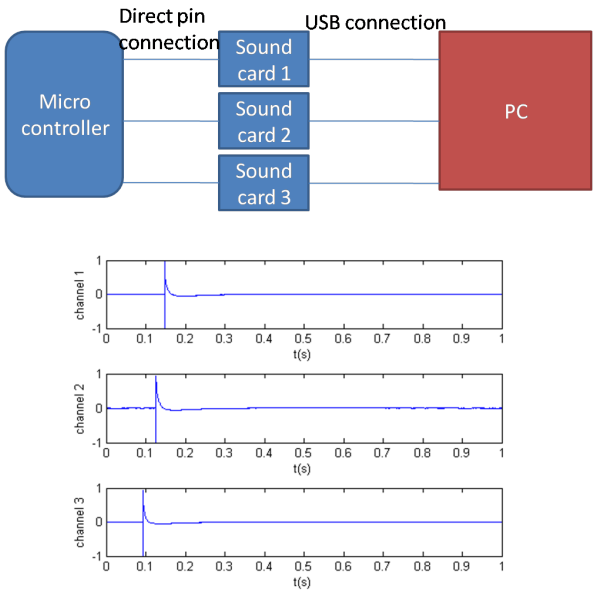


Fig. 1. Experimental setup for testing clock differences between pair of sound cards and recorded signal. The figure above shows the setup and the plot below shows the pulse generated per channel.

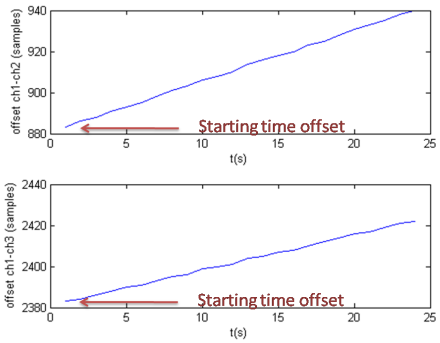


Fig. 2. Detected differences of peak arrival time. These signal time offsets are a combination of the starting time offsets and clock difference rate by the elapsed time.

which implies estimating the position, the time offset and clock difference of each microphone simultaneously. The trajectory of the sound source is also recovered at the same time. The SLAM problem is formulated as a sparse, least-squares minimisation problem, where the minimum is found iteratively using Gauss-Newtown algorithm. This is equivalent to finding the Maximum-Likelihood (ML) estimate of the sequence of sound sources locations and the array calibration under the assumption of Gaussian noise. Simulation and experimental results for a random walker sound source and arrays of variable number of microphones show the viability of the approach.

The rest of this paper is organised as follows. In section II, the detailed explanation of the proposed method is presented. Section III details the implementation of how the method is applied to solve the asynchronous microphone array calibration problem together with its initialisation and termination conditions. In section IV, comprehensive simulations and experimental results are presented. Section V presents the conclusion and discussion about further work.

II. The Proposed Approach

In our system, graph-based SLAM aims at recovering a sequence of sound source positions and the static location of the microphone array given a set of relative measurements. As the sound source moves around, the microphone array senses it from multiple viewpoints. Enforcing consistency between the different views gives rise to the location constraints (of sound source and microphone array).

A. System Model

Let $\mathbf{x}_{mic}$ be the state of the microphone array and $\mathbf{x}_{src\_k}$ be the position of the sound source at the time $t_k = t_1 ... t_K$. Thus the full state is given by,

$$\mathbf{x}^T = \left(\mathbf{x}^T_{mic}\mathbf{x}^T_{src\_1} \dots \mathbf{x}^T_{src\_K}\right), \tag{1}$$

where

$$\mathbf{x}^T_{mic} = \left(\mathbf{x}^T_{mic\_1} \dots \mathbf{x}^T_{mic\_N}\right), \tag{2}$$

and N is the total number of microphones.

Note that in this case the state of each microphone

$$\mathbf{x}^T_{mic\_n} = (x^x_{mic\_n} x^y_{mic\_n} x^\tau_{mic\_n} x^\delta_{mic\_n}) \tag{3}$$

for $n = 1 ... N$, where the location is given by the variables with subscripts x and y and the variables with subscripts τ and δ represent the starting time offset and the clock difference per second of each microphone respectively.

In a similar way, the state of the sound source

$$\mathbf{x}^T_{src\_k} = (x^x_{src\_k} x^y_{src\_k}) \tag{4}$$

contains only two variables that represent x and y position, as the orientation is not estimated in this case.

Let the microphone 1 be used as the reference, then time offsets and clock differences of other microphones are computed relative to the microphone 1. Hence, $x^\tau_{mic\_1} = 0$ and $x^\delta_{mic\_1} = 0$. Moreover, in order to define the position and orientation of the reference frame, the origin and, x and y axes need to be defined. As microphone 1 is the reference its position is set as $(0, 0)$. Let also another microphone (for instance the 2nd) define the positive direction of the x axis. This will fully define our reference frame, however, if the microphone array is bi-dimensional there will be two possible solutions for the position of the microphone array $\pm y$. In practice if the structure of the array is known, it can be exploited to remove the ambiguity in the y direction, *i.e* another microphone can define the positive direction of y axis. Note that it is assumed in any case that the number of microphones N is known and fixed.

To make the analogy to a standard SLAM framework as shown in Fig. 3, the sound source locations are treated as robot poses and the microphone array is treated as a single landmark. This landmark has the particularity of being

Preprint submitted to 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems. Received March 4, 2015.

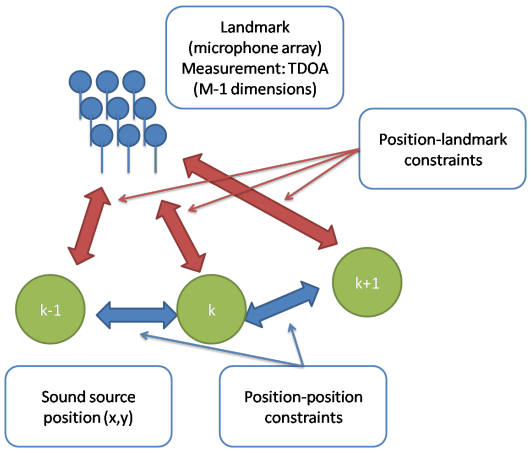


Fig. 3. Description of the sound source positions, microphone array and constraints in a SLAM framework.

observable at all sound source positions[1]. The microphone array becomes the first node in the graph-based SLAM framework and each sound source position becomes one node.

For the position-position constraints, in our case the trajectory of the sound source is assumed to be arbitrary (no-odometry prior is considered), with the only constraint of two adjacent locations set to be not too distant to each other. Therefore, we use a random walker model in which the sound source position of the next time instance is expected to be at the same location as the previous time with a large uncertainty associated as

$$\mathbf{z}^{p-p}_{k-1,k} = \mathbf{0} \tag{5}$$

$$I^{p-p}_{k-1,k} = \frac{1}{\sigma^2_{p-p}} I\,, \tag{6}$$

where $\mathbf{z}^{p-p}_{k-1,k}$ and $I^{p-p}_{k-1,k}$ denote the measurement and information matrices between positions $k-1$ and k for $k = 1 \dots K$. σ_{p-p} is the standard deviation of the random walker model within which the location of next sound source should fall. I denotes the identity matrix.

Regarding position-landmark constraints, the measurement represents TDOA values at each position of sound source. Specifically, the measurement is defined as

$$\mathbf{z}^{p-l^T}_{k} = (TDOA_{mic\_2,mic\_1} \dots TDOA_{mic\_N,mic\_1}) \tag{7}$$

where $TDOA_{mic\_n,mic\_1}$ for $n = 2 \dots N$ is the TDOA between microphone n and microphone 1, which is used as the reference as mentioned before. The information matrices for this position-landmark constraint is given by

$$I^{p-l}_{k} = \frac{1}{\sigma^2_{p-l}} I\,, \tag{8}$$

[1]Note that the main difference with a standard landmark-pose SLAM system is that here all the microphones can be observed at any time. In a standard SLAM system only part of the landmarks are observed at any time. This fact allows the microphone array to be treated as a single landmark with a large state that contains all the microphones. The same solution, however, is achieved if the the microphones are considered independently.

where σ_{p-l} the standard deviation of Gaussian distribution within which the error of each TDOA measurement should be.

B. Graph-Based SLAM Optimisation

In the least square problem of the graph-based SLAM, the estimated state vector is found by minimising the error over all position-position constraints and position-landmarks constraints [13]

$$\mathbf{x}^* = argmin \sum_{ij} e^T_{ij} \Omega_{ij} e_{ij} \tag{9}$$

where i and j mean ith and jth nodes in the graph-based SLAM.

This estimated $\mathbf{x}^*$ can be obtained by iterative Gauss-Newton optimisation [13].

$$\mathbf{x} = \mathbf{x} + \Delta\mathbf{x} \tag{10}$$

where

$$H\Delta\mathbf{x} = -b \tag{11}$$

where H and b are called coefficient vector and coefficient matrices. the computation of these two variables are computed as follow[13],

$$\begin{aligned} \bar{b}^T_i &= \sum e^T_{ij} \Omega_{ij} A_{ij} \\ \bar{b}^T_j &= \sum e^T_{ij} \Omega_{ij} B_{ij} \end{aligned} \tag{12}$$

and

$$\begin{aligned} \bar{H}_{ii} &= \sum A^T_{ij} \Omega_{ij} A_{ij} \\ \bar{H}_{ij} &= \sum A^T_{ij} \Omega_{ij} B_{ij} \\ \bar{H}_{ji} &= \sum B^T_{ij} \Omega_{ij} A_{ij} \\ \bar{H}_{jj} &= \sum B^T_{ij} \Omega_{ij} B_{ij} \end{aligned} \tag{13}$$

where $\bar{b}_i$ and $\bar{b}_j$ are ith and jth element of the coefficient vector b. $\bar{H}_{ii}$,$\bar{H}_{ij}$,$\bar{H}_{ji}$ and $\bar{H}_{jj}$ are sub block matrices parts of the coefficient matrices H. A_{ij} and B_{ij} are the Jacobian matrices of e_{ij} over the graph node $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively

$$\begin{aligned} A_{ij} &= \frac{\partial e(\mathbf{x}_i, \mathbf{x}_j)}{\partial \mathbf{x}_i} \\ B_{ij} &= \frac{\partial e(\mathbf{x}_i, \mathbf{x}_j)}{\partial \mathbf{x}_j}\,. \end{aligned} \tag{14}$$

In particular for the asynchronous microphone array, calibration problem are computed as shown below. For each position-position constraint of the sound source at t_{k-1} to t_k, the error function is computed as

$$\begin{aligned} e^{p-p}_{k-1,k} &= (\mathbf{x}_{src\_k} - \mathbf{x}_{src\_k-1}) - \mathbf{z}^{p-p}_{k-1,k} \\ &= \begin{bmatrix} x^x_{src\_k} - x^x_{src\_k-1} \\ x^y_{src\_k} - x^y_{src\_k-1} \end{bmatrix}. \end{aligned} \tag{15}$$

Then, Jacobian matrices for this position-position constraint is

$$A^{p-p}_{k-1,k} = \frac{\partial e^{p-p}_{k-1,k}}{\partial \mathbf{x}_{src\_k-1}} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \tag{16}$$

and

$$B_{k-1,k}^{p-p} = \frac{\partial e_{k-1,k}^{p-p}}{\partial \mathbf{x}_{src\_k}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{17}$$

Regarding the position-landmark constraint of the microphone array to the sound source location at t_k, the error function is defined as

$$\begin{aligned} e_k^{p-l} &= \hat{\mathbf{z}}_k^{p-l} - \mathbf{z}_k^{p-l} \\ &= \begin{bmatrix} \frac{\sqrt{(x_{mic\_2}^x - x_{src\_k}^x)^2 + (x_{mic\_2}^y - x_{src\_k}^y)^2}}{c} \\ \vdots \\ \frac{\sqrt{(x_{mic\_n}^x - x_{src\_k}^x)^2 + (x_{mic\_n}^y - x_{src\_k}^y)^2}}{c} \end{bmatrix} \\ &- \begin{bmatrix} \frac{\sqrt{{x_{src\_k}^x}^2 + {x_{src\_k}^y}^2}}{c} \\ \vdots \\ \frac{\sqrt{{x_{src\_k}^x}^2 + {x_{src\_k}^y}^2}}{c} \end{bmatrix} + \begin{bmatrix} x_{mic\_2}^\tau \\ \vdots \\ x_{mic\_n}^\tau \end{bmatrix} \\ &+ k\Delta t \begin{bmatrix} x_{mic\_2}^\delta \\ \vdots \\ x_{mic\_n}^\delta \end{bmatrix} - \mathbf{z}_k^{p-l} \end{aligned} \tag{18}$$

where $\hat{\mathbf{z}}_k^{p-l}$ denotes the predicted TDOA from the current state vector, c refers to speed of sound and Δt means the time interval between each sound source position.

Then, the Jacobian matrices for this error function result in

$$A_k^{p-l} = \frac{\partial e_k^{p-l}}{\partial \mathbf{x}_{mic}} = [\mathbf{0} J_{mic\_2}^{p-l} \dots J_{mic\_N}^{p-l}] \tag{19}$$

and

$$B_k^{p-l} = \frac{\partial e_k^{p-l}}{\partial \mathbf{x}_{src\_k}} = [J_{k\_x}^{p-l} J_{k\_y}^{p-l}] \tag{20}$$

where $J_{mic\_n}^{p-l}$ for $n = 1 \dots N$ is the partial derivative of e_k^{p-l} with respect to the state of the microphone n. Since the microphone 1 is used as a reference, its state is constant value. Thus, the Jacobian is equal to zero. $J_{mic\_n}^{p-l}$ is only nonzero at row n and this nonzero row is computed as

$$\begin{aligned} &J_{mic\_n}^{p-l}(n,:) = \\ &\begin{bmatrix} \frac{x_{mic\_2}^x - x_{src\_k}^x}{c\sqrt{(x_{mic\_2}^x - x_{src\_k}^x)^2 + (x_{mic\_2}^y - x_{src\_k}^y)^2}} \\ \frac{x_{mic\_2}^y - x_{src\_k}^y}{c\sqrt{(x_{mic\_2}^x - x_{src\_k}^x)^2 + (x_{mic\_2}^y - x_{src\_k}^y)^2}} \\ 1 \\ k\Delta t \end{bmatrix}^T \end{aligned} \tag{21}$$

$J_{k\_x}^{p-l}$ and $J_{k\_y}^{p-l}$ in Eq. 20 represent the Jacobian matrices of e_k^{p-l} with respect to the state of x and y locations of the sound source at time t_k respectively. These matrices are computed as

$$\begin{aligned} J_{k\_x}^{p-l} &= \begin{bmatrix} \frac{x_{src\_k}^x - x_{mic\_2}^x}{c\sqrt{(x_{mic\_2}^x - x_{src\_k}^x)^2 + (x_{mic\_2}^y - x_{src\_k}^y)^2}} \\ \vdots \\ \frac{x_{src\_k}^x - x_{mic\_N}^x}{c\sqrt{(x_{mic\_N}^x - x_{src\_k}^x)^2 + (x_{mic\_N}^y - x_{src\_k}^y)^2}} \end{bmatrix} \\ &- \begin{bmatrix} \frac{x_{src\_k}^x}{c\sqrt{{x_{src\_k}^x}^2 + {x_{src\_k}^y}^2}} \\ \vdots \\ \frac{x_{src\_k}^x}{c\sqrt{{x_{src\_k}^x}^2 + {x_{src\_k}^y}^2}} \end{bmatrix} \end{aligned} \tag{22}$$

$$\begin{aligned} J_{k\_y}^{p-l} &= \begin{bmatrix} \frac{x_{src\_k}^y - x_{mic\_2}^y}{c\sqrt{(x_{mic\_2}^x - x_{src\_k}^x)^2 + (x_{mic\_2}^y - x_{src\_k}^y)^2}} \\ \vdots \\ \frac{x_{src\_k}^y - x_{mic\_N}^y}{c\sqrt{(x_{mic\_N}^x - x_{src\_k}^x)^2 + (x_{mic\_N}^y - x_{src\_k}^y)^2}} \end{bmatrix} \\ &- \begin{bmatrix} \frac{x_{src\_k}^y}{c\sqrt{{x_{src\_k}^x}^2 + {x_{src\_k}^y}^2}} \\ \vdots \\ \frac{x_{src\_k}^y}{c\sqrt{{x_{src\_k}^x}^2 + {x_{src\_k}^y}^2}} \end{bmatrix} \end{aligned} \tag{23}$$

Finally, the block corresponding to the 1st microphone in H is set to identity matrices,

$$H(1:4, 1:4) = I\,. \tag{24}$$

III. Implementation

A. Application Setup

As mentioned above, we are interested in simultaneous sound source localisation and calibration of an asynchronous microphone array. In our implementation a sound-source moves randomly or following a pre-defined path around a room, where an array of microphones is fixed and recording. Once the array of microphones starts recording, one person holding the sound emitter (*e.g. a smart phone*) moves around the room. Then, recorded audio signals from all microphone channels are processed using our graph-based SLAM method, and sound source positions and the locations, starting time offsets and clock differences of all microphones are estimated. Note that once the array is calibrated, popular synchronous microphone array processing techniques can be applied for separation or localisation of multiple sound sources.

B. Initialisation and Termination Conditions

No prior knowledge of the microphone positions or sound source trajectories are assumed for initialization of the state

TABLE I

PARAMETERS SETTING IN SIMULATION

Parameters	Values
Number of microphones	9
Distance between microphones	0.5m
Maximum starting time offset	0.1s
Maximum clock difference	0.1ms
Observation (TDOA) noise STD	0.333ms
Sampling frequency	44.1 KHz
Random walker STD	0.333m
Adjacent sound source distance	0.05m
Maximum iterations	50
ϵ for $\Delta\mathbf{x}$	0.0001

vector. Therefore, the initial values correspond of those variables are randomly selected within the workspace. Moreover, zero starting time offsets and zero clock differences are assumed for the initialisation. Like any other optimization problem, least square optimization based graph SLAM also suffers from non-convergence from a bad initial value, a situation that can be minimized if approximate priors can be supplied.

The termination condition is based on the maximum number of iterations and change of the state vector $\Delta\mathbf{x}$. If the algorithm reaches a predefined maximum number of iterations, or the change of the state vector is smaller than a predefined threshold ϵ, the algorithm stops.

IV. VALIDATION RESULTS

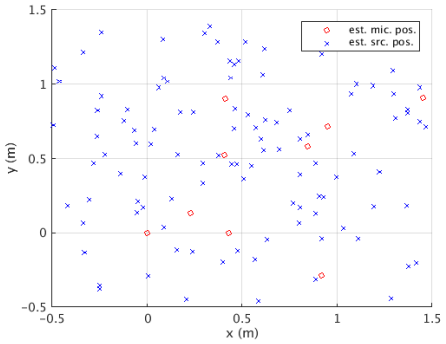
The validation of the proposed methodology is studied first in a simulation environment, where the performance of the proposed algorithm is tested under a variety of conditions with known ground truth. An experiment with a set of ordinary microphones was then conducted to show the effectiveness under realistic conditions.

A. Simulation Results

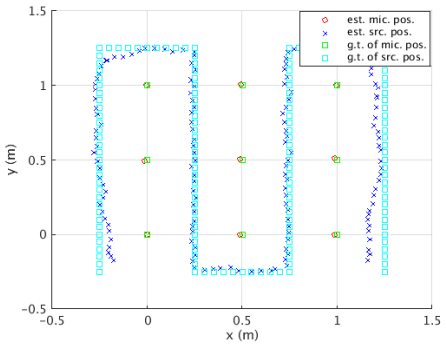
The parameters used in simulations are summarised in Table I. The observation noise and random walker model noise in simulations are obtained by empirical observation multiplied by conservative factor to deal with the worst case scenario. In all simulations, in order to obtain unique solutions, the position of microphone 1 is fixed at the origin of the coordinate system, microphone 3 in the 3×3 microphone array (microphone 2 in the 3×2 microphone array and microphone 4 in the 4×4 microphone array) is fixed at positive x axis and the y coordinate of microphone 4 in the 3×3 microphone array (microphone 3 in the 3×2 microphone array and microphone 5 in the 4×4 microphone array) is set to be positive.

Firstly, we performed a 10-run Monte Carlo simulation, which considers an array of 9 (3×3) microphones and sound source moving as (5). The results of the 1st Monte Carlo run are shown in Fig. 4. From the figure, it can be seen that, despite random initialisation, the proposed method is able to converge with good accuracy to the simulated values.

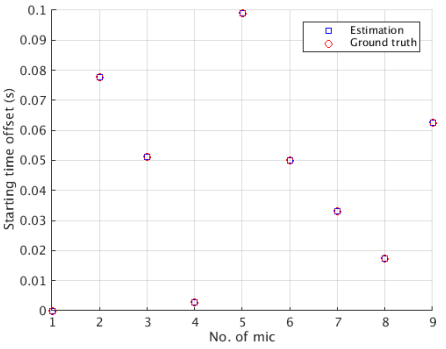
Secondly, in order to show the influence of the number of microphones over the estimation accuracy, another two



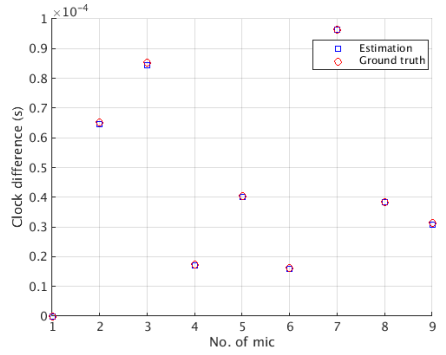
(a) Initialisation of the state vector.



(b) Final estimation results for microphone and sound source positions after convergence over 17 iterations.

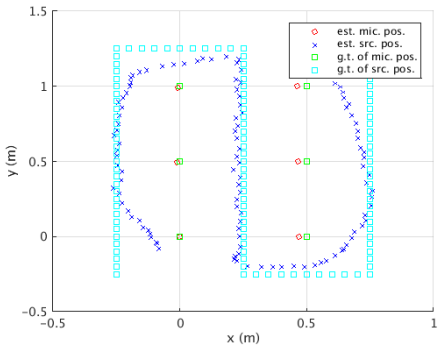


(c) Final estimation results for starting time offset.

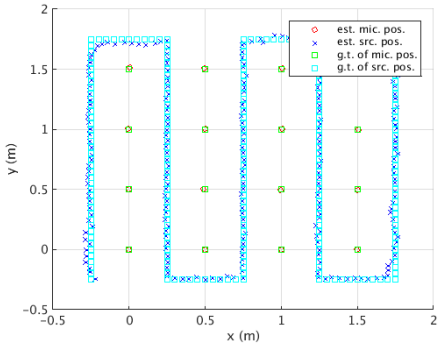


(d) Final estimation results for clock difference.

Fig. 4. Initialisation and final estimation results for a 3×3 asynchronous microphone array.



(a) Estimation results for one of 3×2 microphone array.



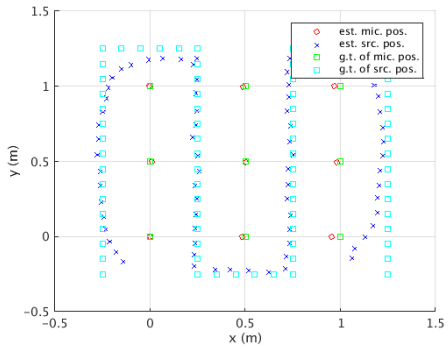
(b) Estimation results for one of 4×4 microphone array.

Fig. 5. Estimation results of 3×2 and 4×4 microphone arrays.

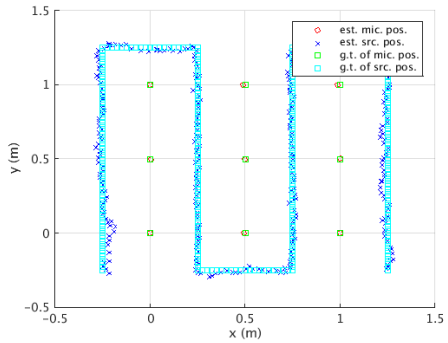
TABLE II

ESTIMATION RMS ERRORS OVER 3 DIFFERENT MICROPHONE ARRAYS OF THE 10-RUN MONTE CARLO SIMULATIONS FOR EACH CONFIGURATION

Arrangement	3×2	3×3	4×4
mean RMS error(m)	0.0826	0.0218	0.0100



(a) Estimation results of one of a 3×3 microphone array with half of the sound source positions.



(b) Estimation results of one of a 3×3 microphone array with twice of the sound source positions.

Fig. 6. Estimation results of half and twice of the sound source positions.

TABLE III

ESTIMATION RMS ERRORS WITH DIFFERENT NUMBERS OF SOUND SOURCE POSITIONS OF THE 10-RUN MONTE CARLO SIMULATIONS FOR EACH CONFIGURATION

number of sound source positions	half	original	twice
mean RMS error(m)	0.1083	0.0218	0.0074

10-run Monte Carlo simulations for a 3×2 and a 4×4 arrays are performed. The results are shown in Fig. 5. The comparison of the root mean square (RMS) errors for microphone positions in the 3×2, 3×3 and 4×4 arrangements is given in Table II. From the figure and the table, it can be seen that more numbers of microphones result in better estimation accuracy of microphone position and the usage of 9 microphones is sufficient, under the simulated TDOA observation noise, to recover the trajectory of the sound source with low RMS.

Finally, to test the influence of the number of sound source positions over the estimation accuracy, another two 10-run Monte Carlo simulations with half of the sound source positions and twice the number of the sound source positions are performed using the 3×3 microphone array. The results are shown in Fig. 6. The comparison of RMS errors of the microphone positions is given in Table III. The figure and table show that more number of sound source positions can lead to better estimation accuracy.

B. *Experimental Results*

To validate the proposed methodology, the following experimental set-up in an indoor seting was devised: an array of 6 microphones was fixed at a known location as shown by Fig. 7. These microphones were individually sampled by independent USB sound cards. Relevant parameters of the experimental set-up are summarised in Table IV. The observation noise and random walker model noise are obtained empirically. Again, in order to obtain an unique solution, the position of microphone 1 is fixed at the origin of the coordinate system, microphone 3 is fixed at positive x axis and the y coordinate of microphone 4 is set to be positive.

Recording of an incoming sound signals (a short time chirp) would then commence. A hand-held sound emitter (a smart phone producing a known soundwave) would move around the microphone array following two trajectories, one trajectory similar to the one in simulations and another different trajectory.

Fig. 7. Experimental setup of the asynchronous microphone array. Each channel of the array is sampled independently using individual USB sound card.

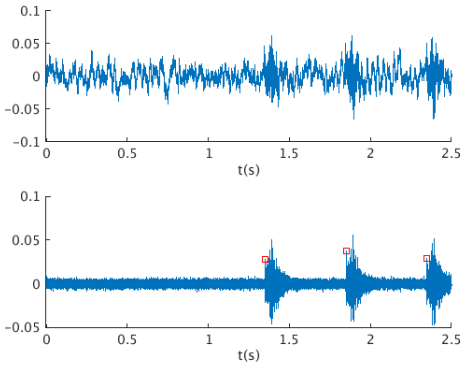


Fig. 8. Pre signal processing and detection of signal arrival (plot below) for raw audio data (plot above).

TABLE IV
EXPERIMENTAL SET-UP PARAMETERS

Parameters	Values
Number of microphones	6
Distance between microphones	0.5m
Observation (TDOA) noise STD assumed	0.167ms
Sampling frequency	44.1 KHz
Random walker STD	0.167m
sound source	Samsung Galaxy S4 phone
sound wave	short time chirp
time interval of sound	0.5s
total duration of recording	1min
Maximum iterations	50
ϵ for $\Delta\mathbf{x}$	0.0001

1) Signal Processing: The raw audio recording contains background noise and reverberation as shown in Fig. 8. An Equiripple high pass filter was used to clean the low frequency noise with a frequency lower than the lowest frequency of the emitted chirp signal. The first distinctive peak of the filtered wave was chosen as the arrival time of the signal. It should be noted that any other "peaky" signal rather than chirping soundwaves could have been used instead (e.g. hand clapping) since the algorithm is simply seeking a initial distinctive peak. For the purpose of testing the calibration procedure described here this is sufficient, although more robust mechanisms would be possibly needed in more complex acoustic environments.

2) Accuracies: The final estimation results are shown in the Fig. 9. Since we only have 6 microphones in total, the accuracy of the estimation expected to be similar to the 3×2 array and not as good as those with 9 or 16 microphones. However, final RMS errors for microphone positions are 0.0288m and 0.0204m respectively. These errors are much better than the simulation result (RMS error of 3×2 in Table II). The reason for this is that we have more sound source positions in the experimental setup than the simulation of 6 microphones in Fig. 5 (a). Moreover, the TDOA observation noise in experiment is smaller than the one used in the simulation, which is conservatively assumed to be 0.333ms and this can be easily achieved under 44.1 KHz sampling rate. The uncertainty associated to x and y positions of microphone 1 and y position of microphone 3 are zero since microphone 1 is fixed at origin and microphone 3 is fixed at positive x axis. The error of the estimation result can also come from non-precise measurements of the speed of the sound in the current experimental setup in addition to the observation noise. Using more microphones, like 9 or 16, or moving the sound source slower to have more sound source positions can improve estimation results further.

V. CONCLUSION

A method for simultaneous asynchronous microphone array calibration and sound source localisation using a graph-based SLAM approach is proposed in this paper. The method relaxes two key constraints imposed by traditional techniques employed for microphone array based sound source localisation and separation to obtain synchronous readings of an audio signal: knowledge of accurate geometry information of the microphone array, as well as availability of a multichannel analog-to-digital converter. In comparison with relevant techniques reported in the literature of asynchronous microphone array calibration, the proposed methodology estimates the clock difference of each independent sound card in addition to the starting time offset, thereby making it more suitable for generic applications with standard audio devices. This work is currently focused on embedding the asynchronous microphone array on a mobile robotic platform for sound source localisation to be used in tracking.

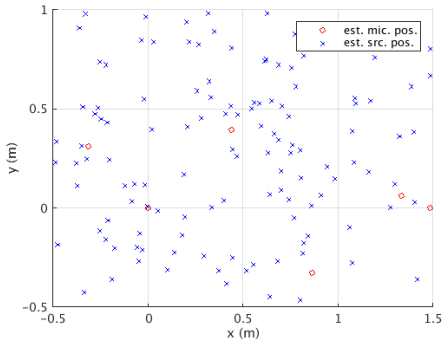
REFERENCES

[1] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *IEEE/RSJ International Conference on Intelligent Robots and Systems 2014 (IROS 2014)*, 2004, pp. 2123–2128.

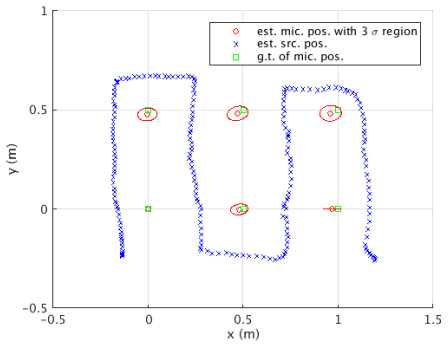
[2] S. Yamamoto, J.-M. Valin, K. Nakadai, J. Rouat, F. Michaud, T. Ogata, and H. G. Okuno, "Enhanced robot speech recognition based on microphone array source separation and missing feature theory," in *The 2005 IEEE International Conference on Robotics and Automation (ICRA 2005)*, 2005, pp. 1477–1482.

[3] K. Nakamura, K. Nakadai, F. Asano, and G. Ince, "Intelligent Sound Source Localization and its application to multimodal human tracking," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011)*, 2011, pp. 143–148.

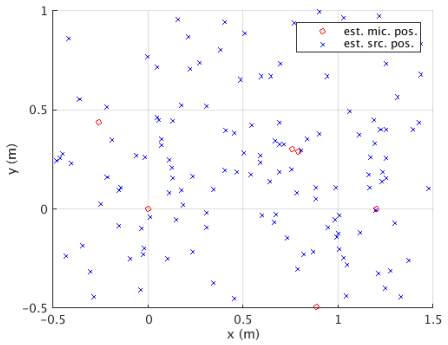
[4] K. Nakadai, S. Yamamoto, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "A robot referee for rock-paper-scissors sound games," in *The 2008 IEEE International Conference on Robotics and Automation (ICRA 2008)*, 2008, pp. 3469–3474.



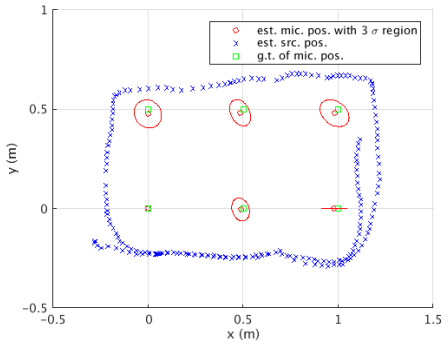
(a) Initialization of the state vector for the 1st trajectory.



(b) Final estimation results for the 1st trajectory. RMS error of microphone positions is 0.0288m.



(c) Initialization of the state vector for the 2nd trajectory.



(d) Final estimation results for the 2nd trajectory. RMS error of microphone positions is 0.0204m.

Fig. 9. Estimation results of experiments using a 2×3 asynchronous microphone array.

[5] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, "SLAM-based online calibration of asynchronous microphone array for robot audition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011)*, 2011, pp. 524–529.

[6] P. Pertila, M. Mieskolainen, and M. S. Hamalainen, "Closed-form self-localization of asynchronous microphone arrays," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011, pp. 139–144.

[7] M. H. Hennecke and G. A. Fink, "Towards acoustic self-localization of ad hoc smartphone arrays," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011, pp. 127–132.

[8] H. H. Fan and C. Yan, "Asynchronous differential TDOA for sensor self-localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007)*, 2007, pp. II–1109–II–1112.

[9] J. Bove, V. Michael, and B. Dalton, "Audio-based self-localization for ubiquitous sensor networks," in *Audio Engineering Society Convention 118*, 2005.

[10] A. Canclini, E. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE Transactions on Audio, Speech and Signal Processing*, vol. 21(2), pp. 439–443, 2013.

[11] V. C. Raykar, B. Yegnanarayana, S. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13(5), pp. 751–761, 2005.

[12] K. Hasegawa, N. Ono, S. Miyabe, and S. Sagayama, "Blind estimation of locations and time offsets for distributed recording devices," *Latent Variable Analysis and Signal Separation*, pp. 57–64, 2010.

[13] G. Grisetti, R. Kummerle, C. Stachniss, and W. Burgard, "A tutorial on graph-based SLAM," *IEEE Intelligent Transportation Systems Magazine*, vol. 2(4), pp. 31–43, 2010.