The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Binaural Sound Localization based on Sparse Coding and SOM

Hong shik Kim, Jongsuk Choi

Abstract— **Many kinds of sound source localization systems have been developed for detecting a direction of sound source. They are commonly using time delay of arrival (TDOA) or interaural time difference (ITD) algorithm for sound source localization where, especially, the ITD is the difference in arrival time of a sound between two ears. It is largely changed depending on frequency components of sound even though the sound source is located in the same place. In this paper we propose a binaural sound localization system using sparse coding based ITD (SITD) and self-organizing map (SOM). The sparse coding is used for decomposing given sounds into three components: time, frequency and magnitude. Moreover we estimate the azimuth angle through the SOM. This localization system is installed in our robot that has two ears, head and body. We use PeopleBot as a body of the robot.**

I. Introduction

THERE are many kinds of sound source localization systems have been developed for detecting a direction of sound source. Generally, sound source localization systems use time delay of arrival (TDOA) [1] or interaural time difference (ITD) [2] as a cue to estimate azimuth angle. Especially the ITD is the difference in the arrival time of a sound source between two ears. It changes depending on frequency components of a sound source even though the source is located in same place. Hence, we have to calculate several ITDs corresponding to different frequency bins which are chosen by Gammgtone Filterbanks and equivalent rectangular bandwidth (ERB) Filter cochlear model [8]. For the purpose of ITD calculation, most approaches [11] [12] use short time frequency analysis based on FFT. However, it has a side effect that small time shift can produce large changes in the representation where a particular sound event falls within a block. This problem causes errors in the result of sound source localization. To overcome this problem, our research employs sparse coding [3] which can decompose a sound signal into three components: time, frequency and magnitude. And the self-organizing map (SOM) [4] is applied to the results of the sparse coding to make the related SITD map. After performing neuron's learning in the SOM, we can estimate the azimuth angle of sound source through the SITD map with a good accuracy. This localization system is installed in a robot that has two ears and head. And we use PeopleBot of ActivRobots Inc. as a body of this robot.

Hong-shik Kim is with Korea Institute of Science and Technology, Seoul, Republic of Korea (e-mail: hskim@kist.re.kr).

Jong-suk Choi is with Korea Institute of Science and Technology, Seoul, Republic of Korea (phone: +82-2-958-5618; fax: +82-2-958-5629; e-mail: cjs@ kist.re.kr).

II. System Configuration

A. Overview

Our robot is composed of two ears, head and body as shown in Fig 1. We use KEMAR ear and the head is manufactured in Korea Institute of Science and Technology (KIST). An 8-channel analog to digital converter (TD-BD-8CSUSB, Tokyo Electron Device Ltd.) is installed in the robot. The body of this robot is PeopleBot.

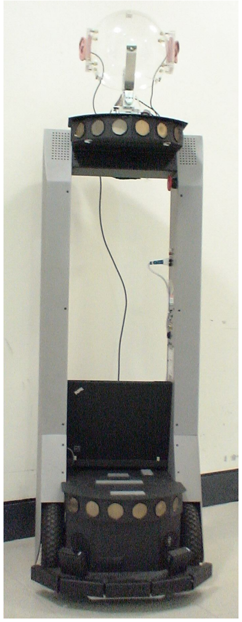


Fig. 1. The robot that has a head, two ears and body

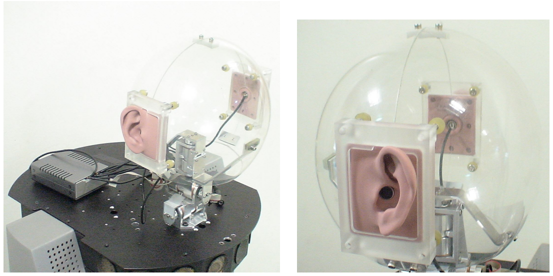
B. Body Part

We used PeopleBot as a body of this robot. It is an intelligent mobile robot specially designed and equipped for human-robot interaction research and application. The software platform to operate PeopleBot is Advanced Robotics Interface for Applications (ARIA) [5] which is developed by Mobile Robots Inc. The ARIA is a C++-based open-source development environment. This robot has 500-tick encoders provide 1% dead reckoning error, corrected with gyroscope. These differential drive platforms are highly holonomic and can rotate in place moving both wheels, or can swing around a stationery wheel in a circle of 32cm radius.

978-1-4244-3804-4/09/$25.00 ©2009 IEEE

C. Head and Ears Part

We designed the robot head such that it can move like human neck such as flexible neck solution [6]. It moves three degrees of freedom (Pan, Tilt, Swing). Two microphones that located in each ear are used to hear the sounds. Fig. 2 shows the real shape of the robot head and ear.



(a) Robot head (b) ear and microphone

Fig. 2. Head and ears.

D. A/D converter

In order to acquire digitally converted sound data, we used an 8-channel analog to digital converter (TD-BD- 8CSUSB, Tokyo Electron Device Ltd.). The feature of TD-BD-8CSUSB is like this,

1) Simultaneous in-phase sampling: Up to 8ch microphone inputs.
2) Power is provided for ECM microphones by USB2.0 or USB1.1 interface from a PC
3) Board size as small as a business card. (93mm x 56mm x 13mm)
4) Operates with USB bus power
5) Available sampling frequencies : 48KHz, 32KHz, 16KHz, 8KHz, 44.1KHz, 22.05KHz, 11.025KHz
6) Microphone input amplifier, built-in ADC gain control. SNR is 60dB or higher

Using this A/D converter we can easily get digital values of sound though USB communication port. Fig. 3 shows the TD-BD- 8CSUSB.

Fig. 3. 8 Channel A/D converter board with microphone cables

III. Sparse Coding

A. Sparse & Kernel Representation

We used a sparse and shiftable Kernel method of sound signal representation [7] since this method can decompose a sound signal into three components: time, frequency and magnitude. In this method, the sound signal $x(t)$ is encoded with a set of kernel functions that can be positioned arbitrarily and independently in time. The sound signal will be expressed as follows.

$$x(t) = \sum_{m=1}^{M} \sum_{i=1}^{n_m} s_i^m \phi_m (t - \tau_i^m) + \varepsilon(t) \quad . \tag{1}$$

, where, τ_i^m is temporal position,
S_i^m is coefficient of the i^{th} instance of kernel ϕ_m ,
n_m indicate the number of instances of ϕ_m ,
M is the number of ϕ_m ,
$\varepsilon(t)$ is a noise.

Here the Kernel function is a Gammatone filterbank whose center frequency and width are set according to an ERB (Equivalent Rectangular Bandwidth) filterbank cochlear model [8]. Fig. 4 shows the Gammatone filterbanks that have 64 channels.

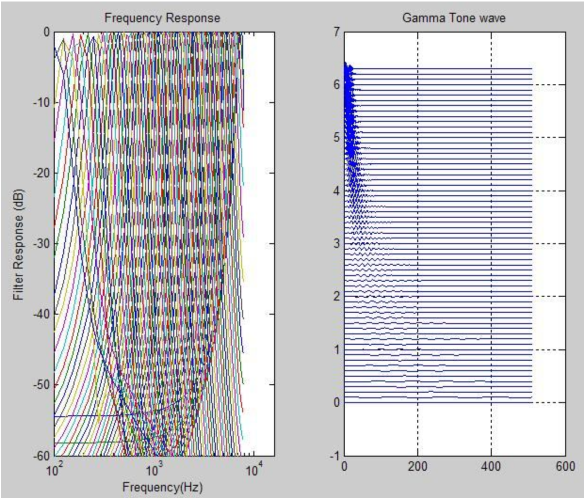
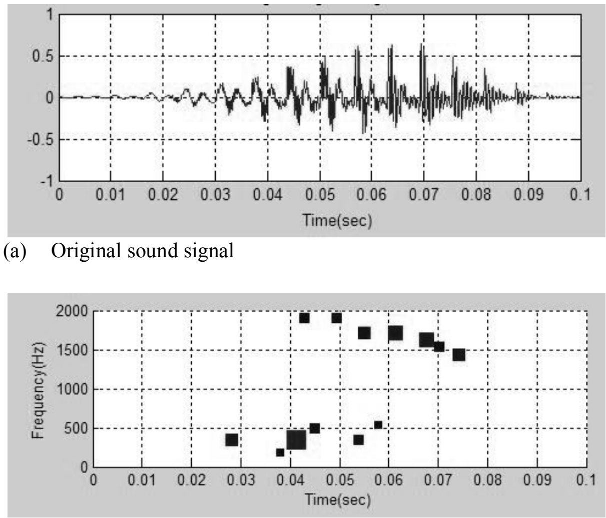


Fig. 4. The Gammatone filterbanks

Fig.5 illustrates a generative model which represents a sound signal as a set of kernel functions; Fig. 5 (a) is the original sound signal and (b) is the representation of the signal as a set of kernel functions. We call this plot as a spikegram, and a rectangular in the spikegram as a spike. In the spikegram, x-axis means the time, y-axis means the frequency, and the size is the magnitude of spike. In order to extract the spikes from the original sound, matching pursuit algorithm [9] is used since to produce a more efficient estimate of the time positions and coefficients.



(a) Original sound signal

(b) Spikegram

Fig. 5. A sound signal and spikegram

B. Interaural Time Difference

As mentioned in the Introduction chapter, we used the ITD as a cue of binaural sound localization. Through eq. (1), we can get, for each microphone, the position of max coefficient at each kernel function which can be described as follows.

$$x(t) \rightarrow \{\tau^m, S^m(\tau^m)\}$$

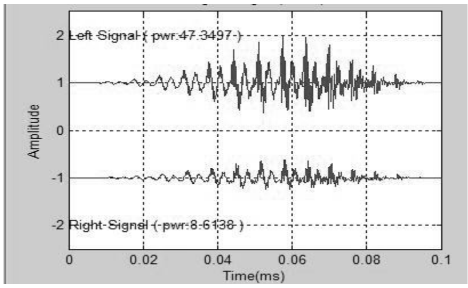
$$\tau^m = \arg\max_{\tau_i^m}\{S_i^m(\tau_i^m)\} \qquad (2)$$

Then, the time difference of two sound signals which are inputs from two microphones is calculated using the difference of max coefficient positions between left and right spikegrams (binaural spikegrams), which we call SITD (Sparse coding based ITD).

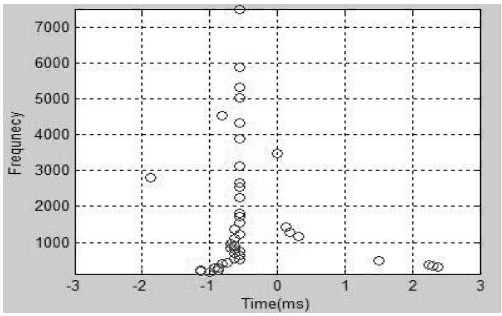
$$SITD^m = \tau_r^m - \tau_l^m \qquad (3)$$

Where, τ_r^m is max coefficient position of right signal,
τ_l^m is max coefficient position of left signal,
m indicate the number of kernel functions.

Fig. 6 shows the SITD of two sound signals.



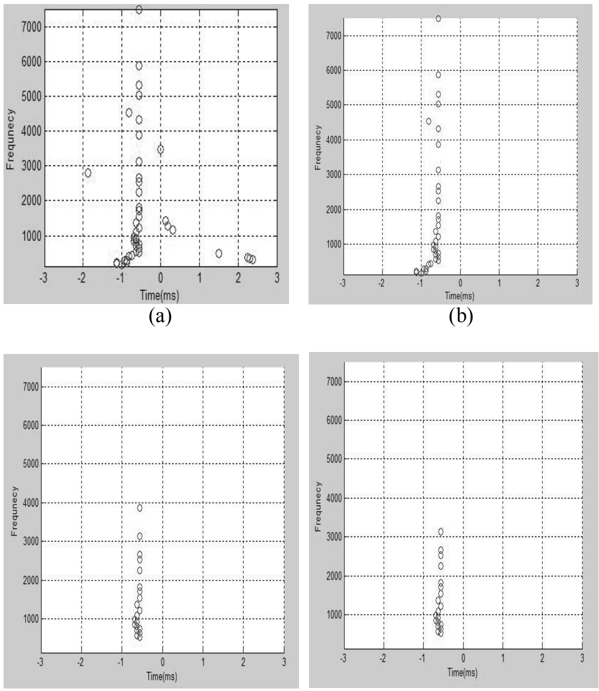
(a) Sound signals (sound source at 30-degrees left of center)



(b) The time difference between left and right

Fig. 6. Sound signals and time difference based on binaural spikegrams.

The time difference has errors that cause failure in the estimation of the direction of sound source. The errors can be dealt with various filtering method. The first thing is mean-variance filter. The mean can be calculated by the summation of Gaussian which has the SITD time positions as mean. The second thing is bandpass filter. We use only 500 ~ 4,000Hz. The third filter is threshold filter in the magnitude of the SITD, i.e., the time difference that has smaller coefficient is filtered out. Using these three kinds of filter, we can remove undesirable spikes and get the precise SITD. Fig. 7 shows the time difference of two sound signals after filtering.



(a)

(b)

(c)

(d)

Fig. 7 (a) is original time difference between two signals. (b) is a result which applied mean variance filter from (a). (c) is applied bandpass filter in (b). (d) is a result after applied three kinds of filter.

IV. Self Organizing Map

A. Sound collection

In our research, we fabricate a robot system that collects sound at -70 degrees to +70 degrees with 5 degree resolution, and build up a database for the learning of SOM. This system is shown in Fig. 8.

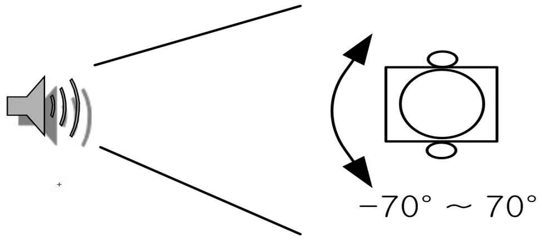


Fig. 8. The robot system that collects sound source

B. Learning

The binaural sound localization system in this research has the Self-Organizing Map [4]. We organize 1-dimensonal SOM at each frequency bin, because SITD has a different characteristic in frequencies. The SOM has 58 neurons for the representation of azimuth angle and 3,000 iterations for the organizing map. The learning algorithm is like this,

Step 1. Initialize weight vectors and neighborhood size.

Step 2. Select a ‘winning node’. The winning node is selected by Euclidean distance between input and weight vector. This equation is like this,

$$C(t) = \arg\min_i (\| X(t) - W_{ij}(t) \|_2) \quad . \tag{4}$$

Where, C(t) is the winning node
W(t) is the weight vector of node i at time t.
X(t) is the input vector.

Step 3. Update the weights using equation (5) ~ (7)

$$W_{ij}(t+1) = W_{ij}(t) + \Delta W_{ij}(t). \tag{5}$$

$$\Delta W_{ij}(t) = \alpha \phi(i,k)(X - W_{ij}(t)) \quad . \tag{6}$$

$$\phi(i,k) = e^{\frac{-d(i,c)^2}{\beta(t)^2}} \quad . \tag{7}$$

Where $\phi(i,k)$ is referred to as the neighborhood size,
α is a learning rate,
$-d(i,c)^2$ is the Euclidean distance form node i to the winning node c,
$\beta(t)^2$ is the neighborhood size at time t.

V. Estimation of Azimuth Angle

In our research, we estimate azimuth angle using the sparse coding based SITD and SOM. The overview of our integrated system is shown in Fig. 9.

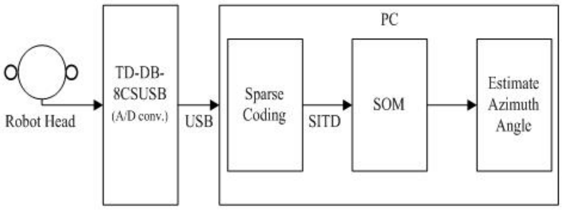


Fig. 9. Integrated System

The trained SOM feed with the sparse coding results required for the calculation of the azimuth angle. As mentioned in Section IV-B, we organized 1-dimensonal SOM at each frequency. The estimated results of the azimuth angle are made from the average of SOM outputs which can be described as follows.

$$Azimuth\_angle = \frac{\sum_{i=1}^{n} \theta_i}{n}. \tag{8}$$

Where, i is SOM number,
θ is the result of SOM,
n is the number of SOM output.

VI. Experiments

A. Environments

We performed an experiment in a demonstration room to verify our binaural sound localization system. The demonstration room looks like Fig 10. The robot is located in the center of the room and the speaker is set in front of the robot. The demonstration room also has a table, a bed, and a partition.

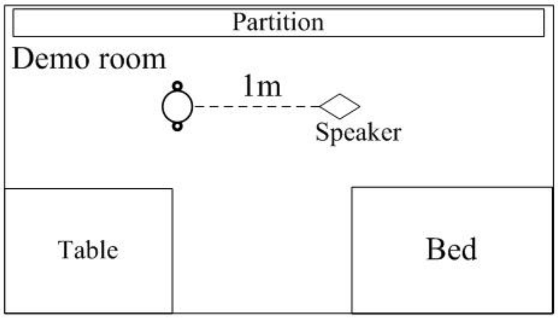


Fig. 10. Integrated System

B. Results

We have performed off-line tests. The off-line means estimating the direction with recorded sound signals. The sound signals are sampled at 16 kHz and a frame is composed of 1,600 samples (100ms). Input voice is “Come on, Tirot” in Korean. Fig. 11 shows the result of the binaural sound localization system.

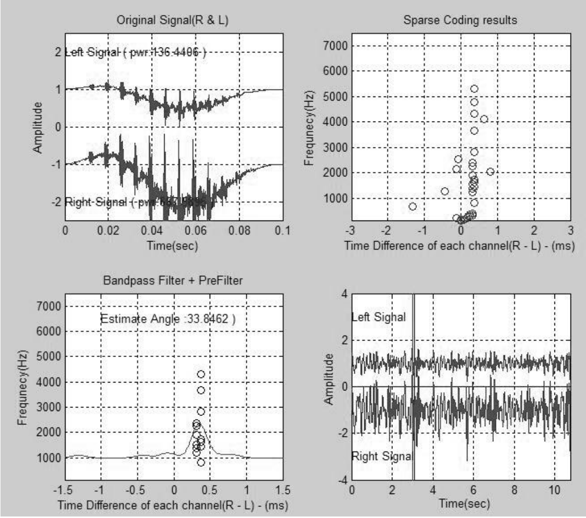


Fig. 11. Result of sound localization (Source location: + 30 degree)

The lower right figure represents total sound signals. Here, one frame near 3 sec. is zoomed in the upper left figure where the left ear’s signals are smaller than right ones, because the sound source is located in right side (+30 degree) of the robot. The upper right figure shows the results of sparse coding at the frame. Since in the first figure which is the representation of a frame (100ms) that we estimated, there are exists low frequency noises we filtered them out by applying a bandpass filter. Also, after applying three filters explained in section III-B into the original SITD of the upper right figure, we got the clean SITD of the lower left figure where the final result of the binaural sound localization is shown. It represents the average of sparse coding results. The mean-variance filter uses this average. In this case, our system estimates about +33 degree which is very close to the real direction (+30 degree).

C. Performances

Performance results are recorded at 29 difference angles. The distance between the robot and a speaker is 1m and 2m. The experimental results are given in table I. The error range for performance measure is ± 10 degrees and the average of success rate is 98 % and 94% (we regard as success the results within the error range). The statistical results are recorded in table II. This table shows mean, SD (Standard Deviation) and RMSE (Root Mean Square Error). SNR of the input signal is around 13 ~ 15dB during the experiments.

TABLE I
EXPERIMENTAL RESULT OF SOUND LOCALIZATION

ANGLE	SUCCES RATE of 1m experiment	SUCCES RATE of 2m experiment
-70°	100 %	96 %
-65°	100 %	100 %
-60°	100 %	94 %
-55°	100 %	98 %
-50°	100 %	92 %
-45°	100 %	100 %
-40°	100 %	100 %
-35°	100 %	94 %
-30°	100 %	100 %
-25°	93 %	90 %
-20°	95 %	94 %
-15°	100 %	83 %
-10°	100 %	100 %
-5°	98 %	100 %
0°	100 %	97 %
5°	94 %	90 %
10°	95 %	96 %
15°	98 %	100 %
20°	100 %	82 %
25°	100 %	100 %
30°	100 %	94 %
35°	100 %	93 %
40°	100 %	90 %
45°	76 %	82 %
50°	92 %	95 %
55°	97 %	94 %
60°	100 %	94 %
65°	100 %	100 %
70°	96 %	89 %

SUCCESS RATE OF BINAURAL SOUND SOURCE LOCALIZATION

TABLE II
EXPERIMENTAL RESULT OF SOUND LOCALIZATION

ANG LE	1m experiment			2m experiment		
	mean	SD	RMSE	mean	SD	RMSE
-70°	-65.95°	1.49°	4.05°	-60.80°	1.80°	9.20°
-65°	-63.80°	2.36°	2.30°	-61.69°	4.98°	1.67°
-60°	-61.45°	2.48°	3.15°	-58.63°	2.88°	3.22°
-55°	-57.84°	2.15°	3.73°	-51.44°	3.70°	4.06°
-50°	-53.46°	2.38°	3.52°	-48.74°	6.86°	7.15°
-45°	-46.77°	2.15°	2.31°	-41.04°	4.44°	4.52°
-40°	-40.50°	1.41°	1.50°	-38.18°	3.13°	3.95°
-35°	-37.22°	1.45°	2.22°	-31.83°	5.55°	6.02°
-30°	-34.66°	4.11°	5.06°	-30.27°	4.62°	4.64°
-25°	-27.31°	2.18°	2.71°	-27.94°	5.16°	5.23°
-20°	-20.29°	1.33°	1.34°	-23.75°	5.20°	5.63°
-15°	-14.74°	1.04°	1.00°	-17.74°	6.59°	6.27°
-10°	-10.35°	1.58°	1.45°	-10.77°	2.75°	2.86°
-5°	-4.32°	0.97°	1.09°	-3.14°	2.10°	2.44°
0°	0.12°	3.52°	3.53°	3.17°	2.26°	3.63°
5°	6.05°	2.93°	2.62°	8.50°	6.17°	6.71°
10°	8.74°	2.21°	2.07°	9.16°	5.24°	5.39°
15°	15.21°	0.77°	0.78°	15.89°	5.02°	5.06°
20°	20.32°	0.98°	0.94°	22.80°	8.14°	8.39°
25°	25.58°	0.77°	0.84°	28.43°	4.26°	4.92°
30°	30.63°	1.20°	1.19°	29.38°	6.98°	6.90°
35°	37.27°	1.26°	2.27°	31.42°	5.38°	5.54°
40°	42.60°	5.83°	6.21°	41.99°	10.43°	10.83°
45°	46.86°	4.31°	4.90°	47.96°	4.69°	4.65°
50°	53.04°	2.10°	3.12°	54.94°	3.03°	5.48°
55°	57.29°	1.67°	2.51°	58.92°	7.14°	8.30°
60°	62.10°	2.32°	3.44°	60.30°	5.53°	5.57°
65°	63.19°	1.69°	1.01°	63.81°	3.06°	2.83°
70°	66.91°	1.01°	3.09°	63.11°	4.16°	6.89°

STATISTIC OF BINAURAL SOUND SOURCE LOCALIZATION

Fig. 12 and 13 show statistical graph of SD and RMSE at 1m, 2m.

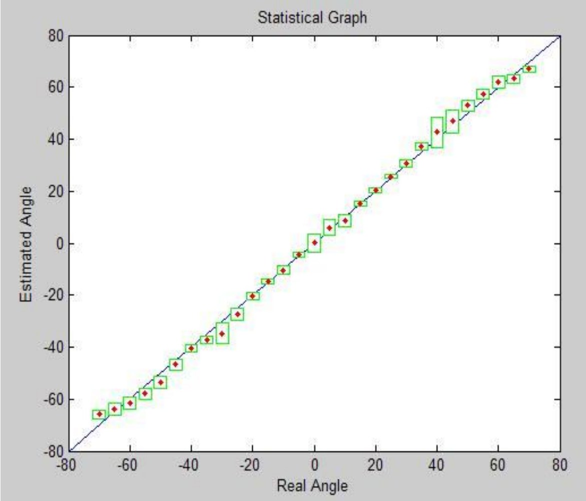


Fig. 12. Statistical graph of 1m experiment.

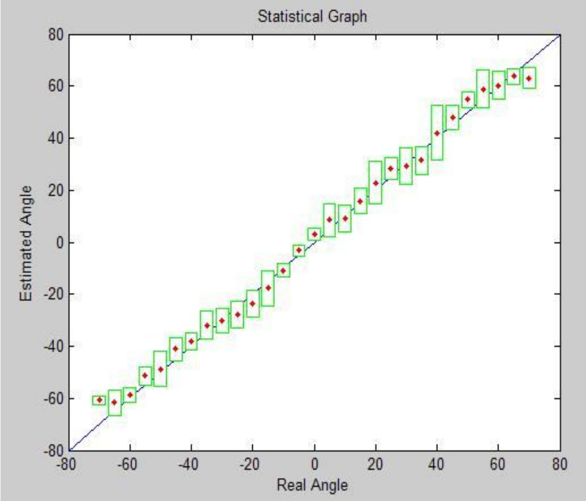


Fig. 13. Statistical graph of 2m experiment.

In the graph, x-axis means real angle and y-axis means estimated angle. The red points represent the mean and the green rectangular show the SD at each angle.

VII. Conclusion

In this paper, we have presented a binaural sound localization system based on the sparse coding and SOM. After getting the spikegrams of each ear's signal, calculating the SITD between the binaraul spikegrams, and applying three kinds of filtering to the SITD, we have got clean SITD data. Also, using SOM's learning method w.r.t the SITD, we could estimate azimuth angle of sound source with high performance. SOM is used in other researches but using it together with the sparse coding with post filters is a new approach in sound source localization. Our next plan is to add online learning to make more robust our binaural sound localization system.

References

[1] Yoon-Seob Lim, Jong-Suk choi and Mun-Sang Kim, "Probabilistic Sound Source Localization," International Conference on Control, Automation and Systems 2007.

[2] Nakashima, H. & Mukai, T, "3D Sound source localization system based on learning of binaural hearing," IIIE international Conference on Systems, Nan and Cybernetics, 4, 3534-3539, 2005

[3] Smith, E. C & Lewichi , M. S, "Efficient coding of time-relative structure using spikes," Neural Comput. 17, 19-45, 2005

[4] T. Kohonen, "The self-organizing map." Proceedings of the IEEE, 78(9): 1464-1480, 9 , 1990

[5] Performance PeopleBot Operations manual

[6] Ricardo Beira, Doutor Jose, Alberto Rosado Dos Sandos Victor, "Mechanical Design of and Anthropomorphic Robot Head"

[7] Lewicki, M. S. and Sejnowski, T. J, "Coding time-varying signals using sparse, shift-invariant representations," In Advances in Neural Information Processing Systems, volume 11, 730-736, MIT Press, 1999

[8] Slaney, M. Auditory toolbox. Technical Report 1998-010, Interval Research Corporation

[9] Mallat, S.G. and Zhang, Z, "Matching Pursuits with time-frequency dictionaries," IEEE Transctions on Signal Processing, 41(12):339-3415, 1993

[10] L. Calmens, G.Lakemeyer, H.Wagner, "Azimuthal sound localization using coincidence of timing across frequency on robotic platform", J.Acoust. Soc. Am., Vol.121, No4,2034-2048,2007

[11] F. Keyrouz, K. Diepold, "An Enhanced Binaural 3D Sound Localization Algorithm", isspit, pp.662-665, 2009 IEEE International Sympo

[12] sium on Signal Processing and Information Technology, 2006