A Bayesian Approach for Sound Source Estimation

Krittameth Teachasrisaksakul[1],
Surapa Thiemjarus[2], and **Chantri Polprasert**[3], Non-members

ABSTRACT

The goal of this study is to equip a mobile service robot with an ability to navigate to the human who commands the robot by speech. Based on time delay of arrival (TDOA) features, a Bayesian approach to sound source direction estimation is proposed. The method requires low computational complexity and is feasible for real-time robot navigation. Based on an experiment with various parameter settings in an indoor environment, different factors that affect the classification accuracy have been analyzed. The experiment results have illustrated that the proposed Bayesian method outperforms both that of the Microsoft Kinect sensor and the trigonometric approach in terms of classification accuracy.

Keywords: Acoustic Source Localization, Bayesian Network, Microsoft Kinect

1. INTRODUCTION

Recent advances in electronic, mechanic and sensor technology have made the concept of human-assisted robot become realizable. Most of the robots can only perform their assigned tasks to assist human with low degree of interactivity with human. There exist, however, many studies [1, 2, 3] that aim toward developing an adaptive robot which can work in a space filled with human and other robots.

To provide more intelligent services to users, indoor localization and human interaction should be seamlessly integrated into the robot navigation system. Information sources commonly used for robot navigation include computer vision [4, 5], sound [6, 7, 8], WiFi [9, 10], Radio-frequency Identification (RFID) [11], and low-power radio signal from wireless sensor network [12].

Based on information from different input modalities, such as a microphone array, camera(s), wireless signals, sensors, or a joystick, the robot can be equipped with an ability to interact with, moving its part(s) and navigate according to human commands. Robot control can be performed either or both autonomously or manually. Different modes of robot control can be designed upon the context of use so as to bring about the most convenient and natural experiences to users.

To achieve this goal, a framework for indoor robot navigation is proposed. As illustrated in Fig. 1, the framework consists of two modes of navigation. In the automatic navigation, a camera is used for detecting objects and obstacles, WiFi signal for room detection, and the microphone array for direction of the user's command. Based on the context detected by these multiple input modalities, the robot will perform autonomously according to the pre-programmed tasks. For example, once the direction of the speaker is detected by the microphone array, the camera on the robot's head can be turned to target the caller for depth detection. For more detailed robot control for achieving an un-predefined task, manual navigation can be carried out using a joystick or a mobile phone which is already carried by most users on a daily basis.

This paper focuses on the caller tracking module using acoustic source localization. A Bayesian approach for sound direction based on the time delay of arrival (TDOA) features has been proposed. The advantage of the technique is its ability to handle with uncertainty (e.g. due to noises or sound reflections). An experiment on various parameter settings has been conducted to evaluate the effect of each parameter on the model accuracy. For validation, both recorded sound samples and real human sound sources were used.

2. A REVIEW ON ACOUSTIC SOURCE LOCALIZATION

Acoustic source localization (ASL) [13, 14, 15] is a technique for estimating the direction and/or location of the sound source by exploiting audio cues extracted from sound signals received by one or more pairs of microphones. ASL is an active research topic and has been applied in various speaker-tracking applications, ranging from source localization micro-sensor [16], camera pointing system for video conferencing [17] to sound localization module in robots. The

Manuscript received on October 15, 2012 ; revised on April 5, 2013.

[1] The author is with School of Information, Computer, and Communication Technology Sirindhorn International Institute of Technology (SIIT), Thammasat University, Pathumthani, Thailand. , E-mail: krittameth.teachasrisaksakul@studentmail.siit.tu.ac.th

[2,3] The authors are with National Electronics and Computer Technology Center, Pathumthani, Thailand., E-mail: surapa.thiemjarus@nectec.or.th and chantri.polprasert@nectec.or.th

Fig. 1: *Indoor Robot Navigation Framework.*

multi-modal humanoid robot, CB, [18], and the service robot, Golem, [19] are the two examples of robot that use ASL for enhancing the interactivity.

ASL techniques can be divided into three major categories: 1) those maximizing a beamformer's steered response power (SRP), 2) those applying spectral estimation concepts, and 3) those using time-difference of arrival (TDOA). Examples of different types of location estimators used in the steer beamformer-based approaches include a maximum likelihood (ML) location estimator, a delay-and-sum beamformer, and a weighted beamformer. The theoretical analysis of ML estimator is presented in [20]. A delay-and-sum beamformer is extended to the case of multiple sources in [21]. The major drawback of this approach is inaccuracy stemmed from indistinct and many maxima of response power. This problem is alleviated in [22]. Examples of spectral-estimation-based approaches include the use of spectral analysis such as minimum variance (MV) spectral estimation [23] and eigenanalysis-based methods like multiple signal classification (MUSIC) algorithm [24].

Several time delay estimation (TDE) approaches have been explored over the past three decades. Among various approaches, generalized cross correlation (GCC) is the most prevalent. Many works studied GCC and its variations to enhance its accuracy [8, 13, 14] as well as compare their performances with those of other TDE algorithms such as Moddemeijer information theoretic delay criterion (MODD), and cochlear filtering (COCH) [18]. GCC-PHAT is implemented as a part of the system employing audiovisual information [25]. Other TDE algorithms include new approach using linear interpolation and multichannel cross correlation coefficient (MCCC) [26], maximum-likelihood approach [20], cross-power spectrum phase analysis [27], and creation of new spatial likelihood function by combining operation [28]. However, TDE performance is mainly impacted by room reverberation and noise. Several algorithms are proposed to cope with these causes of inaccuracy [29, 30, 31].

Two widely used cues in ASL include the inter-aural level difference (ILD) and inter-aural time difference (ITD). ILD and ITD measure the difference in loudness and time-of-arrival of an audio signal received at two microphones placed at different positions, respectively. Under free space or low reverberation environment, ITD and ILD yield direct relationship to the acoustic source position [14, 26]. However, under rich reverberant environment or no line-of-sight between the source and microphone array, reflections on boundary or sound scattering obstruct the system to directly locating the acoustic source. In [7, 32], this problem was addressed by matching of ITD and ILD features between the microphone pair. Liu et al. proposed the use of ITD and binaural signal processing scheme to extract directional acoustic information [33]. Another impediment to sound localization is a requirement of switching between near field and far field. In [34], this issue was addressed by finding the grid that is most matched with the current time difference feature vector.

ASL system that mimics human auditory system or binaural localization system is another attractive option for robot localization since it can localize sounds in three-dimensional environment using only two microphones. In [35], Rodemann et al. used binaural spectral cues for azimuth and elevation localization, and classified front-back source direction. Zhang et al. [36] determined the three-dimensional sound direction based on feature differences and Self-Organizing Map (SOM). Hornstein et al. [37] proposed the use of Head-Related Transfer Functions (HRTFs) to extract audio cues for creating audio-motor maps. Avni and Rafaely [38] studied the effect of incomplete representation of the sound field in the spherical harmonics domain. Since HRTFs rely on discrete measurement, Nakadai et al. [39] employed Active Direction-Pass Filter (ADPF), which performs ASL by auditory epipolar geometry, as an alternative method in a moving robot application.

3. ACOUSTIC SOURCE LOCALIZATION METHODS

Bayesian approach has been widely and successfully used for audio signal processing, especially for blind source separation. Hsin-Lung et al. [40], for example, incrementally identified dynamic sound sources and estimated independent component analysis (ICA) parameters by using online Bayesian learning. A few recent studies have applied Bayesian approach on acoustic source localization problem. In [41], ITD and IID were estimated before the Bayes-Rule was applied for final result. The approach has highest accuracy of azimuth localization in noisy environment compared to other three approaches. In [42], Yan et al. applied Bayesian theory to localization of sound source in a planar structure. TDOA is initially measured from sensor information then Bayesian updating of posterior probability distributions is performed to identify parameters about the source. The experimental results demonstrate that Bayesian approach has higher capability to deal with uncertainties.

In this study, a Bayesian approach for sound direction estimation based on the time delay of arrival (TDOA) features has been proposed. To provide a benchmark for measuring the performance of the proposed method, it is compared against two standard ASL methods, i.e. the ASL module provided in the Kinect sensor and a trigonometric approach.

3.1 A Bayesian Approach to Sound Source Estimation

The data analysis process comprises of four main steps, namely, feature extraction, data quantization, model learning and model inference. For each pair of the microphones, time delay of arrival (TDOA) features are extracted using the *generalized cross-correlation (GCC) method* described in [29]. Given that $X_i(\omega)$ and $X_i(\omega)$ are the Fourier transform of the sound signals received at the i^{th} and j^{th} microphones, their GCC function, $R_{ji}(\tau)$, can be expressed as:

$$R_{ij}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} (W_{ij}(\omega)X_i(\omega)X_j(\omega)^*)e^{j\omega\tau}d\omega \quad (1)$$

where $W_{ji}(\omega)1/|X_i(\omega)X_i(\omega)^*|$ is the phase transform (PHAT) weighting factor. $R_{ij}(\tau)$ shows an explicit global maximum at the lag value τ corresponding to the relative time delay. Therefore, TDOA, $\hat{\tau}_{ij}$, is estimated from:

$$\hat{\tau}_{ij} = \underset{\tau\in T}{\arg\max}\ \ R_{ij}(\tau) \quad (2)$$

where T denotes a set of all possible lag values.

For detection of the eight sound source directions, a discrete nave Bayesian network is used. The three TDOA features from each microphone pair are first quantized into eight states using mean and standard deviation calculated from the training data after outlier removal. To enhance model generalizability, a small constant value of 0.0001 was added to each parameter in the link matrix prior to normalization.

Assuming uniform prior probability, classification for sound source direction is made as follows:

$$\begin{aligned}\hat{\theta} &= \underset{k}{\arg\max}\ P(\theta_k|\hat{\tau}_{12},\hat{\tau}_{13},\hat{\tau}_{23}|)\\ &= \underset{k}{\arg\max}\ P(\hat{\tau}_{12}|\theta_k)P(\hat{\tau}_{13}|\theta_k)P(\hat{\tau}_{23}|\theta_k)\end{aligned} \quad (3)$$

where $\hat{\theta}$ denotes the predicted azimuth angle, $\hat{\tau}_{ji}$ denotes the TDOA value between a pair of microphones i and j.

3.2 A Trigonometric Approach

The sound source direction, θ, can be determined from an equilateral triangular microphone array using a trigonometric approach described in [43]. As shown in Fig. 2, the incidence angles between each microphone pair can be calculated as follows:

$$\alpha = \cos^{-1}\left(\frac{\hat{\tau}_{12}}{D}\right) \quad (4)$$

$$\beta = \cos^{-1}\left(\frac{\hat{\tau}_{23}}{D}\right) \quad (5)$$

$$\gamma = \cos^{-1}\left(\frac{\hat{\tau}_{13}}{D}\right) \quad (6)$$

where D denotes the distance between the microphones. Since the range of the function $\cos^{-1}$ is $[0,\pi]$, to obtain θ in 360°, the following conditions are used:

$$\theta = \begin{cases} \alpha - \frac{\pi}{6} & if \quad 0 \le \beta < \frac{\pi}{3} \\ \alpha - \frac{\pi}{6} & if \quad \frac{\pi}{3} \le \beta < \frac{2\pi}{3}, 0 \le \gamma < \frac{\pi}{3} \\ \gamma + \frac{\pi}{6} & if \quad \frac{\pi}{3} \le \beta < \frac{\pi}{2}, \frac{2\pi}{3} \le \gamma < \frac{5\pi}{6} \\ \gamma - \frac{11\pi}{6} & if \quad \frac{\pi}{2} \le \beta\frac{2\pi}{3}, \frac{5\pi}{6} \le \gamma < \pi \\ -\alpha - \frac{\pi}{6} & if \quad \frac{2\pi}{3} \le \beta \le \pi \end{cases} \quad (7)$$

3.3 ASL Capabilities of the Kinect Sensor

Fig. 3 illustrates different components of a Kinect sensor [44]. Kinect contains an array of 4 microphones and equipped with speech recognition capability and echo cancellation. The 16-bit audio from each of the four channels are processed at a rate of 16 KHz. For sound localization, the sensor employs the adaptive beamforming technique [45]. It supports 11 fixed beams, ranging from -50 to +50 degrees with 10 degree increments. The image sensing features of the sensor have been utilized in recent studies on robot navigation [46, 47].

4. DATA COLLECTION

To evaluate the performance of the proposed Bayesian method for sound source, acoustic source localization experiments were conducted in the Ambient Intelligence and Intelligent Informatics (AI3) Laboratory. The datasets have been collected in two experimental scenarios. For the first scenario, sound samples were played from the loudspeaker and recorded twice at each of the several different combinations azimuth angle, sound type, elevation, volume, and distance from the microphone as shown in Table 1. Sound recording was made using both an array of three omni-directional microphones and the Kinect sensor.

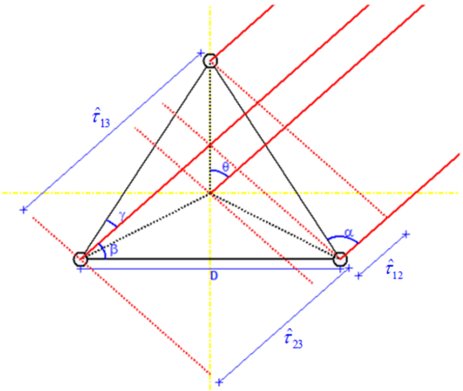


Fig.2: *Parameters involved in sound source estimation based on a trigonometric approach (Modified from [43]).*

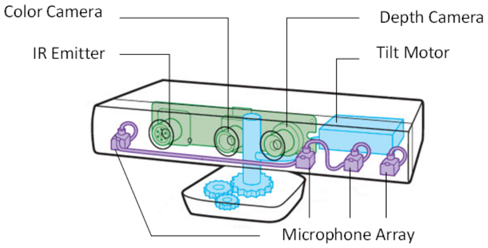


Fig.3: *Different parts of Kinect sensor (Modified from[44]).*

Table 1: *Experimental Parameters.*

Parameter Name	Values
Angle between microphone and loudspeaker	0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°
Sound type	clapping and the phrase "Come here"
Elevation of loudspeaker from floor	0.08, 0.1, 0.1395 meter
Loudness/volume of sound samples	122.47, 110.9, 98.7 dBA
Distance between microphone and loudspeaker	0.5, 1, 2, 3 meter(s)

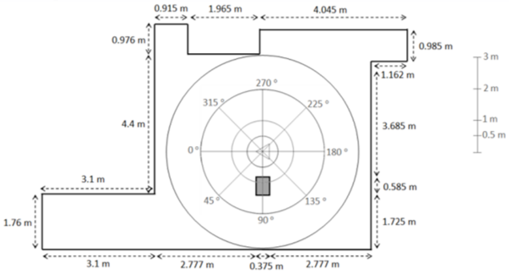


Fig.4: *AI3 room dimension and sound source locations.*

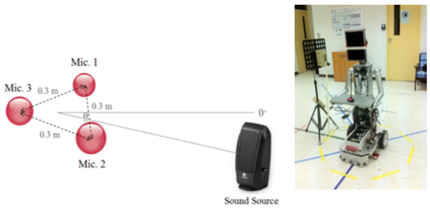


Fig.5: *Experimental setup: setting of the equilateral triangular microphone array (left) and its placement on the robot (right).*

An array of three omni-directional microphones was placed in an equilateral triangular configuration with a distance of 30cm apart as shown in Fig. 5(left). The microphone array was attached to the robot located in the middle of the room at 1.11m above the floor as illustrated in Fig. 5(right). At 90?, 1m, sound recording was not possible due to the existence of a column. A total of 1728 sound records were obtained in each iteration. The described experiment was repeatedly conducted using a Kinect sensor placed at the center of the room.

In the second experimental scenario, real male and female voice uttering the robot name, Dinsaw, at each point on the map was recorded to assess the model performance in a real-world environment. The heights of the male and female subjects are 1.70m and 1.52m, respectively. Another set of 192 sound records was obtained for model testing.

5. EXPERIMENTS AND RESULTS

The nave Bayesian model was trained with all the data collected during the first iteration of the first experimental scenario and tested with a subset of data collected during the second iteration. Fig. 6 (a-e) illustrate a comparison of the overall model accuracy of the Bayesian approach for sound source direction classification for different azimuth angles, sound types, elevation of the sound source, volumes, and distance from the microphone array.

According to Fig. 6(a), at 225°, 270° and 315°, an accuracy of above 90% can be achieved with the Bayesian network. However, the classification accuracy of sound signals acquired at 0o to 180o are degraded since the sound paths were partially blocked and reflected at the column located at 90 o, 1m from the middle of the room as shown in Fig. 4. With the Bayesian network, the overall classification accuracy of the test datasets is 82.29%. Fig. 6(b) shows that direction can be detected with higher accuracy for clapping sound compared to the speech sound. This is due to the distinct peaks of sound amplitudes reflected in the GCC coefficients. Fig. 6(c) shows that the classification accuracy increases as the sound source elevation become higher. Among the three elevation levels, sound direction can be detected with highest accuracy at 0.1395m since the loudspeaker is located higher than the height of the microphone array and the sound wave can directly travel to the transducer. Fig. 6(d) shows that the records with higher sound volume can be detected with higher accuracy due to the higher signal-to-noise ratio when the sound intensity is higher. Fig. 6(e) illustrates the accuracy of the sound source detection for varying distances from the microphone array. The accuracy is lowest at 0.5m, which is possibly due to the inaccurate plane wave assumption used in time delay estimation. The lower accuracy at 3m can be due to the attenuation of the sound signal as the source is moved further away. The best performance occurs in the setting with clapping sound, 0.1395m elevation of the speaker, high volume, and 2-meter distance from sound source.

To further assess the performance of the caller tracking module in a real-world environment, the dataset collected during the second experimental scenario was used for model testing. The overall accuracy for sound source detection is 93.75% and 90.63%, for female and male respectively.

Fig. 7(a-e) illustrate a comparison of the overall model accuracy of the Bayesian approach for sound source direction classification versus the trigonometric approach, and the ASL capabilities of Kinect sensor for different values of parameter settings. The trigonometric method described in Section 3.3 was applied on the 1728 sound records obtained from the second iteration of the first dataset. To enable the comparison, the continuous output angles from the trigonometric method and the Kinect sensor were first quantized into one of the eight different angles. The Bayesian approach has the overall accuracy of 82.29% which is higher compared to the trigonometric approach (70.14%) and that of Kinect sensor (29.17%).

The results in Fig. 7(b-e) demonstrate that both the Bayesian and the trigonometric approaches have the similar trend of the classification accuracy in different values of the four parameters, i.e. sample type, elevation, volume, and distance. However, in these four parameters, the Bayesian approach performs better than the trigonometric approach. Moreover, the trigonometric approach cannot estimate the angle from some part of datasets due to incorrect estimation of the TDOA value(s) from one or more microphone pair. The percentages of unclassifiable data in different parameter settings are shown in Table 2. The values are inversely proportional to the classification accuracy shown Fig. 7. As described in its specification [48], the Kinect sensor is capable of estimating angles between ?50 to +50 degrees. It is therefore unable to detect the sound source from sideward or backward areas.

The experimental results also show that both the Bayesian and the trigonometry approaches are more accurate than the sound source localization of Kinect sensor in all parameter settings. The accuracy trends of the Kinect sensor in different values of the four parameters are also different from the two approaches. For instance, the accuracy of detecting clapping sound and speech sound are almost equal and the accuracy of different volumes is also equal since the sensor detects the sound based on the confidence level. If the sound is evaluated at confidence level higher than 0.5, the sensor computes the angle. Otherwise, it does not compute the angle. Therefore, the recordings with different volumes and different sample types are interpreted the same by the sensor if they are evaluated at confidence level higher than 0.5. The accuracy of 1 m elevation is greater than that of 1.395m elevation. The accuracy of 2 m distance is lower than 3m distance.

Table 2: *Trigonometric Approach's Unclassifiable Percentage of Each Experimental Parameter (* denotes the lowest percentage in each parameter).*

Parameter Name	Parameter Value (degree)	Unclassifiable
Azimuth angle (degree)	0	4.17*
	45	11.11
	90	38.89
	135	11.11
	180	37.50
	225	25.00
	270	6.94
	315	11.11
Sample type	Clap	16.32*
	Speech	20.14
Elevation (meter)	0.08	23.96
	1	23.96
	1.395	6.77*
Volume (dB)	122.47	17.19
	110.9	16.67*
	98.7	20.83
Distance (meter)	0.5	21.53
	1	26.39
	2	10.42*
	3	14.58
All datasets	–	18.23

6. CONCLUSION

In this paper, we proposed a Bayesian approach to sound source localization. This approach can detect the direction of the speaker by using TDOA between three microphones as the input features and a nave Bayesian network as the classification model. The method is simple and yet requires low computational complexity.

Based on an equilateral-triangular microphone array, a dataset collected in various settings of sound type, source elevation, sound volume and the distance from the sound source. The datasets acquired during the first iteration is used for model construction. A detailed analysis of how different parameters affect the classification accuracy of the sound source has been performed by applying the model on the datasets acquired during the second iteration of the experiment. Further validation on both male and

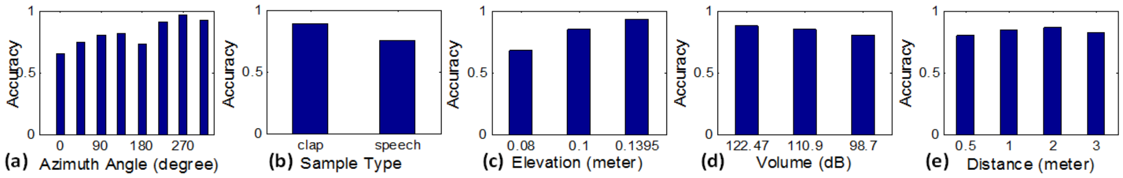


Fig.6: *A comparison of the classification accuracy of the test dataset versus different values of five parameters, i.e., a) azimuth angle, b) sound type, c) elevation, d) volume, and e) distance from the microphone.*

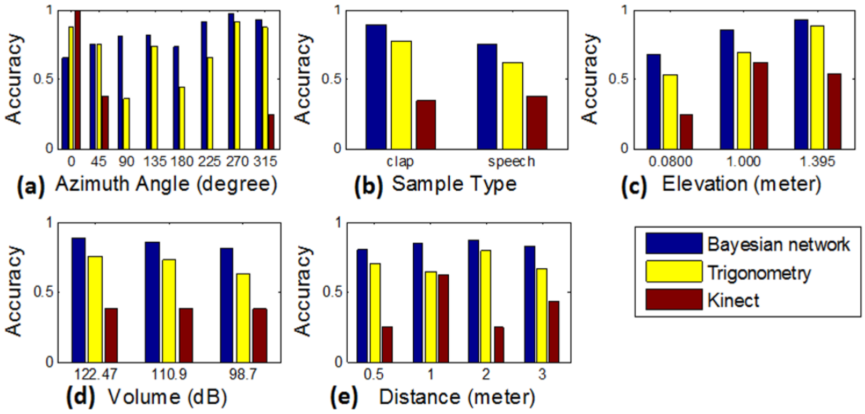


Fig.7: *A comparison of the classification accuracy of the test dataset versus different values of five parameters, i.e., a) azimuth angle, b) sound type, c) elevation, d) volume, and e) distance from the microphone.*

female subjects has shown that the model yields >90% overall classification accuracy.

Compared to the trigonometry method and the ASL module of the Kinect sensor, the proposed Bayesian approach yields a better performance in almost all the parameter settings. Apart from its simplicity and effectiveness, Bayesian approach also combines the advantages of both trigonometric approach and the ASL capabilities of Kinect sensor, which are ability to estimate direction of sound in 360o and using low computational load, respectively. Therefore, all of these comparisons can support the reliability of such method which is compelling for future usage and integration to multi-modal robot navigation system.

7. ACKNOWLEDGEMENT

The authors would like to thank Ms. Niracha Iemcha-od for her help during data collection. The equipments used in this study were supported by the Telecommunication Research and Industrial Development Institute (TRIDI) and Telecommunications Commission (NTC) of Thailand and CT Asia Robotics.

References

[1] T. Kruse, A. Kirsch, E. A. Sisbot, and R. Alami, “Exploiting human cooperation in human-centered robot navigation,” *the IEEE International Symposium on Robots and Human Interactive Communications.*, Viareggio, Italy, pp. 192–197, 2010.

[2] C.-P. Lam, C.-T. Chou, K.-H. Chiang, and L.-C. Fu, “Human-centered robot navigation-towards a harmoniously human-robot coexisting environment,” *IEEE Transactions on Robotics*, vol. 27, pp. 99–112 2011.

[3] J. Rios-Martinez, A. Spalanzani, and C. Laugier, “Understanding human interaction for probabilistic autonomous navigation using Risk-RRT approach,” *the International Conference on Intelligent Robots and Systems.*, San Francisco, CA, USA, pp. 2014–2019, 2010.

[4] H. Casarrubias-Vargas, A. Petrilli-Barcelo?, and E. Bayro-Corrochano, “EKF-SLAM and machine learning techniques for visual robot navigation,” *the International Conference on Pattern Recognition.*, Istanbul, Turkey, pp. 396–399, 2010.

[5] V. Nguyen, A. Harati, A. Martinelli, R. Siegwart, and N. Tomatis, “Orthogonal SLAM a step toward lightweight indoor autonomous navigation,” *the International Conference on Intelligent Robots and Systems.*, Beijing, China, pp. 5007–5012 2006

[6] J.-S. Hu, C.-Y. Chan, C.-K. Wang, and C.-C. Wang, “Simultaneous localization of a mobile robot and multiple sound sources using a microphone array,” *the International Conference on Robotics and Automation.*, Kobe, Japan, pp. 29–34, 2009.

[7] L. W. Wu, C.-C. Cheng, W.-H. Liu, and J. S. Hu,

"Gaussian mixture-sound field landmark model for robot localization " *the International Conference on Mechatronics and Automation.*, Niagara Falls, Ontario, Canada, vol. 1, pp. 438–443, 2005.

[8] H. Liu and M. Shen, "Continuous sound source localization based on microphone array for mobile robots," *the International Conference on Intelligent Robots and Systems.*, Taipei, Taiwan, vol. , pp. 4332 - 4339, 2010.

[9] J. Biswas and M. Veloso, "WiFi localization and navigation for autonomous indoor mobile robots," *the International Conference on Robotics and Automation.*, Anchorage, AK, USA, pp. 4379–4384, 2010.

[10] M. Ocana, L. M. Bergasa, M. A. Sotelo, and R. Flores, "Indoor robot navigation using a POMDP based on WiFi and ultrasound observations," *the International Conference on Intelligent Robots and Systems.*, pp. 2592–2597 2005

[11] W. Gueaieb and S. Miah, "An intelligent mobile robot navigation technique using RFID technology," *IEEE Transactions on Instrumentation and Measurement*, vol. 57, pp. 1908–1917, 2008.

[12] M. A. Batalin, G. S. Sukhatme, and M. Hattig, "Mobile robot navigation using a sensor network " *the International Conference on Robotics and Automation.*, Barcelona, Spain, vol. 1, pp. 636-641, 2004

[13] M. Azaria and D. Hertz, "Time delay estimation by generalized cross correlation methods," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 280–285, 1984.

[14] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, pp. 320–327, 1976.

[15] E.-E. Jan and J. Flanagan, "Sound source localization in reverberant environments using an outlier elimination algorithm," *the International Conference on Spoken Language.*, Philadelphia, PA , USA, vol. 3, pp. 1321–1324, 1996.

[16] A. P. Lisiewski, H. J. Liu, M. Yu, L. Currano, and D. Gee, "Fly-ear inspired micro-sensor for sound source localization in two dimensions," *Journal of the Acoustical Society of America*, vol. 129, pp. 166–171, 2011.

[17] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," *the International Conference on Acoustics, Speech, and Signal Processing.*, Munich, Germany, vol. 1, pp. 187–190, 1997.

[18] V. M. Trifa, A. Koene, J. Moren, and G. Cheng, "Real-time acoustic source localization in noisy environments for human-robot multimodal interaction," *the IEEE International Symposium on Robot and Human interactive Communication.*, Jeju, Korea, pp. 393–398, 2007.

[19] C. Rascon, H. Aviles, and L. A. Pineda, "Robotic orientation towards speaker for human-robot interaction," *the Ibero-American Conference on Artificial Intelligence.*, Bahía Blanca, Argentina, pp. 10–19, 2010.

[20] J. C. Chen, K. Yao, and R. E. Hudson, "Acoustic source localization and beamforming theory and practice," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 359–370, 2003.

[21] Y. Sasaki, S. Masunaga, S. Thompson, S. Kagami, and H. Mizoguchi, "Sound localization and separation for mobile robot tele-operation by tri-concentric microphone array," *the National Institute of Advanced Industrial Science and Technology Digital Human Symposium.*, Tokyo, Japan, 2009

[22] J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *the International Conference on Acoustics, Speech and Signal Processing.*, Toulouse, France, pp. IV – IV 2006

[23] J. Dmochowski, J. Benesty, and S. Affes, "Linearly constrained minimum variance source localization and spectral estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 1490–1502, 2008.

[24] C. T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments," *the International Conference on Intelligent Robots and Systems.*, St. Louis, MO, USA, pp. 2027–2032, 2009.

[25] B. C. Park, K. C. Kwak, K.-D. Ban, and H.-S. Yoon, "Sound source localization based on audio-visual information for intelligent service robots," *the International Symposium on Advanced Intelligent Systems.*, Sokcho, South Korea, pp. 364–367, 2007.

[26] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 509–519, 2004.

[27] U.-H. Kim, J. Kim, D. Kim, H. Kim, and B.-J. You, "Speaker localization on a humanoid robot's head using the time delay of arrival-based feature matrix," *the International Symposium on Robot and Human Interactive Communication.*, Munich, Germany, pp. 610–615, 2008.

[28] P. Pertila, T. Korhonen, and A. Visa, "Measurement combination for acoustic source localization in a room environment," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, pp. 1–15, 2008.

[29] J. Dibiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., pp. 157–180, 2001.

[30] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, “Robust sound source localization using a microphone array on a mobile robot,” *the International Conference on Intelligent Robots and Systems*, vol. 2, pp. 1228–1233, 2003.

[31] S. Doclo and M. Moonen, “Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments,” EURASIP Journal on Applied Signal Processing, vol. 2003, pp. 1110–1124, 2003.

[32] J.-S. Hu, W.-H. Liu, and C.-C. Cheng, “Indoor sound field feature matching for robot’s location and orientation detection,” *Pattern Recognition Letters*, vol. 29, pp. 149–160, 2008.

[33] R. Liu and Y. Wang, “Azimuthal source localization using interaural coherence in a robotic dog: modeling and application,” Robotica, vol. 28, pp. 1013-1020, 2010.

[34] X. Li, H. Liu, and X. Yang, “Sound source localization for mobile robot based on time difference feature and space grid matching ” *the International Conference on Intelligent Robots and Systems.*, San Francisco, California, USA, pp. 2879–2886, 2011.

[35] T. Rodemann, G. Ince, F. Joublin, and C. Goerick, “Using binaural and spectral cues for azimuth and elevation localization,” *the International Conference on Intelligent Robots and Systems.*, Nice, France, pp. 2185–2190, 2008.

[36] Z. Zhang, K. I., T. Miyake, and T. Imamura, “Three-dimension sound localization by binaural model using self-organizing map,” *International Journal of Innovative Computing, Information and Control*, vol. 6, pp. 361–371, 2010.

[37] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda, “Sound localization for humanoid robot - building audio-motor maps based on the head-related transfer function,” *the International Conference on Intelligent Robots and Systems.*, Beijing, China, pp. 1170–1176, 2006.

[38] A. Avni and B. Rafaely, “Sound localization in a sound field represented by spherical harmonics,” *the International Symposium on Ambisonics and Spherical Acoustics.*, Paris, France, 2010.

[39] K. Nakadai, H. G. Okuno, and H. Kitano, “Real-time Sound Source Localization and Separation for Robot Audition,” *the International Conference on Spoken Language Processing.*, Denver, Colorado, USA, 2002.

[40] H. Hsin-Lung and C. Jen-Tzung, “Online Bayesian learning for dynamic source separation,” *the IEEE International Conference on Acoustics Speech and Signal Processing.*, Dallas, Texas, USA, pp. 1950-1953, 2010.

[41] H. Liu, Z. Fu, and X. Li, “A two-layer probabilistic model based on time-delay compensation for binaural sound localization,” *the IEEE International Conference on Robotics and Automation.*, Karlsruhe, Germany, pp. 2705-2712, 2013.

[42] G. Yan and J. Tang, “A Bayesian approach for damage localization in plate-like structures using Lamb waves,” *Smart Materials and Structures*, vol. 22, pp. 035012, 2013.

[43] N. Bacani, A. Beauvillier, F. Kasting, and R. Gruger, “Final Report: Development Of A Web-Based Design Activity Monitoring and Collaboration Tool ”, University Of Victoria, Victoria,1 April 2011, 2011.

[44] Wired.com. (2011, 10 July). *Kinect Hackers Are Changing the Future of Robotics.* Available: http://www.wired.com/magazine/2011/06/mf_kinect/2/

[45] M. Zhang and M. H. Er, “Adaptive beamforming by microphone arrays,” *the IEEE Global Telecommunications Conference.*, Singapore, vol. 1, pp. 163-167, 1995.

[46] P. Benavidez and M. Jamshidi, “Mobile robot navigation and target tracking,” *the International Conference on System of Systems Engineering.*, Albuquerque, New Mexico, USA, pp. 299 - 304, 2011.

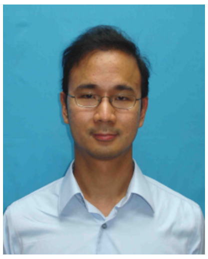
[47] J. Rios-Martinez, A. Renzaglia, A. Spalanzani, A. Martinelli, and C. Laugier, “Navigating between people: a stochastic optimization approach,” *the IEEE International Conference on Robotics and Automation.*, St. Paul, MN, USA, 2012.

[48] Microsoft. (2011, 10 July). Audio Stream. Available: http://msdn.microsoft.com/en-us/library/jj131026

Krittameth Teachasrisaksakul received the BSc degree in Information Technology from Sirindhorn International Institute of Technology, Thammasat University, Thailand in 2012. He is the recipient of Anandamahidol Scholarship under the Royal Patronage of His Majesty the King of Thailand, and currently studies PhD in Computing at Imperial College London, UK. His research interests include signal processing, motion estimation and humanoid robots.

Surapa Thiemjarus received her BSc degree with first honor in Information Technology from Sirindhorn International Institute of Technology, Thammasat University, Thailand, in 2001. She also received MSc in Advanced Computing from Imperial College, UK and MPhil in Speech, Text Processing and Internet Technology from University of Cambridge, UK, in 2002 and 2003, respectively. She was the recipient of Anandamahidol Scholarship under the Royal Patronage of His Majesty the King of Thailand from 2001 to 2007 and received her PhD in Computing from Imperial College, UK in 2007. She served as a faculty member in the School of Information, Computer, and Communication Technology, Sirindhorn

International Institute of Technology, Thammasat University and currently works as a researcher at the National Electronics and Computer Technology Center, Pathumthani, Thailand. Her research interests include machine learning, pattern recognition, Body Sensor Networks, context-aware and pervasive sensing.



Chantri Polprasert received the B.S. degree in electrical engineering from Chulalongkorn University, Bangkok, Thailand, in 1999, the M.S. degree in telecommunications from Asian Institute of Technology (AIT) in 2000, and the Ph.D. degree in eletrical engineering from the University of Washington, Seattle, in 2009. Since 2010, he has been a researcher at the National Electronics and Computer Technology Center, Pathumthani, Thailand. His research interests include communications over time- and frequency-selective fading channels, channel estimation and equalization, acoustical signal processing, brain-computer interfaces.