2D Binaural Sound Localization: for Urban Search and Rescue Robotics

A. R. Kulaib*, M. Al-Mualla, and D. Vernon

Khalifa University
Sharjah, UAE
**E-mail: akulaib@kustar.ae*
www.kustar.ac.ae

Being able to localize the position of a sound source is an important issue in robotics and many other application areas, since it enables those systems to interact with the environment. For example, USAR robotics can use sound to search for hidden victims that are shouting for help. The 2D binaural sound localization system in this paper is inspired by the human auditory system and are based on the Interaural Time Difference (ITD) and the Head Related Transfer Function (HRTF). The ITD is used to localize a sound source in the horizontal plane relaying on the difference of arrival times of the sound signal between the two microphones and notches in frequency spectra (HRTF) are used to localize a sound source in the vertical plane. To easily and accurately extract notches we used a spiral shaped pinna that allows notches in the frequency spectra to change linearly as a sound source moves in the vertical plane, giving a relationship between notches and elevation angle. The two models used in this paper are tested to assess their accuracy, show their limitations and we concluded by noting how accuracy and repeatability can be improved.

Keywords: 2D Sound Localization, ITD, HRTF.

1. Introduction

Sound systems can sometime provide more useful information than other senses. For example, in a dark room, sound systems are capable of detecting the position of a sound source, while vision systems have strong limitation in those situations. Sound localization techniques are used in C4ISR (Command, Control, Communications, Computers and Intelligence, Surveillance and Reconnaissance) systems [1,2], Urban search-and-rescue (USAR) [3], and systems that help blind people [4] and children with autism [5]. Also it can be used in surveillance systems [6] and to improve virtual reality.

The goal of this paper is to show how sounds can be localized in space using a binaural microphone setup. These localized sounds can be used to guide a USAR robot toward a potential victim. Then the robot will inform the emergency centre in order to take the appropriate action.

This paper is divided into two main sections. Section 2 provide a description of the ITD model for localizing sound in the horizontal plane, stating its advantages, features, and its limitation, while Sec. 3 explains the spiral ear model and how it can be used to measure elevation angle in the vertical plane.

2. Measurement of the Angle of Arrival using Interaural Time Difference

Using a binaural (two-microphone) configuration, the azimuth angle of arrival of a sound sensed by a robot can be computed from Eq. (1) [7]:

$$\theta = \sin^{-1}\left(\frac{\Delta l}{l}\right) \tag{1}$$

where l is the known interaural distance, $i.e$ the horizontal distance between the two microphones, and Δl is the difference in distance travelled by the sound wave when it arrives at the left and right microphones.

In turn, Δl is simply the speed of sound c multiplied by the time taken to travel this distance. This time is know as the Interaural Time Difference (ITD). Hence

$$\theta = \sin^{-1}\left(\frac{ITD \times c}{l}\right) \tag{2}$$

ITD can be computed straight from the cross-correlation of the left and right microphone signals. Specifically, ITD is given by Eq. (3) [7].

$$ITD = \frac{n}{F_s} \tag{3}$$

where n is the position (in samples) of the maximum of the cross-correlation function and F_s is the sampling frequency (samples/unit time).

Figure 1 shows a schematic overview of this measurement process. The main software components are Capture sound block, Cross-correlation block, Interaural Time difference block and Azimuth computation block. On the other hand, the hardware components used in the model are a sound source (e.g. loudspeaker), a dummy head (iCub) designed by the RobotCub consortium and an omni-directional condenser microphones with pre amplifier circuit. The microphones are connected to the line-in input of a computer and software is used to sample the left and right channels with the desired sampling rate. These samples represent an acoustic sound that can be noise or someone who is shouting for help. A simple threshold equation is used to decide if the signal will be passed for further processing or not. The threshold equation computes the energy of the two signals using the following equation: $\sum_n (s_l^2(n) + s_r^2(n))/n$ [7]. During a calibration phase the maximum threshold is calculated in a test room without producing any sound. Then if the sound is louder than the maximum threshold value, it will pass to the next step to compute the normalized correlation between the left and the right signals.

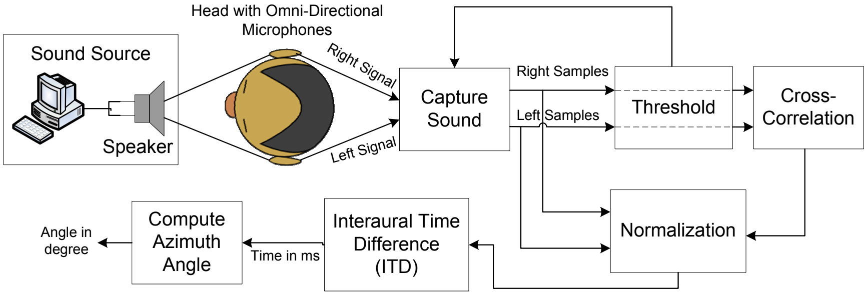


Fig. 1. ITD system model.

The ITD model were tested in a medium size room that has a temperature of about $25°C$. Different types of sound were placed at different distances (1m, 2m and 3m) to test the ability of the system to accurately detect the position of a sound source under different situations. Sound signals used to test the model are white noise which is extended in time and frequency, claps (similar to an impulse) which is local in time and extended in frequency, instantaneous pure tone (similar to sine wave) which is local in time and frequency, and finally a continuous pure tone which is extended in time and local in frequency. To get the measurements for different angles we placed the head on a servo motor and rotated the head in respect to the sound source. Measurements were taken at every 10° on the left and the right side of the microphones. Those measurements were repeated several times to ensure the stability of the system. For these tests the following parameters are known: $F_s = 44100\ Hz$, $c = 346\ m/s$ at room temperature

$(25°C)$, $l = 0.14\ m$ and n is computed from the cross-correlation of the left and the right signals. Finally θ is computed from Eq. (2) and Eq. (3).

In the beginning, the microphones are fixed in space back to back without having any barrier between them (free-field sound measurement). Figure 2 shows a plot of measured angle vs actual angle, while Table 1 shows the relative accuracy of each angular measurement. The average accuracy is 96.7%. The accuracy of our measurement is similar to the results published in the literature. For example, Murray *et al.* [8] implemented a system that have an accuracy of about 92%. In addition, Eq. (2) and Eq. (3) shows that increasing the interaural distance l increases the ITD. Therefore, we increased l from 14 cm to 45 cm. This process result in increasing the accuracy to 98.7% and enables the system to measure angles beyond 70 degree.

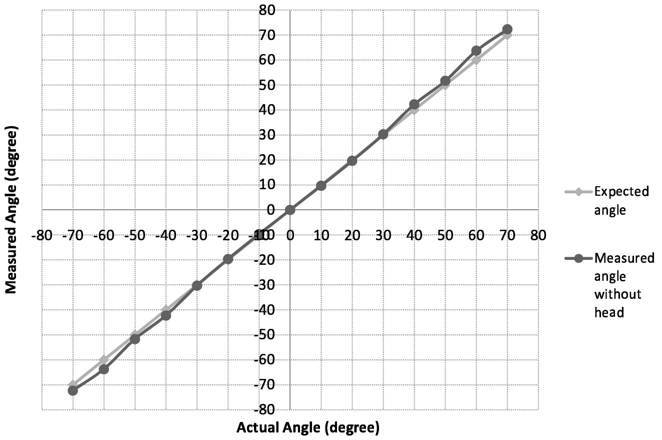


Fig. 2. Measured angle Vs Actual angle (without head)($\sigma = 0$).

After testing the ITD model in free-field environment, the microphones were fixed on an iCub head. The distance between the two microphones is approximately 14 cm. Measurements at every 10° were recorded and listed in Table 2. The accuracy of the measurement reduces as the angle increase above 20° and −20°. This is because the head covers the direct line between the source and the further microphone, resulting in reflection and diffraction, which in turn increases the time delay more than the expected value. Similarly, the measurement were repeated 5 times at every 10° to make sure that the results are repetitive. The average of the five measurement is

Table 1. Measured angles for $l = 0.14\ m$ (free space measurement).

Delay samples				
Cross Correlation	**Mathematically Calculated**	**Actual Angle**	**Calculated Angle**	**Accuracy %**
17	16.76	-70	-72.3	96.7
16	15.45	-60	-63.72	93.8
14	13.66	-50	-51.68	96.6
12	11.46	-40	-42.26	94.3
9	8.92	-30	-30.28	99.06
6	6.10	-20	-19.64	98.2
3	3.09	-10	-9.67	96.7
0	0	0	0	100
3	3.09	10	9.67	96.7
6	6.10	20	19.64	98.2
9	8.92	30	30.28	99.06
12	11.46	40	42.26	94.3
14	13.66	50	51.68	96.6
16	15.45	60	63.72	93.8
17	16.76	70	72.3	96.7

shown in Fig. 4. In addition, ITD is plotted against azimuth angle in order to understand the effect of the head on the accuracy of the measurement as shown in Fig. 3. It is obvious that the head increases the time delay, making a big difference between it and the theoretical measurement that agrees with the measurement taken in free space.

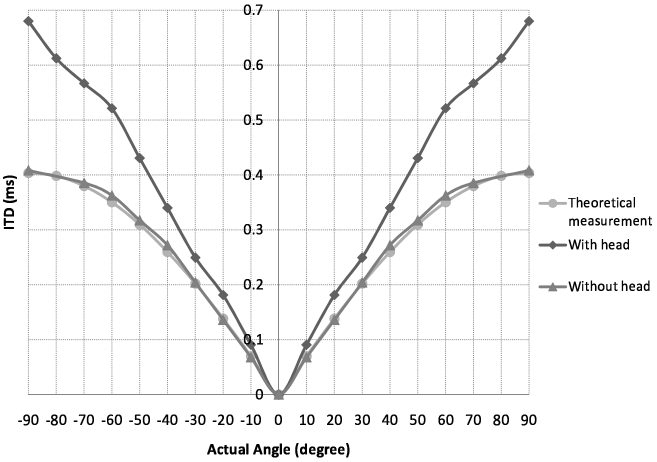


Fig. 3. Interaural Time Difference Vs Actual angle.

Introducing the head reduces the average accuracy to 68%. We have addressed this problem by adjusting the ITD by dividing it by a constant that compensates for the diffraction and reflection effects. For our case we have selected a constant equal to 1.35 to be divided by ITD in Eq. (theta2). This constant is computed from the average of ratios of ITD at different angles (ITD with head divided by without head). This method increases the accuracy of the system from 68% to 97% and enables the system to measure accurately above 20° and $-20°$ as shown in Table 3.

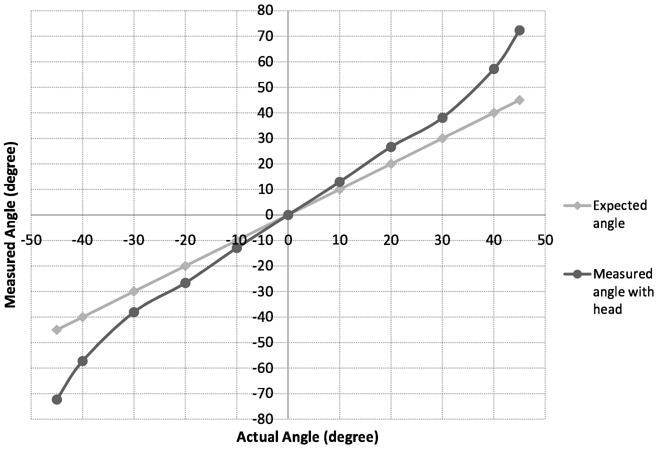


Fig. 4. Measured angle Vs Actual angle (with head)($\sigma = 0$).

The model explained in this section is only capable of localizing a sound source in azimuth angle. Therefore, we need another model that can localize a sound source in elevation direction. A model that can measure elevation angle will be explained in next section.

3. Spiral Ear Model

Before explaining the model it is necessary to define the Head Related Transfer Function (HRTF), since it is the data used to measure elevation angle. "HRTFs capture the sound localization cues created by the scattering of incident sound waves by the body" [9]. It was found that humans can localize sound in the median plane relying on the filtering of a sound due to the head, torso, and external ear, or pinna. The question is what spectral frequencies or combination of frequencies correspond with which location

Table 2. Measured angles for $l = 0.14\ m$ with iCub head.

Delay samples		Actual Angle	Calculated Angle	Accuracy %
Cross Correlation	**Mathematically Calculated**			
19	13.66	-50	—	—
17	12.61	-45	-72.3	39.3
15	11.46	-40	-57.2	57
11	8.92	-30	-38.05	73.1
8	6.10	-20	-26.63	66.85
4	3.09	-10	-12.95	70.5
0	0	0	0	100
4	3.09	10	12.95	70.5
8	6.10	20	26.63	66.85
11	8.92	30	38.05	73.1
15	11.46	40	57.2	57
17	12.61	45	72.3	39.3
19	13.66	50	—	—

Table 3. Measured angles for $l = 0.14\ m$ with iCub head after altering Eq. (2).

Delay samples		Actual Angle	Calculated Angle	Accuracy %
Cross Correlation	**Mathematically Calculated**			
—	20.97	-60	—	—
19	18.55	-50	51.68	96.64
16	15.56	-40	41.35	96.62
12	12.10	-30	29.70	99
8	8.28	-20	19.29	96.4
4	4.20	-10	9.50	95
0	0	0	0	100
4	4.20	10	9.50	95
8	8.28	20	19.29	96.4
12	12.10	30	29.70	99
16	15.56	40	41.35	96.62
19	18.55	50	51.68	96.64
—	20.97	60	—	—

in space. The answer to this question will be clear in this section.

The pinna is an important object for localizing a sound source in the elevation direction. This is because the pinna spectral features like peaks and notches in HRTFs are caused by the direction dependent acoustic filtering due to the pinna [10]. Notches are created in the frequency spectra when incident wave is cancelled by the reflected wave from the artificial concha. Many researchers used an artificial ear that has a spiral shape [7,10–12]. This is because a spiral shaped artificial ear is a special type of pinna that can vary the distance from a microphone placed in the centre of the spiral to a concha wall linearly according to the sound direction. Therefore, this model is capable of producing notches at different frequencies for different elevations and the position of the notch is expected to vary linearly with elevation. “A notch is created when a quarter of the wavelength of the sound λ plus any multiple of $\lambda/2$ is equal to the distance d between the concha and the microphone” [7]:

$$n * \frac{\lambda}{2} + \frac{\lambda}{4} = d(n = 0, 1, 2...) \tag{4}$$

For these wavelengths, the incident wave are cancelled by reflected waves. Thus notches will appear at the following frequencies [7]:

$$f = \frac{c}{\lambda} = \frac{(2 * n + 1) * c}{4 * d} \tag{5}$$

The spiral ear model shown in Fig. 6 is similar to the ITD model in the first few blocks. After passing the Threshold block, the power spectra density is calculated for the right and left samples using the Welch spectra [13]. An example of the output of the power spectra density $H_l(f)$ and $H_r(f)$ is shown in Fig. 5. From this output it is difficult to extract notches, since they disappear in the complex spectra. To solve this problem it is essential to calculate the Interaural spectral difference as [7]:

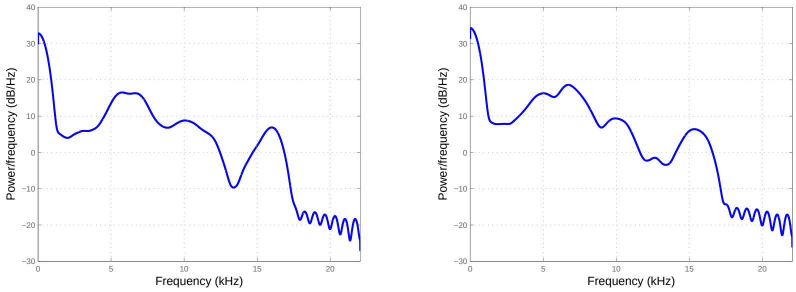
$$\Delta H(f) = 10 \log_{10} H_l(f) - 10 \log_{10} H_r(f) = 10 \log_{10} \frac{H_l(f)}{H_r(f)} \tag{6}$$

Before doing that it is recommended to have a small asymmetry between the two ears. This asymmetry is used to amplify the notch at the left ear, by having a maximum at the same frequency at the right ear. This can be done by selecting the distance for the right ear d_r as [7]:

$$d_r(\phi) = 2\frac{m_r + 1}{2 * n_l + 1} * d_l(\phi) \tag{7}$$

where m_r=maxima number for right ear, n_l=notch number for left ear, and d_l=distance between the microphone and concha for left ear.

However, using two identical ears makes it difficult to have a maximum at the right ear at the same frequency that the left ear has a notch for all elevation. Therefore, the best solution is to have a difference in angle between the two ears that is sufficient to give a maxima at the right ear at the same frequency that the left ear has a notch when the sound comes



(a) (b)

Fig. 5. Power spectral density of the (a) right ear and (b) left ear.

from the front. The optimum angle for the spiral ears used in this model is 18° [7].

Then a 14 degree polynomial is fitted to the interaural spectral difference as shown in Fig. 7. The fitted curve is used to make the system more stable, since the frequency of the notch changes alot in the interaural spectral difference and becomes more stable in the fitted curve. The minima in the fitted curve may not exactly correspond to the notch frequency, but it changes with the elevation of the sound source. Finally, the minima is easily extracted from the fitted curve and is compared to a set of notch frequencies stored in a database to specify the elevation angle. This database contains the notch frequency for different positions in space specified in terms of azimuth and elevation angle.

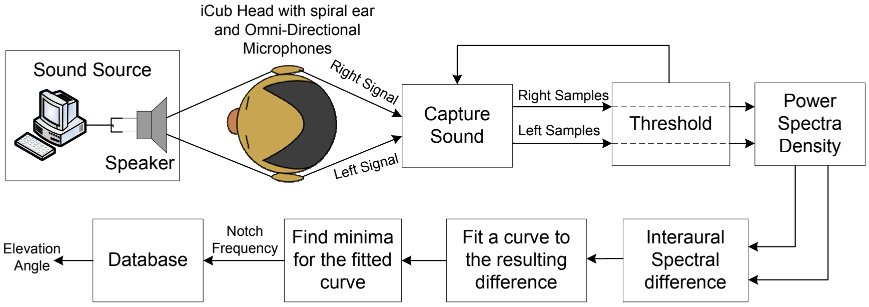


Fig. 6. Spiral ear system model.

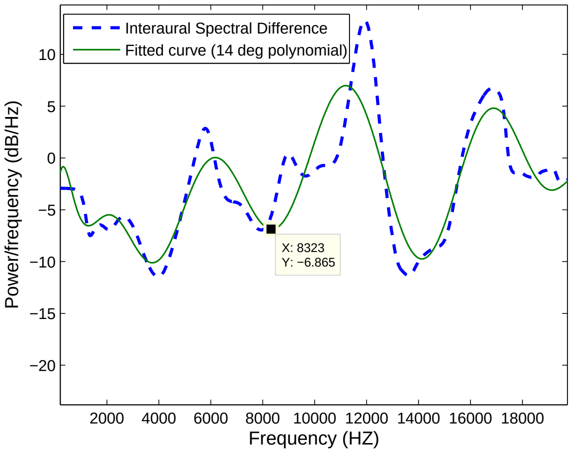


Fig. 7. Interaural spectral difference and a fitted curve.

To test the spiral ear model, a sound source was placed $1m$ away from the iCub head (with spiral ears). To get the measurements at different azimuth and elevation angles we placed the head on a servo motor and rotated the head in respect to the sound source. Measurement was taken at every $5°$ from $-30°$ to $30°$ in the horizontal plane and from $-30°$ to $45°$ in the vertical plane, which counts up to 208 positions in space. We have selected the first minima to be at 6000 Hz, which corresponds to the second notch. This is not always the case, since different type of ears will require a different frequency range.

Five measurement of the notch frequency were taken at different position in space and the average were plotted as shown in Fig. 8. Figure 8 shows the mean of the five measurement at specific position in space, but it dose not show the variation in notch frequency. Therefore, in Fig. 9 we plotted the mean and the standard deviation of the notch frequency at azimuth $20°$ and $-25°$. From Fig. 9 we can see that the accuracy of model is higher when the sound source is above the head and it becomes lower as the sound source moves below the head. The is because we are working in a normal acoustic environment where reverberation and noise can degrade the performance of the model. Also the accuracy of measurement differs for different azimuth angles. To asses the accuracy of the system, the azimuth angle were fixed at $20°$ and the measurement were repeated 5 times at every $5°$ from $-30°$

to 45° in the vertical plane. Then the measured elevation angle is specified by comparing the measured notch frequency with the notch frequencies stored in the database. The comparison method can be based on the closer notch frequency or on the maximum likelihood. The error in the estimated elevation angle is mostly less than 5 degree as shown in Fig. 10. Also we have changed the degree of the polynomial from 14 to 12 [7] and we have found that there is no significant difference in terms of accuracy and repeatability.

Finally, we can say that the spiral ear model and the ITD model can work together to measure the position of a sound source in two dimensions using two microphones connected on an iCub head that have a spiral shaped ears.

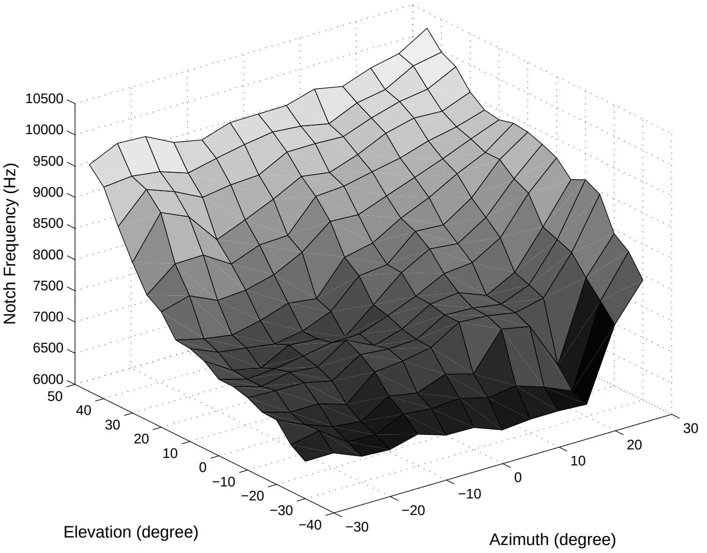
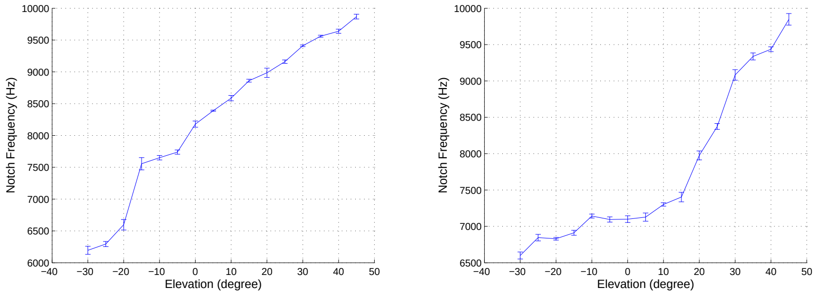


Fig. 8. Average notch frequency at different positions in space.

4. Conclusion

In this paper we have presented two models that are used to localize the direction of a sound source in space. The ITD model is used to measure azimuth angle and the spiral ear model is used to measure elevation an-



(a) Azimuth = 20°

(b) Azimuth = −25°

Fig. 9. Mean and the standard deviation of the notch frequency.

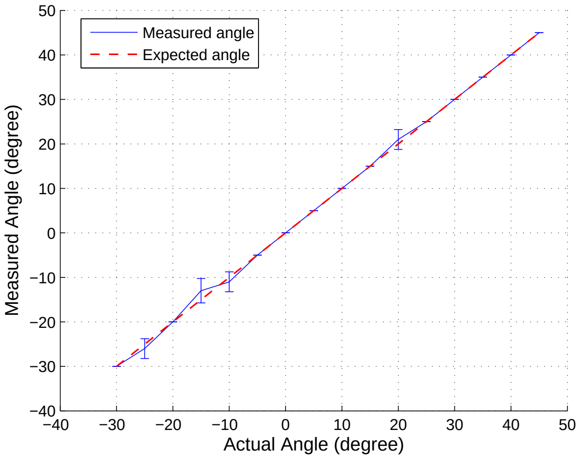


Fig. 10. Measured elevation angle Vs Actual angle at azimuth=20°.

gle. It was found that the accuracy of the ITD model depends on several factors such as the distance between the microphones and the usage of a head. Increasing the distance between the two microphones increases the accuracy from 96.7% at $l = 14$ *cm* to 98.7% at $l = 45$ *cm*, and the range of measurement, enabling the model to measure up to 80 degree to the right and to the left of the microphones. However, fixing the microphones on the head reduces the accuracy to 68% and we have shown that it is possible to increase the accuracy to 97% by manipulating Eq. (2) to compensate for the effect of the head. The spiral ear model accuracy changes for different azimuth angles and the error in the estimated elevation angle is mostly less than 5 degree.

Future work that can be done to improve the models in this paper, include solving the problem of front back ambiguity without rotating, enabling the system to detect multiple sound sources at the same time, and making the system robust even in an environment with high noise level and reverberation. Also the human sound is unique and it can be used to identify the gender and the person identity. This kind of information will be very useful for human robot interaction.

References

1. A. H. Dekker, *C4ISR Architectures, Social Network Analysis and the FINC Methodology: An Experiment in Military Organizational Structure*, tech. rep., DSTO Electronics and Surveillance Research Laboratory (Edinburgh, South Australia, 2002).
2. D. H. Cropley, *SESA/INCOSE SE98* , 1 (1998).
3. B. Shah and H. Choset, *Survey on Urban Search and Rescue Robotics*, tech. rep., Carnegie Mellon University (Pittsburgh, 2003).
4. J. M. Loomis, J. R. Marston, R. G. Golledge and R. L. Klatzky, *Jonrnal of Visual Impairment and Blindness* , 219(April 2005).
5. K. Dautenhahn and I. Werry, *Pragmatics & Cognition* **12**, 1 (2004).
6. M. Menegatti, E. Mumolo, M. Nolich and E. Pagello, *A Surveillance System based on Audio and Video Sensory Agents cooperating with a Mobile Robot*, tech. rep., Smartlab Department of Electrotechnics, Intelligent Autonomous Systems Laboratory (2004).
7. J. Hornstein, M. Lopes and J. Santos-Victor, *IEEE/RSJ international conference on intelligent robots and systems* , 1170(Oct 2006).
8. J. Murray, H. Erwin and S. Wermter, *Robotics Sound-Source localization and Tracking Using Interaural Time Difference and Cross-Correlation*, center for hybrid intelligent systems, University of Sunderland, Workshop on NeuroBotics (Sunderland, 2004).
9. V. Algazi, R. Duda and D. Thompson, *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics* , 99(Oct 2001).
10. Y. Park and S. Hwang, *16th IEEE International conference on Robot and Human interactive Communication* , 405(Aug 2007).
11. S. Hwang, Y. Park and Y. Park, *International Conference on Control, Automation and Systems* , 1906(Oct 2007).
12. S. Hwang, K. Shin and Y. Park, *IEEE Sensors* , 1460(Oct 2006).
13. P. D. Welch, *IEEE Transactions on Audio and Electroacoustics* **15**, 70(June 1967).