“© 2015 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”
Abstract—This paper presents a strategy for sound source localisation using an asynchronous microphone array. The proposed method is suitable for target tracking applications, in which the sound source with a known frequency is attached to the target. Conventional microphone array technologies require a multi-channel A/D converter for inter-microphone synchronization making the technology relatively expensive. In this work, the requirement of synchronization between channels is relaxed by adding an external reference audio signal. The only assumption is that the frequencies of the reference signal and the sound source attached to the target are fixed and known beforehand. By exploiting the information provided by the known reference signal, the Direction Of Arrival (DOA) of target sound source can be calculated in real-time. The key idea of the algorithm is to use the reference source to "pseudo-align" the audio signals from different channels. Once the channels are "pseudo-aligned", a dedicated DOA estimation method based on Time Difference Of Arrival (TDOA) can be employed to find the relative bearing information between the target sound source and microphone array. Due to the narrow band of frequency of target sound source, the proposed approach is proven to be robust to low signals-to-noise ratios. Comprehensive simulations and experimental results are presented to show the validity of the algorithm.

I. INTRODUCTION

Target localization and tracking is a well-studied topic in robotics and industry automation [1] [2] [3] [4]. Many successful applications based on laser range finder, camera and sound [5] [6] are found in the literature. Diverse scenarios include museum guidance [1], hospital assistance [7], pedestrian tracking [2] and search and rescue scenario [8]. Among these applications, target tracking based on sound is of increasing interest since the acoustic waves travel in all directions, can be detected at long distances from the sound source, and beyond the line of sight [9]. Moreover, the sound source can be detected in night time, cluttered environments or in fog, dust, smoke and dense forests, which makes it more reliable than visual cues.

There exists a number of examples of sound source localization, which rely on signal processing algorithms for microphone arrays to estimate the Direction Of Arrival (DOA) such as [10] [11] [12]. The signal processing algorithms for microphone arrays are capable to deal with sound source separation for automatic speech recognition and sound source localization independently. Most of these studies, however, require hardware synchronization of each independent microphone channel. Specifically, synchronization needs a special sound capturing device such as a multi-channel A/D converter. While several commercial products exist, they are either too expensive or too large in size to be integrated inside the robotic platform [13].

The aim of this work is to estimate the DOA of the sound source using an asynchronous microphone array, therefore relaxing the need of hardware synchronization. The asynchronous microphone array, however, has two main problems. Firstly, the time delay between each channel is unknown. Secondly, the clock from different sound cards for independent channels has slight differences. These slight differences can accumulate over long periods and dramatically influence the DOA estimation. Hence, the conventional DOA estimation algorithm can not be used directly with an asynchronous microphone array.

In this paper a DOA estimation algorithm for an asynchronous microphone array is proposed. The solution is based on adding a reference audio signal perpendicular to the microphone array as shown in Figure 1. It is assumed that the frequencies from both the reference signal and the target sound source are two known different single frequencies. The main idea is to use the reference signal to "pseudo align" different
channels of audio signals. In here, "pseudo alignment" means that the signals are aligned with closest cycle. This is not exactly as aligning the different channels of audio signal in order to compensate the time delay between each channel. Instead it aligns the reference signal within a wavelength of that particular frequency, as shown in Fig 2. As can be seen from the figure, the pseudo alignment results in the same Phase Difference (PD) of target sound source with exact alignment. The pseudo alignment can be easily obtained by directly applying the DOA estimation algorithm between every two audio channels at the particular frequency of reference signal. After all channels are pseudo aligned, DOA estimation algorithm will be employed again at frequency of the target sound source.

The remainder of the paper is organised as follow. In the section II, the related work on target tracking, sound source localisation and asynchronous microphone array is summarised. In section III, the details of the proposed method is illustrated. In section IV, comprehensive simulations and experimental results are presented. Section V presents the conclusion and discussion about further work.

II. RELATED WORK

There is a variety of target tracking methods reported in the literature. These methods mainly differ in the selection of sensors and features that are extracted to accomplish the tracking task. Aggarwal et al. [14] provide an overview of interpreting and tracking human motions using cameras. Gockley et al. [15] make use of a laser based approach to implement natural person following behaviours in their social robots. Valls Miro et al. compared laser and camera information for estimating indoor the human relative localisation for tracking tasks [4]. As discussed before, sound source localisation based target tracking has the advantage that acoustic waves travel omni-directionally and can be detected at long distance from the sound source, beyond line of sight, in night time or in fog, dust, smoke, dense forests or cluttered environment.

In the field of sound source localisation, there exist many approaches based on microphone arrays. The phase transform (PHAT) histogram method [16] is able to localise multiple sources, although source signals are not reconstructed. In [17], multiple speech sources localisation is obtained by using sinusoidal tracks to model speech and clustering the inter-channel phase differences between the dual channels of the tracks. High signal-to-noise ratio (SNR) is a requirement in this algorithm. Multiple signal classification (MUSIC) [18] and estimation of signal parameters via rotational invariance technique (ESPRIT) [19], known as subspaces methods, are used to estimate the directions of arrival (DOAs) of source signals. These methods are noise-robust but need more sensors than sources. However, most of these implementations need hardware synchronization of all audio channels.

In more practical applications, recent methods to relax these assumptions have started to be reported. Most of them focus on computing the time delays between different microphone channels. Specific self-localisation methods for ad-hoc arrays of such devices have been proposed in [20] [21] [22] [23]. These methods can achieve high accuracy self-localisation geometry. Ref. [20] and [23] provide closed-form estimators in contrast to iterative solutions. The method presented in [20] also considers acoustically determined orientation estimate of the device, which contains a microphone array. This method has been used in localisation of an asynchronous source in [24].

Raykar et al. [25] work on self-localisation formulates a maximum likelihood estimation for the unknown parameters (time offsets, microphone positions) and measurements (TDOA or TOF) by utilizing active emissions. Ono et al. [26] present a TDOA based cost function approach, which does not required controlled calibration signal, for estimating self-localisation, source localisation, and temporal offset estimation. However the fundamental issues of the cost function minimisation are the high dimensionality of the search space and the need of good initial guess in order to avoid local minimum. An online approach utilising simultaneous localisation and mapping (SLAM) is presented by Miura et al. [13], which used extended Kalman filtering and delay-and-sum beamforming to calibrate the stationary array.

Despite of the fact that all these calibration based methods are able to localise the sound source, they have two main problems. Firstly, the microphone arrays used are usually large in size, which makes it hard to be integrated into robotic platform. Secondly, the calibration based method has a basic assumption: the time delays between different channels are constant. However, according to our finding, the ad-hoc microphones sampled by different sound cards have slight differences in their clocks. This slight differences can accumulated and, over long periods, can potentially influence DOA estimation result. This situation becomes more significant when cheap, small, low quality ad-hoc sound cards are used.

III. PROPOSED APPROACH

The set up of the proposed method is shown in Fig 1. As can be seen from the figure, there is a microphone array that consists of M microphones. Each channel of the microphone is independently sampled by its own sound card, which makes the microphone array asynchronous. A reference signal of frequency \( F_{ref} \) is set perpendicular to the microphone array. The reference sound source is placed far enough so that it satisfies the far field condition. The sound source attached to the target is of frequency \( F_{tar} \).
In order to tackle the problem mentioned above, our method aims to pseudo align different channels of audio signal as shown in Fig. 2. The pseudo alignment of the asynchronous audio channels is obtained by the computation of the phase difference (PD) of the two channels as described in [27]. The detail of this PD calculation is described below.

Considering the situation of $N$ sources and $M$ microphones, the recorded signal at the $m$th microphone can be written using the convolutive mixing model as

$$x_m(t) = \sum_{n=1}^{N} \sum_{l=0}^{L} h_{mn}(l)s_n(t-l) \tag{1}$$

where $m = 1, \ldots, M$, $s_n$ is the signal from the source $n$, $x_m$ is the mixture signal captured by the microphone $m$, and $h_{mn}$ is the impulse response from the source $n$ to the microphone $m$ with $L$ being the maximum time delay due to the reverberation.

The PD is performed in the Time-Frequency (TF) domain, which is obtained by splitting the entire signal into short segments (frames) and taking the Short Time Fourier Transform (STFT) of each frame. In this domain as detailed in [27], the convolutive mixtures can be approximated as instantaneous mixtures and the mixing model can then be written in matrix notation as

$$X(f,l) = H(f)S(f,l) \tag{2}$$

where $f$ is the frequency index and $l$ is the frame time index. $H(f)$ is $M$ by $N$ mixing matrix corresponding to the impulse response $h_{mn}$, and $X(f,l) = [X_1(f,l), \ldots, X_M(f,l)]^T$ and $S(f,l) = [S_1(f,l), \ldots, S_M(f,l)]^T$ denote the vectors of the STFT of the mixture signals at all the microphones and of the source signals, respectively, at frequency $f$ and frame time $l$. In practice, the $X(f,l)$ can be obtained by STFT operation, while $H(f)$ and $S(f,l)$ are normally unknown.

If the TF cell $X(f,l)$ contains practical audio information, in this case if it comes from either reference signal or sound source attached to the target, the PD between channel $m_1$ and $m_2$ associated to this TF cell can be computed as

$$PD_{(m_1,m_2)}(f,l) = \text{angle} \left( \frac{X_{m_1}(f,l)}{X_{m_2}(f,l)} \right). \tag{3}$$

Here, the function $\text{angle}()$ computes the angle of a complete number. This phase difference $PD_{(m_1,m_2)}(f,l)$ corresponds to the minimum shift of two reference signals from microphone $m_1$ and $m_2$ since

$$-\pi < PD_{(m_1,m_2)}(f,l) \leq \pi \tag{4}$$

and this is exactly the pseudo alignment we are after.

In practical implementations, due to the limitations of the hardware, the STFT of the reference signal might not be exactly located in the frequency rule of the reference signal frequency in the STFT matrix. To deal with this, it is necessary to check the neighbouring frequency rules of STFT matrix, as shown in Fig. 5. Here, we use the weighted sum PD in the STFT cells that correspond to these frequency rules to obtain the estimated PD using

$$PD_{(m_1,m_2)}^{ref} = \sum_{l=1}^{L} \sum_{f=F_{ref}+N_{freq}}^{F_{ref}-N_{freq}} w_{(m_1,m_2)}(f,l)PD_{(m_1,m_2)}(f,l) \tag{5}$$

As shown in the Fig. 3, due to the asynchronous sampling, there is time delay between two channels of audio signal. If the exact time delay can be obtained by the cross-correlation of reference signal, the asynchronous array can be treated exactly as an synchronous array. Unfortunately, this is not the case when dealing with low SNR scenarios. As it can be seen from the Fig. 4, the recorded reference signal is quite different from the ideal scenario shown in the Fig 3. This is due to hardware limitations of the sound emitter and receiver and the reverberation of the sound in the environment. In this situation, purely cross-correlation (or generalised cross-correlation [26]) based on time of delay estimation can very easily get affected. Since most of the time the TDOA of different microphones is less than one wave cycle, the miss alignment of one cycle of wave from cross-correlation will lead to wrong DOA estimation. Moreover, when the reference signal is mixed with target sound source, the filtered reference signal is even noisier, which leads to failure in the cross-correlation methods.

A. Pseudo alignment of asynchronous channels using a reference signal

If the exact time delay can be obtained by the cross-correlation based method, the asynchronous array can be used to align two channels exactly using cross-correlation based method.
where $N_{\text{freq}}$ denotes number of adjacent frequency rules (normally $N_{\text{freq}} \leq 5$, although depends on the hardware and number of FFT in STFT) and $L$ denotes the number of time frame columns in STFT matrix. The weight

$$w_{(m_1,m_2)}(f,l) = \min(\|X_{m_1}(f,l)\|, \|X_{m_2}(f,l)\|)$$

(6)

as the minimum amplitude of STFT cells $X_{m_1}(f,l)$ and $X_{m_2}(f,l)$ from the audio channels $m_1$ and $m_2$. This implies that the cell with higher amplitude can provide better PD estimation.

### B. DOA estimation of the target sound source

Once the audio channels are pseudo aligned, the PD computation is performed once again between two channels using Eq. 3. Although this time instead of performing it at the frequency bin of reference signal, it needs to be done at the frequency of sound source attached to the target. In practical implementations, again, the adjacent frequency bins in STFT matrix also need to be considered. The PD of sound source on target $PD_{(m_1,m_2)}$ is computed as

$$PD_{(m_1,m_2)} = \sum_{l=1}^{L} \sum_{f=F_{\text{tar}}+N_{\text{freq}}}^{F_{\text{tar}}-N_{\text{freq}}} w_{(m_1, m_2)}(f,l)PD_{(m_1,m_2)}(f,l)$$

(7)

Note that the weight $w_{(m_1, m_2)}(f,l)$ is the same as in Eq. 6. This PD of the sound source on target $PD_{(m_1,m_2)}$ is without the pseudo alignment. After the pseudo alignment, the PD of the sound source on target $PD_{(m_1,m_2)}$ is

$$PD_{(m_1,m_2)}^{\text{pseudo aligned}} = PD_{(m_1,m_2)} - PD_{(m_1,m_2)}^{\text{ref}} \left( \frac{F_{\text{tar}}}{F_{\text{ref}}} \right)$$

(8)

Here, since the PD value for the pseudo alignment is frequency related, the PD of reference signal $PD_{(m_1,m_2)}^{\text{ref}}$ cannot be directly subtracted. It requires to be multiplied by the proportion of the different frequencies of the target sound source $F_{\text{tar}}$ and the reference signal $F_{\text{ref}}$.

In the last step, the DOA estimation of the target sound source from the audio data of channel $m_1$ and $m_2$ is computed as described in [27],

$$DOA_{(m_1,m_2)} = \cos^{-1} \left( \frac{PD_{(m_1,m_2)}^{\text{pseudo aligned}}}{2\pi F_{\text{tar}} c^{-1} d_{(m_1,m_2)}} \right)$$

(9)

where $d_{(m_1,m_2)}$ is the distance between microphone $m_1$ and $m_2$ and $c$ is the speed of sound.

### IV. EXPERIMENTAL RESULTS

In this section, validation of our method is presented. Firstly, we use a simulation environment to show the performance of the proposed algorithm. Then, we conduct an experiment to show its effectiveness in real world environments.

#### A. Simulation Results

In order to validate the performance of the method under different frequencies of the reference signal and the sound source on the target, a set of simulations varying these frequencies have been performed. All of simulations assumed that both the reference signal and the sound source attached to the target are in the far field of the microphone array. The parameters used in the simulation are summarised in Table I.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of microphones</td>
<td>4</td>
</tr>
<tr>
<td>Reference signal shape</td>
<td>sinusoid</td>
</tr>
<tr>
<td>Target signal shape</td>
<td>sinusoid</td>
</tr>
<tr>
<td>Time duration</td>
<td>0.25s</td>
</tr>
<tr>
<td>Source angle (degree)</td>
<td>45, 60, 90, 120, 135</td>
</tr>
<tr>
<td>Noise type</td>
<td>White Gaussian noise</td>
</tr>
<tr>
<td>Sampling frequency</td>
<td>44.1 KHz</td>
</tr>
<tr>
<td>Mixture type</td>
<td>Pure delay mixture</td>
</tr>
<tr>
<td>Mixture domain</td>
<td>Time domain</td>
</tr>
<tr>
<td>Frame length in STFT</td>
<td>4096</td>
</tr>
<tr>
<td>Frame overlap in STFT</td>
<td>2048</td>
</tr>
<tr>
<td>FFT window</td>
<td>Harming</td>
</tr>
<tr>
<td>Monte Carlo runs</td>
<td>20</td>
</tr>
</tbody>
</table>

Finally, since there are multiple pairs in a microphone array, the final DOA estimation is obtained by average the value of the DOAs from all pairs of the audio channels

$$D\hat{O}A_{(m_1,m_2)} = \frac{M^2}{\sum_{m_1=1}^{M} \sum_{m_2=1}^{M} a(m_1, m_2) DOA_{(m_1,m_2)}}$$

(10)

where

$$a(m_1, m_2) = \left\{ \begin{array}{ll} 1 & \text{if } m_1 \neq m_2 \\ 0 & \text{if } m_1 = m_2 \end{array} \right.$$  \hspace{1cm} (11)

TABLE I. PARAMETERS SETTING IN SIMULATION

In the simulation, there is no noise added to the mixture signal. As figure shows, the proposed method has a reasonably accurate DOA estimation when the frequency of target sound source is less than 4000Hz. Since there is no artificial noise added, the error mainly comes from discretisation of the signal. For the target sound source frequency equals or larger than the 4000Hz, these frequencies has exceeded the maximum frequency $F_{\text{max}}$ that the DOA estimation algorithm can deal with, as computed by

$$F_{\text{max}} = \frac{c}{2d}.$$  \hspace{1cm} (12)

Therefore, producing a large error in the results of these frequencies. Essentially, it means that the TDOA due to the different DOA has exceeded PD limit of $\pm \pi$.

We also are interested in analysing the performance of the approach under different SNRs. In this case, we performed multiple 20 runs Monte Carlo simulations for different noise levels. Different white Gaussian noises are added to the simulated mixture signal. As the performance measure, we...
used the root mean squared (rms) error of DOA estimates for
ilocalisation. The rms error of DOA estimate \( e_{rms} \) is computed
as
\[
e_{rms} = \sqrt{\frac{\sum_i (\tilde{\text{DOA}_{\text{tar},i}} - \text{DOA}_{\text{gt}})^2}{I}}
\]  
(13)

where \( i \) is Monte Carlo index, \( I \) is number of Monte Carlo
runs and \( \text{DOA}_{\text{gt}} \) is the ground truth of DOA.

The \( e_{rms} \) errors of the simulated scenario under various
SNRs are shown in Fig 8. In this scenario, the frequency of the reference signal is \( F_{\text{ref}} = 1500 \text{Hz} \) and the frequency of
the sound source attached to the target is \( F_{\text{tar}} = 3000 \text{Hz} \).
The ground truth of the DOA of the target is set to a constant
\( \text{DOA}_{\text{gt}} = 60(\text{deg})/135(\text{deg}) \). In each choice of SNR, 20
Monte Carlo runs are simulated. The results show that the
proposed method is robust to background noise (less than 3
degree rms error up to -5dB background noise).

\section*{B. Experimental Results}

Our experimental validation of the proposed method is
performed in an indoor environment. Fig. 9 shows the ex-
perimental setup with an array of 3 microphones is used. The
reference signal is set on top of the microphone array. The
reference audio signal is produced by a Nokia 5230
mobile phone. The frequency of the reference signal is set to
\( F_{\text{ref}} = 1500 \text{Hz} \). The target source is produce by a Samsung
Galaxy S4 mobile phone, which emits audio of frequency
\( F_{\text{tar}} = 3000 \text{Hz} \). The target is set static at 135 DOA. The
parameters of the experiment are summarised in Table II.

The experimental results is shown in Fig 10. As can be seen
from the figure, the proposed method produces a reasonable
accuracy of the DOA estimation. The error comes from rever-
beration of a small size room, background noise, not perfect
calibration and hardware quality. In order to improve the
accuracy, it is preferred to operate in an outdoor environment.
Moreover, a greater number of microphones will also improve
the accuracy of the DOA estimation.

\section*{V. Conclusion}

In this paper, the sound source localisation using an asyn-
nchronous microphone array has been investigated. Conven-
tional microphone array technologies require a multi-channel
A/D converter for inter-microphone synchronization. This
requirement of synchronization has been relaxed in our method
by adding a known reference signal. The proposed method
assumes that both reference signal and target sound source
has single frequency band. This method allows cheap/small
TABLE II. PARAMETERS SETTING IN EXPERIMENT

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of microphones</td>
<td></td>
</tr>
<tr>
<td>Reference audio hardware</td>
<td>Nokia 5230</td>
</tr>
<tr>
<td>Target audio hardware</td>
<td>Samsung Galaxy S4</td>
</tr>
<tr>
<td>Reference singal shape</td>
<td>sinusoid</td>
</tr>
<tr>
<td>Target singal shape</td>
<td>sinusoid</td>
</tr>
<tr>
<td>Time duration</td>
<td>0.25s</td>
</tr>
<tr>
<td>Source angle (degree)</td>
<td>135</td>
</tr>
<tr>
<td>Sampling frequency</td>
<td>44.1 KHz</td>
</tr>
<tr>
<td>Frame length in STFT</td>
<td>2048</td>
</tr>
<tr>
<td>FFT window</td>
<td>Harming</td>
</tr>
<tr>
<td>Monte Carlo runs</td>
<td>20</td>
</tr>
</tbody>
</table>

![Fig. 10. Experimental results of the proposed method with the target sound source at 135 degree. The rms error is 4.8213(degree).](image)

microphone arrays to be used for real time target tracking applications. The requirement is that target generates a single frequency acoustic signal. Comprehensive simulations and the a experiment results have shown the effectiveness of our method and its robustness to environmental noise. The future work includes exact alignment of the different channels by changing the reference signal.

REFERENCES


