Sound Localization: Human Vs. Machine

W.G. Nuwan Jayaweera, A.G. Buddhika P. Jayasekara, A.M. Harsha S. Abeykoon
Department of Electrical Engineering
University of Moratuwa
Moratuwa, Sri Lanka
jayaweerawgn@yahoo.com, buddhika@elect.mrt.ac.lk, harsha@elect.mrt.ac.lk

***Abstract*—The Human shows a remarkable capability in localizing a sound source and navigating towards it. In the current context of robotic applications, machinery models have been developed, so that they can be used in sound source localization. But, it is not yet quantified the accuracy of human's sound source localization in different frequencies and distances at the free field. Thus, the aim of this paper is to estimate the error of human's sound source localization at different frequencies and distances on the horizontal plane and the paper presents the characteristics of ear by taking each individual's localization ability into consideration.**

An experiment is conducted to investigate the individual ability to predict the sound incident direction. Ten samples of asian young adults from the age group of 20 - 30 years are taken into the experiment and their responses for localization cues are recorded. The experiment is also conducted for different sound source locations such as 1m, 2m and 3m and different sound source frequencies of 1 kHz and 5 kHz. The results show the individual responses for the direction prediction and they are unique from each individual to the other. The average percentage errors for direction prediction at 1 kHz frequency sound signal give 0.20, 0.93 and 5.20 for 1m, 2m and 3m distances respectively. Also, the average percentage errors for direction prediction at 5 kHz frequency sound signal give 3.59, 1.68 and 0.52 for 1m, 2m and 3m distances respectively.

Keywords*— *ear sensation; error quantification; Head Related Transfer Function (HRTF); Inter-aural Level Difference (ILD); Inter-aural Time Difference (ITD)

I. Introduction

Among five sensations, human owes, the ear sensation plays a major role in localizing a sound source. At the same time, by following human ear sound source localization mechanism, several industrial applications such as mobile robot navigation, sound source tracking, automatic fall detection and underground explosions [1]-[2], have been invented and immensely being used in modern world. In the context of military, solders are in an essence of identifying different sound sources positioned in different azimuths and elevations in the free field. Hence, different hearing protection systems, in different configurations, have been tested to increase their sound source localization ability [3]. Thus, in order to develop sound source localization applications, it is very important to observe the characteristic of human auditory system including quantified measures [4].

In aforementioned machine related sound source localization industrial applications, there are two main methods, namely ITD and ILD, which have widely been applied in estimation of sound source localization cues. By associating those two concepts, few algorithms such as cross correlation [5]-[6], general cross correlation [5]-[6], maximum likelihood [5]-[6], adaptive least mean square filter [5]-[6] and Head Related Transfer function (HRTF) which is used Rayleigh's duplex theory [5]-[7] have been developed and even practiced nowadays.

In biological point of view, the complex configuration of human ear is with a remarkable ability in determining the sound source direction and distance, so called sound localization cues, at near field [8] and far field [8] as well. Thus, it is important to identify how accurately and preciously, a human can localize a sound source [9]. Sound, in fact, creates pressure waves in transmission medium such as air, water and solid. Due to the vibration of transmission medium molecules, the sound propagates in omni directional paths as pressure waves. The incident acoustic waves to the ear from different angles pass through the outer ear, middle ear, inner ear and inner hair cells to the auditory cortex of the brain via the auditory nerves. As a result, human predicts the sound localization cues [7].

As medical experts guess, there are six mechanisms of which sound source localization takes place in human auditory system. They are Inter-aural Time (or phase) Difference (ITD), Inter-aural Intensity (or level) Difference (IID), the action of pinna, movement of head, the loudness of the direct sound from the sound source compared to the level of reverberation and the distance to the sound source [7]. Even though, the human owes a complex multifunctional signal processor, it can be noticed that the human is less accurate compared to computer processor based sound source localization systems. Thus, no references found yet to explain human sound source localization behavior at the brain.

Hartmann at el [11], have addressed in their studies the humans' ability of localizing a sound source in free field and in the closed room environments with or without the reverberation present [3]-[12]-[13]. But they have not quantified the error of predicted values of direction over those of actual values. The aim of this research is therefore to identify the error between actual values of the direction over the predicted values of it while the frequency dependency of the error is analyzed.

The organization of the paper is as follows. Section II, machine's sound localization algorithms are explained. Section III explains the experiments and the experimental results. Finally, the paper is concluded in Section IV.

II. Machine's Sound Localization Algorithms

The sound localization cues appear at the input signal receivers are in two major categories known as monaural and binaural cues [8]-[14]. The theory behind the estimation of HRTF is discussed under monaural cues, while five different ITD algorithms together with ILD algorithm are elaborated under binaural cues.

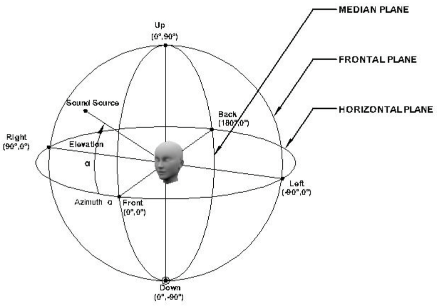


Fig. 1. Typical Sound source localization coordinates

In 2D and 3D sound source localization analysis, the coordinates, which is generally used, consist of horizontal plane, frontal plane and median plane as shown in *Fig. 1* and those coordinates are herein referenced in giving the sound source orientation [8]-[9].

A. Monaural Cues

These cues can be extracted from the received signal at one acoustic receiver and then HRTF is used in estimating the localization cues. HRTF measures the frequency response for different sound incident angles in the spatial location, with respect to the sound appeared at pinna.

The processing algorithm for HRTF can be realized as follows. Let *s(n)* be the known stimulus signal presented at azimuth *Θ* and elevation *φ*. *C(n)* is the known Common Transfer Function (CTF) and $d_{l\Theta\varphi}(n)$, $d_{r\Theta\varphi}(n)$ are the left and right ear Directional Transfer Functions (DTFs) respectively. $h_{l\Theta\varphi}$ *(n)*, $h_{r\Theta\varphi}$ *(n)* are the estimated left and right ear HRTFs. Thus, the equivalent expressions from frequency domain [10],

$$H_{l\,\mathrm{or}\,r,\theta,\varphi}(k) = S(k)_{*}C(k)_{*}D_{l\,\mathrm{or}\,r,\theta,\varphi}(k) \tag{1}$$

Here, it is assumed that the *C(n)* is spatially invariant. Hence the right and left ear DTFs can be estimated as,

$$\left|D_{l\,\mathrm{or}\,r,\theta,\varphi}(k)\right| = \frac{\left|H_{l\,\mathrm{or}\,r,\theta,\varphi}(k)\right|}{\left|S(k)\right|\left|C(k)\right|} \tag{2}$$

$$\angle D_{l\,\mathrm{or}\,r,\theta,\varphi}(k) = \angle D_{l\,\mathrm{or}\,r,\theta,\varphi}(k) - \angle S(k) - \angle C(k) \tag{3}$$

$$D_{l\,\mathrm{or}\,r,\theta,\varphi}(k) = \lfloor D_{l\,\mathrm{or}\,r,\theta,\varphi}(k) \rfloor \exp[j\angle D_{l\,\mathrm{or}\,r,\theta,\varphi}(k)] \tag{4}$$

From the phase information of the computed DTFs, ITD can be calculated, estimating the cross-correlation between $d_{l\Theta\varphi}(n)$ and $d_{r\Theta\varphi}(n)$

$$n_{ITD,\theta T\varphi} = \arg_{\tau}\max\sum_{n} d_{l,\theta,\varphi}(n)d_{r,\theta,\varphi}(n+\tau) \tag{5}$$

B. Binaural Cues

Binaural cues are extracted from the received signals at both sound receivers and hence sound localization is done. There are two kinds of binaural cues known as Interaural Time Differences (ITD) and Interaural Level Differences (ILD) [8].

1) Interaural Time Difference (ITD): When the sound incidents at two receivers which are apart from each other, the time taken to reach sound signal to receivers results in a time difference called ITD. With reference to the ITD estimation, five well known algorithms so called cross-correlation, general cross-correlation, maximum likelihood, average square difference function and least mean square adaptive filter methods can be elaborated as follows,

a) Cross-correlation (CC) method: The cross-correlation function between two received signals $r_1(t)$ and $r_2(t)$ present at the two microphones is calculated and then located the maximum peak of the output response to estimate the time delay D_{CC}. The additive noise at each microphone is identified as $n_1(t)$ and $n_2(t)$ [6]-[8].

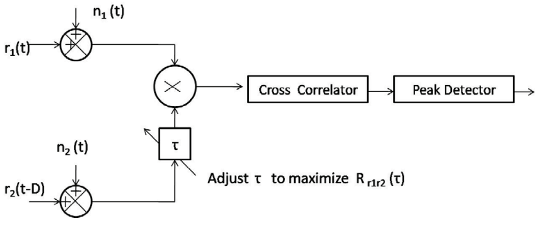


Fig. 2. Cross-correlation processor

$$R_{r1r2}(\tau) = E[r_1(t)r_2(\tau)] \tag{6}$$

$$D_{cc} = \arg\max_{t}[R_{r1r2}(\tau)] \tag{7}$$

CC method gives the median error greater than 3.5 [15]

b) Generalized Cross-Correlation (GCC) Method: To sharpen the cross correlation peak, weighting function $w_P(f)$ is used. This eliminates spreading of the peak of correlation function and hence more accurate than cross correlation. This is also referred to as phase transform (PHAT) [1]. The $G_{r1r2}(f)$ is the cross spectrum of received signals [6] –[8].

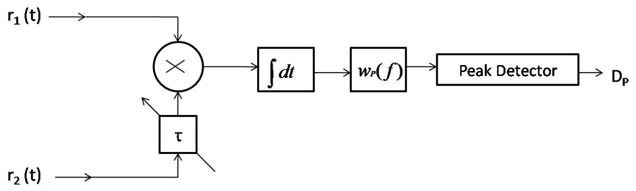


Fig. 3. Generalized cross-correlation processor

$$R_{r1r2}(\tau) = \int_{-\infty}^{+\infty} w_P(f) G_{r1r2}(f) e^{j2\pi 2\pi} df \tag{8}$$

$$w_P(f) = \frac{1}{|G_{r1r2}(f)|} \tag{9}$$

$$D_P = \arg\max_t [R_{r1r2}(\tau)] \tag{10}$$

GCC gives the abs. error less than 0.2^0 for angle of arrival [1].

c) Maximum Likelihood (ML) method: This is an enhanced method from GCC family and gives the maximum likelihood solution for ITD. The weighting function $w_{ML}(f)$ is chosen to improve the estimated delay D_{ML} by attenuating the signals fed into the correlator in spectral region where the Signal to Noise Ratio (SNR) is lowest [16].

$$R_{r1r2}(\tau) = \int_{-\infty}^{+\infty} w_{ML}(f) G_{r1r2}(f) e^{j2\pi 2\pi} df \tag{11}$$

$$w_{ML}(f) = \frac{1}{|G_{r1r2}(f)|} \frac{|\gamma_{r1r2}(f)|^2}{1 - |\gamma_{r1r2}(f)|^2} \tag{12}$$

$$D_{ML} = \arg\min_t [R_{r1r2}(\tau)] \tag{13}$$

The maximum likelihood also achieved good results with absolute value of error less than 2.5 degrees [15].

d) Average Square Difference Function (ASDF) method: This method estimates the position of the minimum error square between the two received noisy signals $r_1(n)$ and $r_2(n)$. Then this position value is considered as the estimated time delay D_{ASDF}.

$$R_{ASDF}[\tau] = \frac{1}{N} \sum_{n=0}^{N-1} [r_1(n) - r_2(n+\tau)]^2 \tag{14}$$

$$D_{ASDF} = \arg\min_t [R_{ASDF}(\tau)] \tag{15}$$

e) Least Mean Square (LMS) adaptive filter method: This is a Finite Impulse Response (FIR) filter which automatically adapts to its coefficient to minimize the mean square difference between its two inputs. It comprises of reference and desired signals $r(n)$ and $s(n)$ respectively. The LMS filter response is,

$$y(n) = W^T(n)X(n) \tag{16}$$

Where T denotes the transpose and $X(n)$ is the filter state consisting of most recent samples of $r(n)$. The $W(n)$ is the L-vector of filter weights at instant n. The error output is,

$$e(n) = s(n) - W^T(n)X(n) \tag{17}$$

The weight vector is updated at every sample,

$$W(n+1) = W(n) + \mu e(n) X^*(n) \tag{18}$$

Where * represents the complex conjugate and μ is the feedback coefficient. The algorithm adapts the FIR filter to insert a delay equal and opposite to the existing between the two signals; in an ideal situation, the filter weight corresponding to the true delay would be unity and all the other weights would be zero [5]-[17].

Information on the mutual delay signals can be integrated into a representation called Coherence Measure (CM) and associated to a function $C(t,\tau)$.

$$C(n,l) = w(n,l_0) \tag{19}$$

Where $w(n,l_0)$ is the $W(n)$ component for lag l_0

The time delay related to lease mean square adaptive filter can be estimated as

$$D_{LMS} = \arg{}_l\max\left(\sum_{n=1}^{N} C(n,l)\right) \tag{20}$$

By using the algorithms explained, the time difference is calculated and by relating ITD to the human head related geometry as shown in *Fig. 4, Fig. 5* and *Fig. 6*, sound source direction is estimated [17].

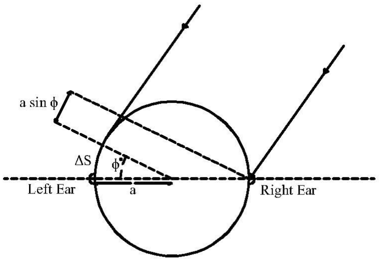


Fig. 4. Far field sound reception at both ears

For frequencies below approximately 500 Hz,

$$\tau = (2a/c) \sin\varphi \tag{21}$$

For frequencies above approximately 2 kHz

$$\tau = (a/c)\left(\varphi + \sin\varphi\right) \tag{22}$$

Where, c is the velocity of sound in air

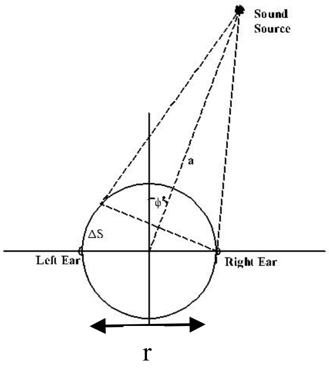


Fig. 5. Close sound source to head - direct path only to one ear

sinφ ≤ r/2a

$$\Delta S = \left[\left(n+\frac{1}{2}\right)\cos\varepsilon + \frac{1}{2}(\varphi+\varepsilon) - \sqrt{n^2+n+\frac{1}{2}-\left(n+\frac{1}{2}\right)\sin\varphi}\right] \quad (23)$$

$$n = \frac{a - r/2}{r} \text{ and } \varepsilon = \arcsin\left(\frac{r}{2a}\right) = \arcsin\left(\frac{1}{1+2n}\right)$$

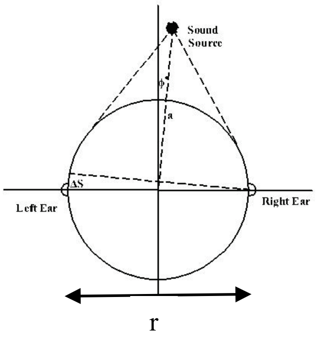


Fig. 6. Close sound source to head - no direct path to either ear

sinφ ≤ r/2a

$$\Delta S = r.\varphi \quad (24)$$

2) Interaural Intensity Difference (IID): The acoustic shadow effect is significant at higher frequencies and hence there exists an intensity difference between two recievers.

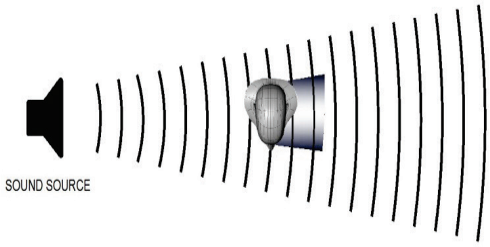


Fig. 7. Acoustic shadow effect

According to the inverse-square-law, the signal received by the microphone can be modeled as [18]. *s(t)* and *d* are the sound signal and the distance to the sound source respectively.

$$x(t) = s(t)/d + \xi(t) \quad (25)$$

The energy received by the microphone can be obtained by integrating the square of the signal over this time interval,

$$E_i = \int_0^W x^2(t)dt = \int_0^W [s^2(t)/d^2 + \xi^2(t)]dt$$
$$= \frac{1}{d^2}\int_0^W [s^2(t)dt + \int_0^W \xi^2(t)]dt \quad (26)$$

Given two microphones, the above equation leads to a simple relationship between the energies and distances,

$$E_1 d_1^2 = E_2 d_2^2 + \eta \quad (27)$$

Where, $\eta = \int_0^W [\xi_1^2(t) - \xi_2^2(t)]dt$ is a zero mean random variable if the variance of $\xi_i(t)$ is constant.

However, if only binaural cues are used in sound source localization some ambiguities occur in estimating localization cues due to several points at the cone which cause the same time difference and level difference as *Fig. 8* depicts. This is called cone of confusion. Thus, in estimating the spatial location at the elevation, monaural cues are used [8].

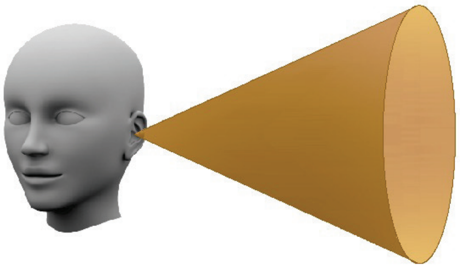


Fig. 8. Cone of confusion

In this paper, the authors estimate the error percentage to observe the error of predicted values of direction.

III. Experiments and Results

An experiment was conducted to investigate the humans' ability in localizing a sound source. In this experiment, 15 numbers of asian young adults in the age group of 20 to 30 years were selected arbitrary and then sited at the centre of the experimental area as shown in *Fig. 9*. 1 kHz and 5 kHz frequency distinct sounds were played at the distances of 1m, 2m and 3m in horizontal plane at 24 different unknown locations to the listener. The sound source was always played 1m above the ground level.

The predicted location of all participants was recorded and the actual values over predicted values of direction prediction with R^2 value were plotted as *Fig. 10* depicts. Then, 10 numbers of samples were selected considering the R^2 value of over 90% and 70% for 1 kHz and 5 kHz respectively.

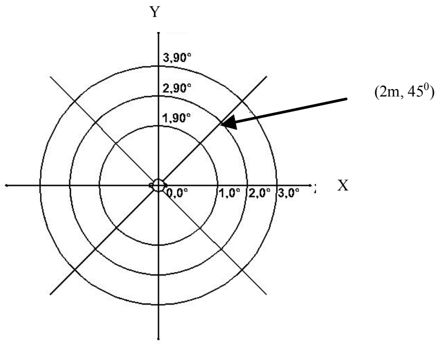


Fig. 9. Experimentation setup

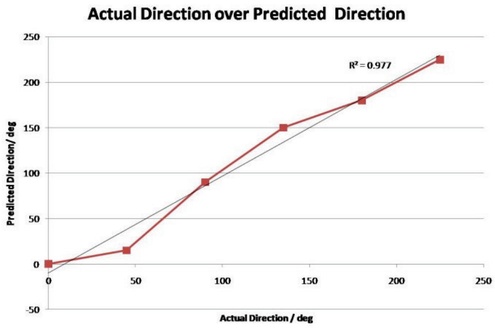
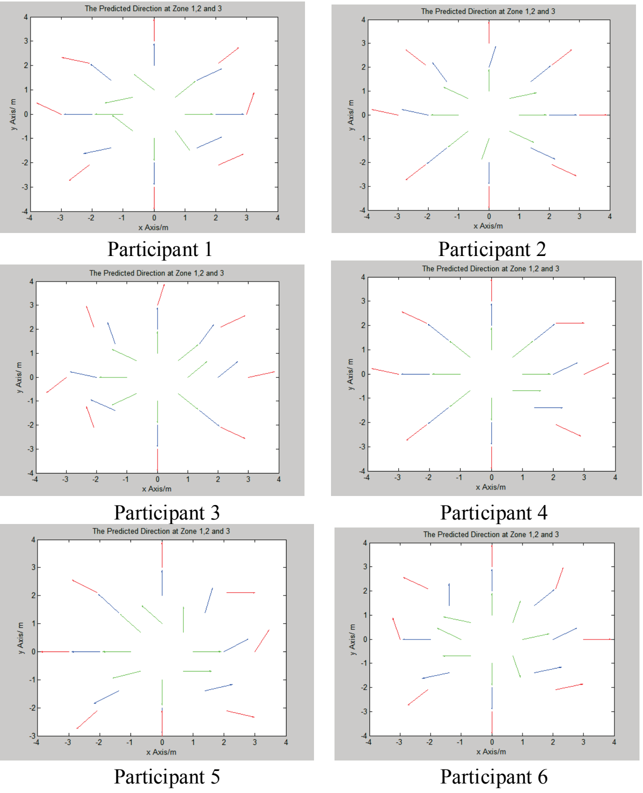


Fig. 10. Sound incident direction predicted results for human



Participant 1

Participant 2

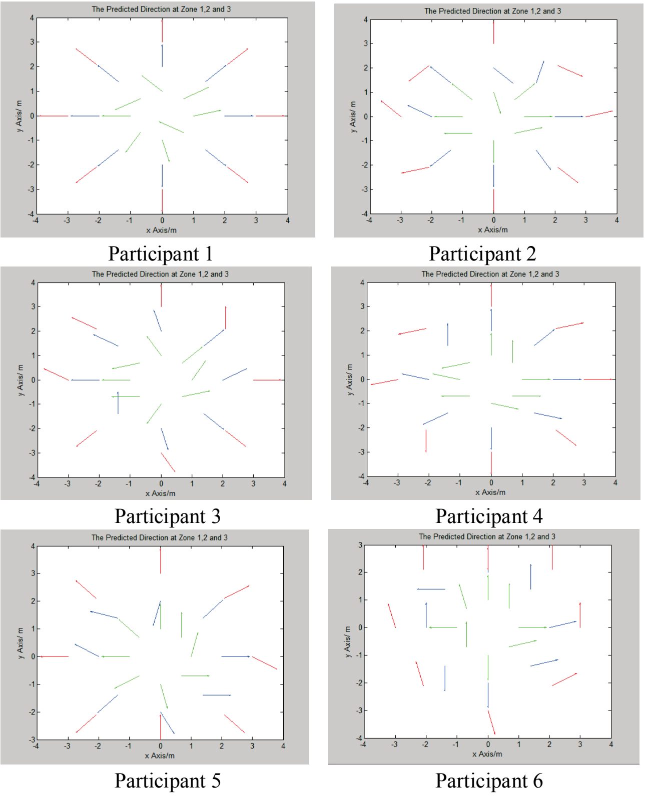
Participant 3

Participant 4

Participant 5

Participant 6

Fig. 11. Sound arrival direction prediction results for 1 kHz sound (1m, 2m & 3m distances in green, blue & red respectively)



Participant 1

Participant 2

Participant 3

Participant 4

Participant 5

Participant 6

Fig. 12. Sound arrival direction prediction results for 5 kHz sound (1m, 2m & 3m distances in green, blue & red respectively)

Fig. 11 and *Fig. 12* are graphical representations of the responses associated with 6 participants out of 10 for 1 kHz and 5 kHz sound signals which are originated at 1m, 2m and 3m distances (green, blue and red color respectively). It can be noticed that each individual has inherent ability to predict the direction. Some of them are capable of accurately predicting sound incident direction at lower frequencies than high frequencies and vice versa. In general, *Fig. 11* and *Fig. 12* show that participants' direction prediction results are more accurate at 1 kHz frequency than 5 kHz frequency.

Fig. 13 and *Fig. 14* depict the average percentage error for sound incident direction prediction responses of individuals for 1 kHz and 5 kHz frequency sound signals which are originated at 1m, 2m and 3m distances. Also, Table I tabulates the average percentage error for sound incident direction prediction responses of individuals for 1 kHz and 5 kHz frequency sound signals which are originated at 1m, 2m and 3m distances. In general, the results show that participants are accurate in predicting the direction at 1m distance for 1 kHz frequency sound signal while 5 kHz sound signal shows the same at 3m distance.

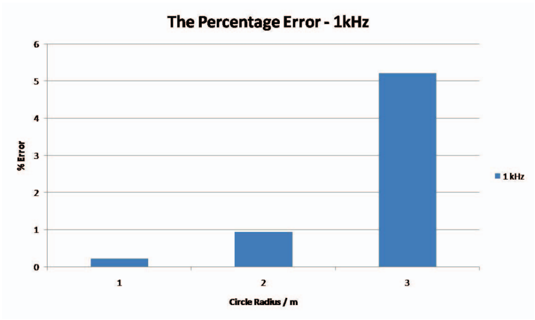


Fig. 13. The average percentage error for direction prediction at 1 kHz sound signal

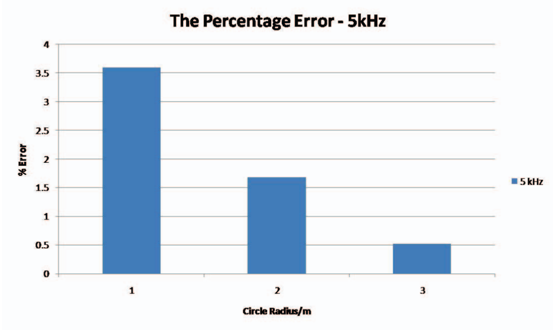


Fig. 14. The average percentage error for direction prediction at 5 kHz sound signal

TABLE I. AVERAGE PERCENTAGE ERROR FOR DIRECTION PREDICTION

Radius (m)	1 kHz	5 kHz
1	0.20	3.59
2	0.93	1.68
3	5.20	0.52

IV. CONCLUSION

This paper estimates the human's average percentage error for direction prediction for the sound sources with 1 kHz and 5 kHz frequencies. The sound signals are generated at 1m, 2m and 3m distances. With reference to the sound source localization algorithms developed for the machines, it can be noticed that the absolute error associated with machines in GCC method is less than 0.2 degrees. Whereas, the percentage errors calculated in the experiment shows that 1 kHz frequency sound signal, originated at 1m, 2m & 3m distance, has 0.20%, 0.93% and 5.20% errors respectively for direction prediction while 5 kHz frequency sound signal, originated at 1m, 2m & 3m distance, has 3.59%, 1.68% and 0.52% error for direction prediction.

It can further be noticed that the error related to 1 kHz frequency signal at 1m distance has the low error value while 5 kHz frequency signal shows the same at 3m distance. Also, the results reflect each individual's inherent ability in sound source localization. Some of them show that they accurately predict the direction at lower frequencies than higher frequencies and vice versa.

REFERENCES

[1] Pourmohammad, A., Ahadi, S.M., "Real Time High Accuracy 3-D PHAT-Based Sound Source Localization Using a Simple 4-Microphone Arrangement," *Systems Journal, IEEE* , vol.6, no.3, pp.455,468, Sept. 2012

[2] Yun Li; Ho, K.C.; Popescu, M., "A Microphone Array System for Automatic Fall Detection," Biomedical Engineering, IEEE Transactions on, vol.59, no.5, pp.1291, 1301, May 2012 doi: 10.1109/TBME.2012.2186449

[3] Sharon M. Abel, Stephen Boyne, Heidi Roesler-Muironey "Sound localization with an army helmet worn in combination with an in-ear advanced communications system", 2009 Oct-Dec;11(45):199-205.

[4] Huettel, Lisa G.; Collins, L.M., "Using computational auditory models to predict simultaneous masking data: model comparison," Biomedical Engineering, IEEE Transactions on, vol.46, no.12, pp.1432, 1440, Dec. 1999

[5] Yushi1 Zhang and Waleed2 H. Abdulla "A Comparative Study of Time-Delay Estimation Techniques Using Microphone Array

[6] Ali Pourmohammad and Seyed Mohammad Ahadi, "TDE-ILD-HRTF-Based 2D Whole-Plane Sound Source Localization Using Only Two Microphones and Source Counting," *International Journal of Information and Electronics Engineering,* vol. 2, no. 3, pp. 307-313, 2012

[7] David J M Robinson, "The human auditory system" Available: http://www.mp3tech.org/programmer/docs/human_auditory_system.pdf

[8] Laurent Calmes, "Biologically Inspired Binaural Sound Source Localization and Tracking for Mobile Robots" pp.11-15, 107-108, Dec 23, 2009.

[9] Tomasz Letowski and Szymon Letowski (2011) "Localization Error: Accuracy and Precision of Auditory Localization", Advances in Sound Localization, Dr. Pawel Strumillo (Ed.), ISBN: 978-953-307-224-1, InTech, Available: http://www.intechopen.com/books/ advances-in-sound-localization/ localization- error- accuracy and-precision-of-auditory-localization

[10] Corey I. Cheng and Gregory H. Wakefield "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space" J Audio Eng Soc, Vol 49, No 4, 2001

[11] W.M. Hartmaan, "Localization of sound in rooms" J. Acoust. Soc. Am. 74, 1380 1391, July 13, 1983

[12] Antje Ihlefeld and Barbara G. Shinn-Cunningham "Effect of source spectrum on sound localization in an everyday reverberant room" Hearing Research Center, Boston University, Boston, May 12, 2011, Pages: 324–333s"

[13] Barbara G. Shinn-Cunningham, Norbert Kopco and Tara J. Martin "Localizing nearby sound sources in a classroom: Binaural room impulse responses", J. Acoust. Soc. Am., Vol. 117, No. 5, May 2005

[14] Tobias Rodemann, G¨okhan Ince, Frank Joublin, and Christian Goerick "Using Binaural and Spectral Cues for Azimuth and Elevation Localization" IEEE/RSJ International Conference on Intelligent Robots and Systems, Sept, 22-26, 2008

[15] Hasan Khaddour "A Comparison of Algorithms of Sound Source Localization Based on Time Delay Estimation" VOL.2, NO.1, APRIL 2011.

[16] Shun Chi Wu; Swindlehurst, A.L.; Wang, P.T.; Nenadic, Z., "Efficient Dipole Parameter Estimation in EEG Systems with Near-ML Performance," Biomedical Engineering, IEEE Transactions on, vol.59, no.5, pp.1339, 1348, May 2012

[17] M. Omologo and P. Svaizer: "Acoustic Event Localization Using a Cross power-Spectrum Phase Based Technique", IEEE Acoustics, Speech, and Signal Processing, vol.2 pp.273-276, April 1994

[18] Birchfield, S.T.; Gangishetty, R., "Acoustic localization by interaural level difference," Acoustics, Speech, and Signal Processing, 2005 Proceedings. (ICASSP '05). IEEE International Conference on, vol.4, no., pp.iv/1109,iv/1112 Vol. 4, 18-23 March 2005