Real-time Sound Source Localisation for Target Tracking Applications using an Asynchronous Microphone Array

Daobilige Su*, Jaime Valls Miro†, Teresa Vidal-Calleja†
Centre for Autonomous Systems
University of Technology Sydney
*Email: daobilige.su@student.uts.edu.au
†Email: {jaime.vallsmiro, teresa.vidalcalleja}@uts.edu.au

***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].

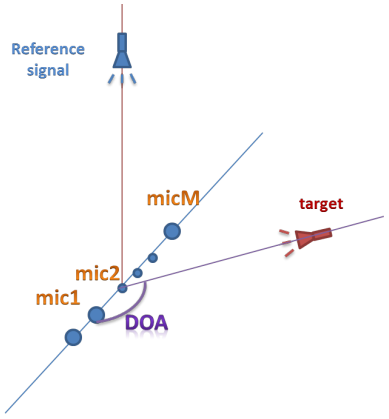

Fig. 1. System setup of the asynchronous microphone array method.

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

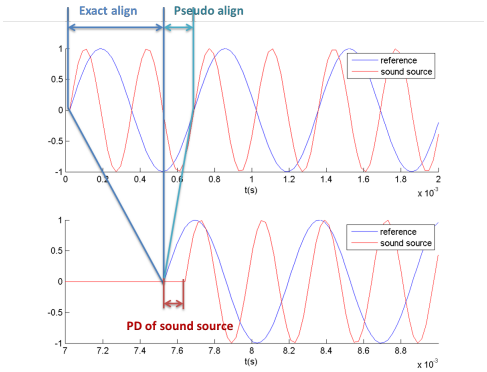

Fig. 2. Exact alignment and pseudo alignment. Both of the alignments result in same PD of the target sound source.

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}.

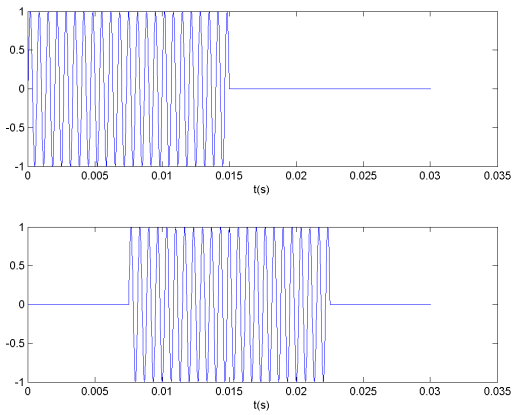

Fig. 3. Perfect reference signals (noise and reverberation free) that can be used to align two channels exactly using cross-correlation based method.

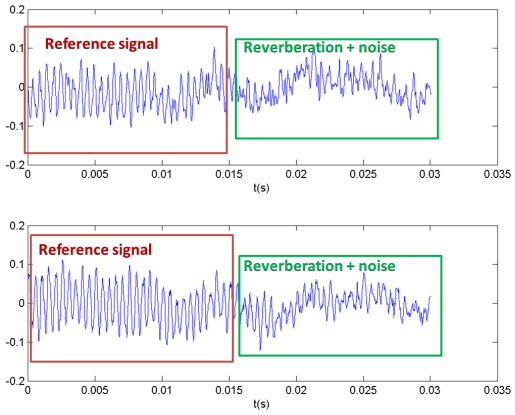

Fig. 4. Raw data of recorded reference signal. Due to the noise and reverberation, the cross-correlation based method can not guarantee a exact alignment.

A. Pseudo alignment of asynchronous channels using a reference signal

As shown in the Fig. 3, due to the asynchronous sampling, there is time delay between two channels of audio signal. If the 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 been 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.

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 mth 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 $m = 1, ..., M$, 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

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

where f is the frequency index and l is the frame time index. $\mathbf{H}(f)$ is M by N mixing matrix corresponding to the impulse response h_{mn}, and $\mathbf{X}(f,l) = [X_1(f,l), ..., X_M(f,l)]^T$ and $\mathbf{S}(f,l) = [S_1(f,l), ..., S_M(f,l)]^T$ 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 $\mathbf{X}(f,l)$ can be obtained by STFT operation, while $\mathbf{H}(f)$ and $\mathbf{S}(f,l)$ are normally unknown.

If the TF cell $\mathbf{X}(f,l)$ 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 $m1$ and $m2$ associated to this TF cell can be computed as

$$PD_{(m1,m2)}(f,l) = angle\left(\frac{\mathbf{X}_{m1}(f,l)}{\mathbf{X}_{m2}(f,l)}\right). \tag{3}$$

Here, the function $angle()$ computes the angle of a complete number. This phase difference $PD_{(m1,m2)}(f,l)$ corresponds to the minimum shift of two reference signals from microphone $m1$ and $m2$ since

$$-\pi < PD_{(m1,m2)}(f,l) \leq \pi \tag{4}$$

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

$$\hat{PD}^{ref}_{(m1,m2)} = \sum_{l=1}^{L} \sum_{f=F_{ref}-N_{freq}}^{F_{ref}+N_{freq}} w_{(m1,m2)}(f,l) PD_{(m1,m2)}(f,l) \tag{5}$$

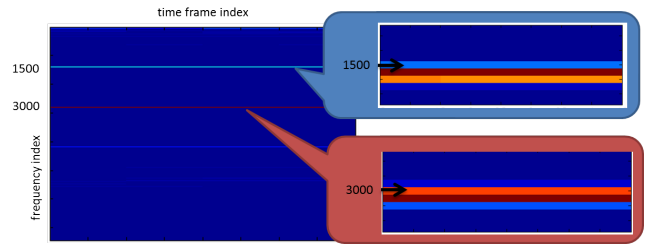

Fig. 5. STFT of recorded signal which consist of reference signal of 1500Hz and target sound source of 3000Hz.

where N_{freq} denotes number of adjacent frequency rules (normally $N_{freq} \leq 5$, 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

$$w_{(m1,m2)}(f,l) = min(\|X_{m1}(f,l)\|, \|X_{m2}(f,l)\|) \quad (6)$$

as the minimum amplitude of STFT cells $X_{m1}(f,l)$ and $X_{m2}(f,l)$ from the audio channels $m1$ and $m2$. 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 $\hat{PD}^{tar}_{(m1,m2)}$ is computed as

$$\hat{PD}^{tar}_{(m1,m2)} = \sum_{l=1}^{L} \sum_{f=F_{tar}-N_{freq}}^{F_{tar}+N_{freq}} w_{(m1,m2)}(f,l) PD_{(m1,m2)}(f,l) \quad (7)$$

Note that the weight $w_{(m1,m2)}(f,l)$ is the same as in Eq. 6. This PD of the sound source on target $\hat{PD}^{tar}_{(m1,m2)}$ is without the pseudo alignment. After the pseudo alignment, the PD of the sound source on target $PD^{tar\_pseudo\_aligned}_{(m1,m2)}$ is

$$PD^{tar\_pseudo\_aligned}_{(m1,m2)} = \hat{PD}^{tar}_{(m1,m2)} - \hat{PD}^{ref}_{(m1,m2)} \left(\frac{F_{tar}}{F_{ref}} \right) \quad (8)$$

Here, since the PD value for the pseudo alignment is frequency related, the PD of reference signal $\hat{PD}^{ref}_{(m1,m2)}$ cannot be directly subtracted. It requires to be multiplied by the proportion of the different frequencies of the target sound source F_{tar} and the reference signal F_{ref}.

In the last step, the DOA estimation of the target sound source from the audio data of channel $m1$ and $m2$ is computed as described in [27],

$$DOA^{tar}_{(m1,m2)} = \cos^{-1} \left(\frac{PD^{tar\_pseudo\_aligned}_{(m1,m2)}}{2\pi F_{tar} c^{-1} d_{(m1,m2)}} \right) \quad (9)$$

where $d_{(m1,m2)}$ is the distance between microphone $m1$ and $m2$ and c is the speed of sound.

TABLE I. PARAMETERS SETTING IN SIMULATION

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

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

$$\hat{DOA}^{tar} = \frac{\sum_{m1=1}^{M} \sum_{m2=1}^{M} a(m1,m2) DOA^{tar}_{(m1,m2)}}{M^2 - M} \quad (10)$$

where

$$a(m1,m2) = \begin{cases} 1 & \text{if } m1 \neq m2 \\ 0 & \text{if } m1 = m2 \end{cases} \quad (11)$$

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.

Simulation results are shown in Fig 6 and Fig 7. 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_{max} that the DOA estimation algorithm can deal with, as computed by

$$F_{max} = \frac{c}{2d}. \quad (12)$$

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

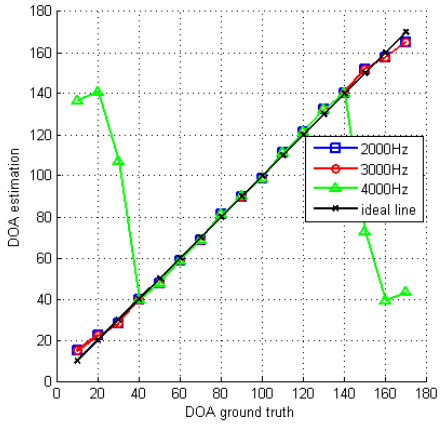

Fig. 6. Simulation results of proposed method with reference signal of frequency 1000Hz. The frequency of target sound source varies from 2000Hz to 4000Hz.

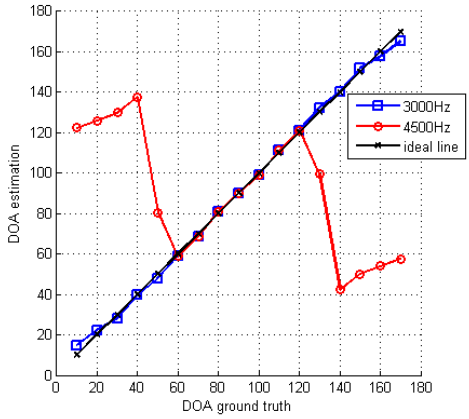

Fig. 7. Simulation results of proposed method with reference signal of frequency 1500Hz. The frequency of target sound source varies from 3000Hz to 4500Hz.

used the root mean squared (rms) error of DOA estimates for localisation. The rms error of DOA estimate e_{rms} is computed as

$$e_{rms} = \sqrt{\frac{\sum_i (\hat{DOA}_i^{tar} - DOA_{gt})^2}{I}} \qquad (13)$$

where i is Monte Carlo index, I is number of Monte Carlo runs and DOA_{gt} is the ground truth of DOA.

The e_{rms} errors of the simulated scenario under various SNRs are shown in the Fig 8. In this scenario, the frequency of the reference signal is $F_{ref} = 1500Hz$ and the frequency of the sound source attached to the target is $F_{tar} = 3000Hz$. The ground truth of the DOA of the target is set to a constant $DOA_{gt} = 60(\text{deg})/135(\text{deg})$. 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).

B. Experimental Results

Our experimental validation of the proposed method is performed in an indoor environment. Fig. 9 shows the experimental 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 $F_{ref} = 1500Hz$. The target source is produce by a Samsung Galaxy S4 mobile phone, which emits audio of frequency $F_{tar} = 3000Hz$. The target is set static at 135 DOA. The parameters of the experiment are summarised in Table II.

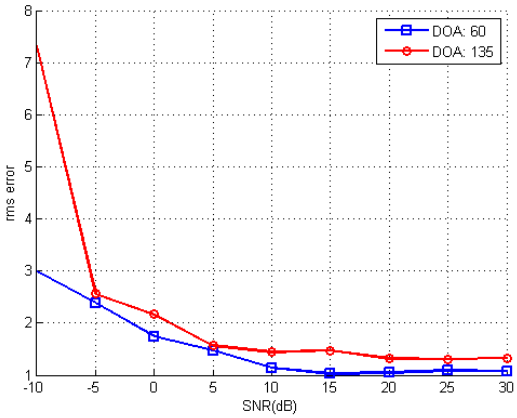

Fig. 8. Simulation results of rms error of the proposed method under different SNR which varies from -10dB to 30dB.

Fig. 9. Experimental setup of the asynchronous microphone array. Each channel of the array is sampled independently using individual USB sound card.

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 reverberation 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.

V. Conclusion

In this paper, the sound source localisation using an asynchronous microphone array has been investigated. Conventional 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

Parameters	Values
Number of microphones	3
Reference audio hardware	Nokia 5230
Target audio hardware	Samsung Galaxy S4
Reference singal shape	sinusoid
Target singal shape	sinusoid
Time duration	0.25s
Source angle (degree)	135
Sampling frequency	44.1 KHz
Frame length in STFT	4096
Frame overlap in STFT	2048
FFT window	Harming
Monte Carlo runs	20

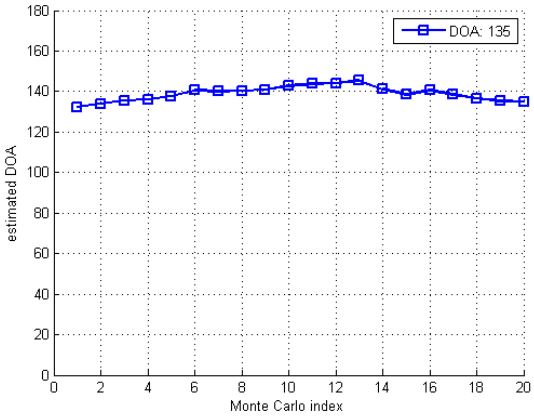

Fig. 10. Experimental results of the proposed method with the target sound source at 135 degree. The rms error is 4.8213(degree).

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

[1] S. Thrun, M. Bennewitz, W. Burgard, A. Cremers, F. Dellaert, D. Fox, D. Haehnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz, "Minerva: A second generation mobile tour-guide robot," in *IEEE International Conference on Robotics and Automation (ICRA 1999)*, 1999.

[2] L. Davis, V. Philomin, and R. Duraiswami, "Tracking humans from a moving platform," in *International Conference on Pattern Recognition (ICPR00)*, 2000, pp. 171–178.

[3] M. Kobilarov, G. Sukhatme, J. Hyams, and P. Batavia, "People tracking and following with mobile robot using an omnidirectional camera and a laser," in *IEEE International Conference on Robotics and Automation (ICRA 2006)*, 2006, pp. 557–562.

[4] J. V. Miro, J. Poon, and S. Huang, "Low-cost visual tracking with an intelligent wheelchair for innovative assistive care," in *12th International Conference on Control Automation Robotics & Vision (ICARCV 2012)*, 2012, pp. 1540–1545.

[5] Z. Chen and S. T. Birchfield, "Person following with a mobile robot using binocular feature-based tracking," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007)*, 2007, pp. 815–820.

[6] 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.

[7] J. F. Engelberger, "Health-care robotics goes commercial: The helpmate experience," *Robotica*, vol. 11, pp. 517–523, 1993.

[8] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, , and H. G. Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, 2013, pp. 3943–3948.

[9] M. Basiri, F. Schill, P. U. Lima, and D. Floreano, "Robust acoustic source localization of emergency signals from micro air vehicles," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012)*, 2012, pp. 4737–4742.

[10] U.-H. Kim, K. Nakadai, and H. G. Okuno, "Improved sound source localization and front-back disambiguation for humanoid robots with two ears," in *Recent Trends in Applied Artificial Intelligence*, 2013, pp. 282–291.

[11] K. Nakamura, K. Nakadai, and G. Ince, "Real-time super-resolution sound source localization for robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012)*, 2012, pp. 694–699.

[12] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21(10), pp. 2193–2206, 2013.

[13] 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.

[14] J. K. Aggarwal and Q. Cai, "Human motion analysis: a review," *Computer Vision and Image Understanding*, vol. 73(3), pp. 428–440, 1999.

[15] R. Gockley, J. Forlizzi, and R. Simmons, "Natural person-following behavior for social robots," in *2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI 2007)*, 2007, pp. 17–24.

[16] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 32(4), pp. 474–484, 2002.

[17] W. Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18(8), pp. 1913–1928, 2010.

[18] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34(3), pp. 276–280, 1986.

[19] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37(7), pp. 984–995, 1989.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] P. Janovi, X. Zou, and M. Kker, "Underdetermined DOA estimation via independent component analysis and time-frequency masking," *Journal of Electrical and Computer Engineering*, vol. 36, 2010.