Spatial Localization of Multiple Sound Sources in a reverberant Environment

Huakang Li, Jiahao Lu, Jie Huang, Takuya Yoshiara
School of Computer Science and Engineering, University of Aizu
d8092104@u-aizu.ac.jp

Abstract*-In this paper, we propose a spatial localization of multiple sound sources using a spherical robot head ranged with four microphones. We obtain the arrival time difference using inter-aural time difference and inter-aural phase difference. Based on the model of precedence effect, arrival temporal disparities obtained from the zero-crossing point is used to calculate the time differences and suppress the influence of echoes in the reverberant environment. To integrate spatial cues of different microphone pairs, a mapping method from the correlation between different microphone pairs to a three dimension map corresponding to azimuth and elevation of sound sources direction is proposed. Experiments show the system provides the distribution of sound sources in azimuth-elevation, even concurrently in reverberant environments.*

Keywords: Acoustic signal processing, cross-correlation, time-delay estimation, precedence effect.

1. Introduction

The next generation robot is expected to adapt to the human society even if some unknown environments. Like human being and animals, to make an excellent work, it is necessary to recognize the environments and the active things for the robot with sensors. Visual sensor is the most popular sensor used for mobile robots [1], however, when a robot generally looks at the external world from a camera, difficulties will occur while the object is not in the visual field of the camera, or the environment lighting is to lowering for the camera to pick-up the objects. In these situations, the most useful information is provided by auditory system, which is considered as the second sensory perception of human beings and animals. For mobile robots, auditory systems can not only recognize the environments, but also cooperate with vision systems [2].

Many robotic auditory systems, similar to the human auditory system, are equipped with two microphones [3]. For spatial localization, because of complexity and ambiguity of spectral cues, the robotic auditory system is very vague in sound elevation localization. Hence multiple channels sound localization system such as the Sony SKD-4X series robots with seven microphones was approved. In this paper, we proposed a new method using a spherical head only with four microphones, three on the head center level and one on the top of the surface. Inter-aural Time Difference and Inter-aural Phase Difference of each microphone pair were used to obtain the Arrival Time Differences Array that exhibits a peak in the location corresponding to the sound source. In the ordinary room, with traditional approaches, the reverberation causes spurious peaks interfere in the localization curs. We supplied the space vector and space scalar quantity summarization restriction to restrain the spurious peaks that may have greater amplitude than the peaks of the real source.

To treat multiple sound sources, an integration method of spatial cues of different microphone pairs was proposed. The geometric averages of different microphone pairs' candidate functions were approved to obtain the accurate sound source locations. The integration of arrival temporal disparities calculated from the zero-crossing points between different pairs of microphones was weighted to exhibit echoes in multiple sound sources localization, and the model of precedence effect was implemented for avoiding reverberation and echoes in the ordinary environments [4]. The results were mapped into 3-Dimension map corresponding to azimuth and elevation for sound sources [5].

2. The Spherical Head with four Microphones

Assuming the side with camera is the front of the mobile robot (Fig.1). While the radius(r) of the spherical robot head is 15cm, the location of four microphones can be denoted as $M_1(15cm, 0°, 90°)$,

$M_2(15cm,180°,0°)$, $M_3(15cm,60°,0°)$, $M_4(15cm,300°,0°)$, where the center of the sphere O is defined as the center of a three-Dimension polar coordinates.

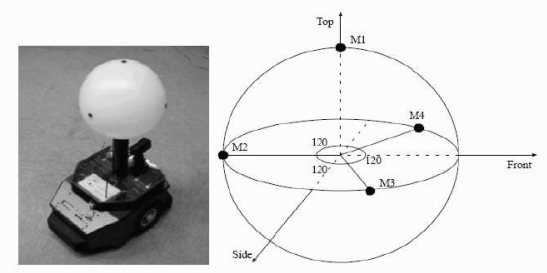

Fig1: A robot with a spherical head and the arrangement of the microphone set

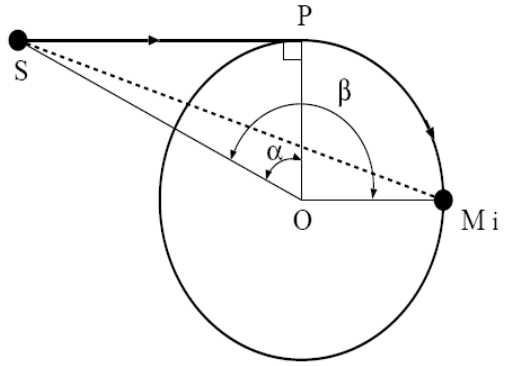

Fig 2: A case, which the sphere is inserted between a sound source and a microphone

Denote the arrival time from the sound source S to each microphone M_i as t_i, the arrival time difference (ATD) between microphone M_i and M_j can be defined as Δt_{ij}, and the ADT vector can be written as

$$\overrightarrow{\Delta T} = \{\Delta t_{ij} \mid i < j \bigcap i, j = 1,2,3,4\}$$

The arrival time difference Δt_{ij} that depends on the azimuth θ and elevation ϕ of the sound source S in the three-dimension polar coordinate, and can be calculated as

$$\Delta t_{ij}(\theta,\phi) = \frac{d_i(\theta,\phi) - d_j(\theta,\phi)}{v}$$

where d_i, d_j denote the distance from the sound source S to microphone M_i and M_j, v is the sound velocity and $i, j = 1,2,3,4 \bigcap i \neq j$.

Denote the direction of microphone M_i and sound source S by $(\theta_{M_i}, \phi_{M_i})$ and (θ,ϕ), the coordinates of M_i and S can be defined as

$$\overrightarrow{OM_i} = (r\cos\phi_{M_i}\cos\theta_{M_i}, r\sin\phi_{M_i}, r\cos\phi_{M_i}\sin\theta_{M_i})$$

$$\overrightarrow{OS(\theta,\phi)} = (D\cos\phi\cos\theta, D\sin\phi, D\cos\phi\sin\theta)$$

Realizable, the sound could not arrive at the microphone directly on some position such as the inverse side, and the change of the pass distance will affect the accuracy of real arrival time difference. Fig.2 shows one of the intersection planes for the spherical head plane that the direct distance SM_i is not the most accurate pass between sound source and microphone. The sound from S arrives at microphone M_i through the direct line segment SP and the arc PM_i transmitted along the surface of the sphere. Thus, the real distance $d_i(\theta,\phi)$ would be calculated as

$$d_i(\theta,\phi) = SP + arc(PM_i)$$

Finally, the arrival distance calculation function by the relation between SP and SM_i can be obtained as

$$d_i(\theta,\phi) = \begin{cases} |\overrightarrow{S(\theta,\phi)} - \overrightarrow{M_i}| & (SP \geq SM_i) \\ SP + ark(PM_i) & (SP < SM_i) \end{cases}$$

3. Estimation of Possibility for Sound Source Direction

3.1 Arrival time difference estimation

Arrival time difference (ATD) and their candidates are calculated from phase difference of each frequency band between different microphone pairs.

$$\overrightarrow{\Delta t_{c_k}} = \{\frac{\Delta\phi}{2\pi} + \frac{n}{f_k} \mid n = \pm 1, 2, ..., N\} \ (N \leq [\frac{d_{\max}}{v} f_k])$$

where $\Delta\phi$ denotes the phase difference in practical measurement, and Δt_{c_k} denotes ATD candidates, f_k is the center frequency of one band frequency, n is the estimate integer smaller than N which is the maximum of the candidate number for this frequency band.

3.2 Restriction between Arrival Time Differences

When the f_k is higher and higher, the candidate N will be big enough to obtain lots of spurious peaks using traditional sound source direction. The space vector summarization and space scalar quantity summarization were proposed to reduce the spurious estimation of ATDs.

$$\begin{cases} \Sigma\Delta t_{c_k}^{(ij)} = 0 \\ \Delta\varphi_{Min} \leq \Sigma \mid \Delta t_{c_k}^{(ij)} \mid \leq \Delta\varphi_{Max} \end{cases} \quad (i, j = 1,2,3,4 \bigcap i \neq j)$$

where $\Delta\varphi_{Min}$ and $\Delta\varphi_{Max}$ are the minimum and maximum of the sum of the spacial vectors between each microphone pairs,

$$\Delta\varphi_{Max} = \frac{\pi r}{C} \quad \Delta\varphi_{Min} = \frac{2r}{C}$$

3.3 Sound Source Direction Estimation

Denote the direction of sound source as azimuth θ and elevation ϕ , the arrival time difference for frequency band f_k is able to be defined as

$$\Delta t_k = q_k(\theta,\phi)$$

And the practical measurement of arrival time difference candidates can be written with a function as

$$C_k = g_k(\Delta t_c)$$

For mapping the arrival time differences candidate histogram to azimuth-elevation plane, the arrival time difference obtained in theory and the time axis of arrival time difference candidates between the same microphone pair were quantized with the same scale. Thus,

$$C_k = g_k(q_k(\theta,\phi))$$

Then, use a new define of the possibility estimation function defined as

$$p_k(\theta,\phi) = C_k(\Delta t_k(\theta,\phi))$$

where $p_k(\theta,\phi)$ is the possibility direction of sound source $S(\theta,\phi)$ .The peak of single microphone pair can not localize the sound source direction exclusively while there are some ambiguous directions by the maximum candidate of arrival time. Since there were multiple microphone pairs between 4 microphone, the possibility estimation of 6 microphone pairs can be obtain by

$$P_k = (\Pi p_k^{(ij)})^{\frac{1}{6}} \qquad i, j = 1,2,3,4 \bigcap i < j$$

$$P = \frac{1}{K}\sum_{k=1}^{K} P_k$$

the estimation provides a possible spatial distribution of sound sources. To get more accurate results, this estimation value of each direction is the geometric average of six candidate function for different microphone pairs.

4. Echo avoidance model of precedence effect

Except within a large, open expanse of snow-covered ground or in a mountain summit, in the rooms, the discrete-time signal received at the sensor contains the direct sound components and reflected components. A particular reflection within a reverberant field is usually categorized as an early reflection or as late reverberation. The arrival time of late reverberation can be ignored while it is usually larger than 80ms while drops 60 db below the direction sound. In this paper, we used the echo avoid model with precedence effect to overcome the interference of early reverberation that only drops within 10 dB.

Denote the sound intensity is $s(f,t)$ and the estimated echo is $s_e(f,t)$. (both $s(f,t)$ and $s_e(f,t)$ are the signals in an independent narrow sub-band *f*) The inhibition will be correspond to the ratio $s(f,t)/s_e(f,t)$. Denote the impulse response from the sound source to a sensor as $h(f,t)$, and the part caused by echoes by $h_e(f,t)$. The echoes $s_e(f,t)$ can be estimated as

$$s_e(f,t) = s(f,t) * h_e(f,t)$$

While the impulse response is non-predicted, we gave a generalized approximation echo value g_e defined as

$$g_e = ke^{-(t-t_0)/\tau}$$

where t_0 is the delay time and τ is the decay rate of the impulse response. Thus, the estimated echoes can be calculated,

$$s_e(f,t) = M\{s(f,t-t'), g_e(t')\} \quad \text{for} \quad 0 < t' < \infty$$

M is the maximum of the convolution. The delay time t_0 and decay factor τ were measured and trained in advance to match the most general cases in an ordinary environment. (in the human auditory system, it should be chosen by learning effect.) The decay factor τ , however, does not severely affect the result of echo estimation. In the above approximation, the signal $s(f,t)$ contains not only direct sound but also the echoes. Thus, $s(f,t)$ itself has a decay feature can be much smaller (about 2 to 10 ms) than that of real environments.

By using the exponential decay feature of the generalized impulse response, we can implement the echo estimation algorithm by a feed-back mechanism as show in Fig.3, where t_s is the sampling time interval, $n_d = t_s/\tau$, $d = e^{-nd}$ and ς is a sigmoid function. This algorithm is very fast, requiring only two multiplications and one comparison to predict the echoes.

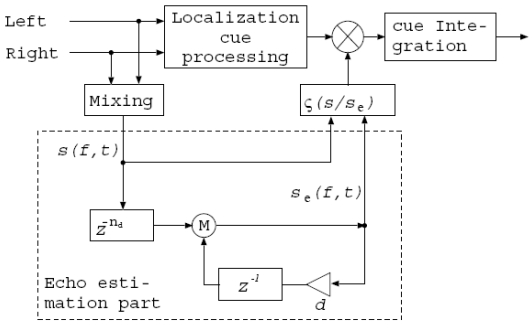

Fig3: echo estimation in the EA model of the precedence effect

A new method, namely, time candidate histogram weighted with EA model based on the precedence

effect was proposed. In this method, the weight based on the EA model was addition to the frequency domain of the signal, and weighted signals were used for the time candidate calculation. However, the weight is calculated as follow,

$$w = (\frac{s(f,t)}{s_e(f,t)} - \delta)/2 \qquad (0 \le w \le 1)$$

Where $s(f,t)$ and $s_e(f,t)$ is the present sound signals and echoes expected from the preceding signal, and δ (it should be chosen by learning effect while the environment is changed) is chosen to be 2 because in the reverberant environment the first reflected sound drops 3 dB than the direct sound source and is determined from the room size [6].

5. Experiments and Results

The experiments were carried out with single sound source and multiple sound sources both in an anechoic chamber and in an ordinary room. The anechoic chamber has a size of $5.5\times5.5\times5.5$ m, and the ordinary room's size is $4\times6\times3$ m. The robot head was set 1 m above the floor, while loudspeakers were set at 1 m away from the center of the robot head, and the microphone array recorded the sound sources for 5 seconds with sampling rate 44.1 kHz. The location results of each frame (250ms) were plotted in a 3-dimension map with azimuth axis, elevation axis (by top, side, front views).

5.1 Localization of a Single Sound Source

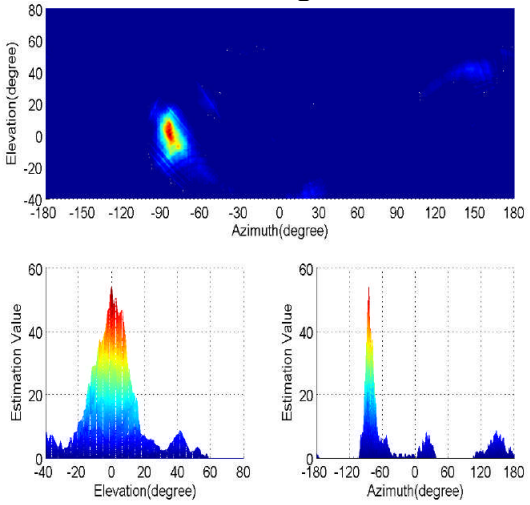

Fig4: localization result of one frame for a single sound source $S(-90^\circ,0^\circ)$ in an anechoic chamber.

Fig.4 shows the localization result of one frame in an anechoic chamber for the source direction $S(-90^\circ,0^\circ)$. Considering the variety of sound sources and unavoidable noise, we shifted 10 frames to obtain the total localization result, and applied a threshold [7] to estimate the exclusive possibility of sound direction, and normalize the total result between 0 and 1 (Fig.5).

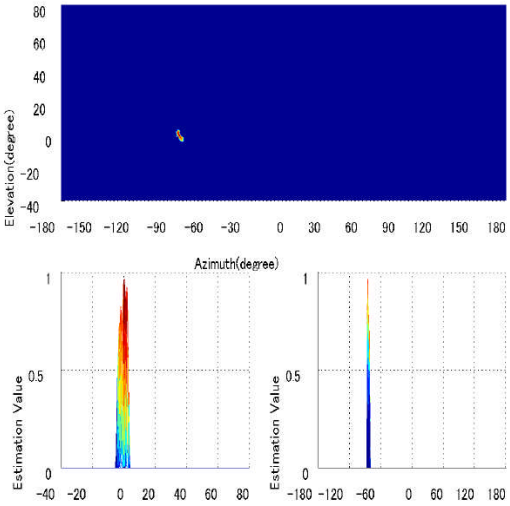

Fig5: localization results with threshold of all time frames for a single sound source $S(-90^\circ,0^\circ)$ in an anechoic chamber.

Table1: Average of peak direction errors of one sound source localization in an anechoic chamber. (azimuth, elevation)

Sound source direction	$(-90^\circ,0^\circ)$	$(0^\circ,25^\circ)$	$(90^\circ,30^\circ)$
Average peak direction errors	$(5^\circ,2^\circ)$	$(3^\circ,3^\circ)$	$S(0^\circ,2^\circ)$

Table 2: Average of peak direction errors of one sound source localization in an ordinary room.

Sound source direction	$(-90^\circ,0^\circ)$	$(0^\circ,25^\circ)$	$(90^\circ,30^\circ)$
Average peak without EA model	$(2^\circ,2^\circ)$	$(1^\circ,5^\circ)$	$(1^\circ,5^\circ)$
Average peak direction errors with EA model	$(1^\circ,3^\circ)$	$(1^\circ,3^\circ)$	$(0^\circ,4^\circ)$

Table.1 shows the peak average direction errors of a single sound source at different position in an anechoic chamber. The system can localize a single sound source direction within 5 degrees errors without any influence in an anechoic chamber correctly.

In the ordinary room, the localization results with threshold for sound source direction $S(-90^\circ,0^\circ)$ was

exactitude both without EA model (Fig.6) and with EA-model (Fig.7). From Table2, we can see that the locations were more unambiguous and verged to the ideal location of sound source.

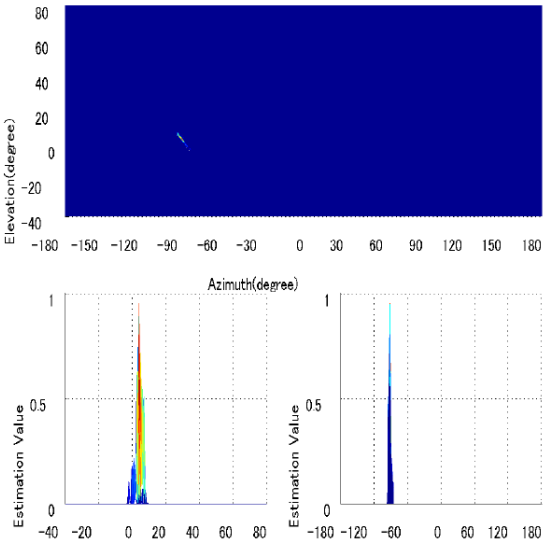

Fig6: localization results with threshold of all time frames for one sound source $S(-90^\circ,0^\circ)$ without EA-mode in an ordinary room.

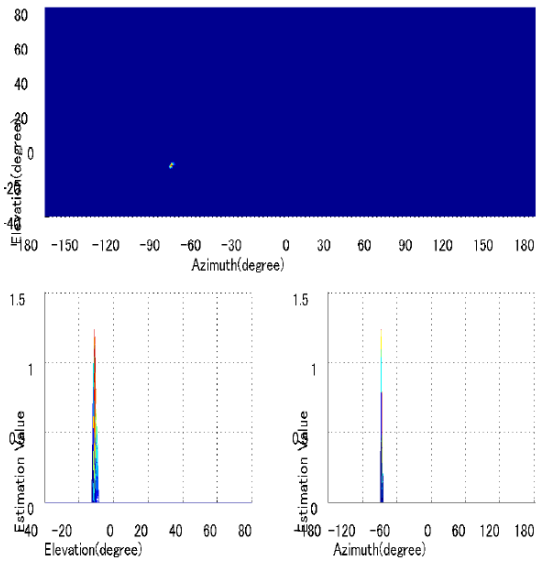

Fig7: localization results with threshold of all time frames for one sound source $S(-90^\circ,0^\circ)$ with EA model in an ordinary room.

5.2Localization for Multiple Sound Sources

For multiple sound source localization, we used one pair of loudspeakers to play two different sound sources at different locations, and the first sound source was played with stronger intensity while the second sound source was also a little later than the first one. Fig.8 shows the localization results with threshold of concurrent sound sources $S_1(0^\circ,0^\circ)$ and $S_2(-90^\circ,0^\circ)$ by arrival time differences candidate histograms in an anechoic chamber, and the average of peaks direction errors were $e_1(8^\circ,7^\circ)$ and $e_2(5^\circ,12^\circ)$. For some frequency characters of the two different sound sources had the same frequency band components in one frame, the interference between each other can't be ignored, however, it was not so distinct. The time candidate histogram method was able to localize multiple sound directions approximately in an anechoic chamber.

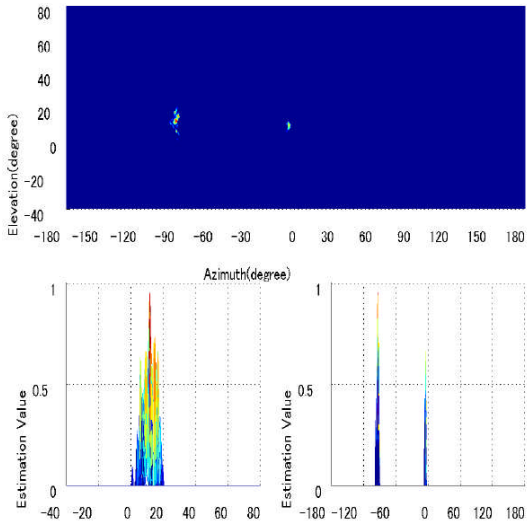

Fig 8: localization results with threshold of all time frames for multiple sound sources $S_1(0^\circ,0^\circ)$ and $S_2(-90^\circ,0^\circ)$ in an anechoic chamber.

Fig.9 shows the localization results without threshold of all time frames for multiple sound sources $S(0^\circ,0^\circ)$ and $S(-90^\circ,0^\circ)$ in an ordinary room without EA-model. Due to the interference between sound sources and echoes, the locations of sound sources were presented as some special lines on the map, which disturbed the peak direction localization. For the location of sound sources were not so determinate, the average of peaks direction errors was either able to presented.

While the EA-model was supplied in the localization of multiple sound sources in the ordinary room, the peak locations of two sound sources were presented very vivid (Fig.10). The average of peaks direction of multiple sound sources at different location in the ordinary room was showed in Table 3. The multiple sound sources localization using candidate histogram with EA-model in the ordinary room can be

presented much more accurate even than the results in an anechoic chamber without EA-model.

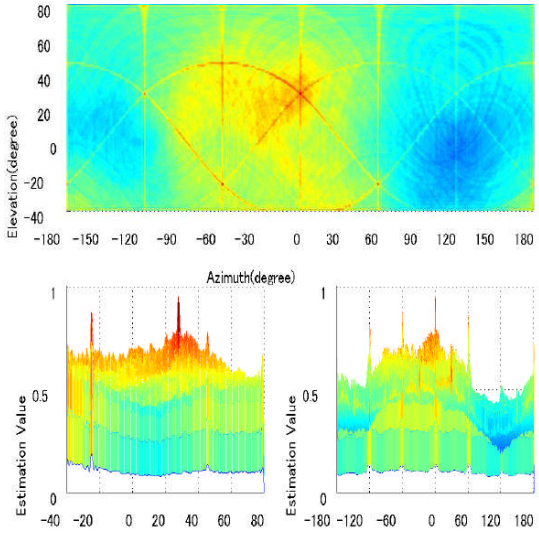

Fig 9: localization results with threshold of all time frames for multiple sound sources $S_1(0^\circ,0^\circ)$ and $S_2(-90^\circ,0^\circ)$ in an ordinary room.

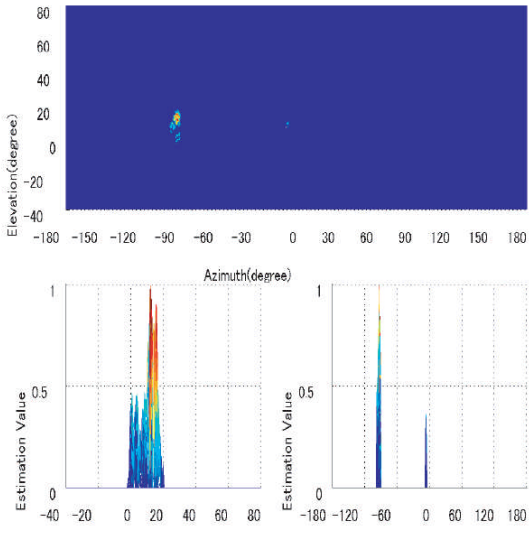

Fig 10: localization results with threshold of all time frames for multiple sound sources $S_1(0^\circ,0^\circ)$ and $S_2(-90^\circ,0^\circ)$ with EA model in an ordinary room.

Table 3: average of peaks direction errors of multiple sound sources localization in an ordinary room (azimuth, elevation).

Sound sources direction	$(0^\circ,0^\circ)$ $(-90^\circ,0^\circ)$	$(0^\circ,15^\circ)$ $(90^\circ,15^\circ)$	$(0^\circ,15^\circ)$ $(90^\circ,30^\circ)$
Average peaks direction errors	$(7^\circ,8^\circ)$	$(4^\circ,2^\circ)$	$(2^\circ,6^\circ)$
	$(4^\circ,12^\circ)$	$(4^\circ,2^\circ)$	$(6^\circ,8^\circ)$

6. Conclusions

In this paper, we proposed a new sound localization system, namely, arrival time differences candidate histograms using four microphones arranged on a spherical robot head. A mapping method projected from histogram of different microphone pairs to an azimuth-elevation plan was proposed. For at the edge of the candidate histogram, the average of peaks direction errors was distinct, we would improve the transition from candidate histogram to mapping system in the future works. This system can localize a single sound source correctly in an ordinary room, and could also localize two concurrent sound sources separately by histogram method weighted with EA-model. Since the frame was short enough, we would extend this method for active sound localization in the ordinary environment.

References

[1] C. Zhao, Y. Ohtake, and J. Huang. “Robot position calibration using colored rectangle signboards” J. three dimensional images, 17(1):166-169, march 2003.

[2] Y. Asai, H. Nakashima, T. Yamamura, J. Huang, and N. Ohnishi. “Acquiring the ability of object localization by vision and audition through motion”. IECIE Trans. D-II, J83-D-II(7): 1676-1684, 2000.

[3] J. Huang, “Developing a Multi-model Tele-robot-spatial auditory Processing”, Journal of Shanghai University, Vol. 5 (Suppl.): 147-151, September 2001.

[4] J. Huang, N. Ohnishi, and N. Sugie. “Sound localization in reverberant environment based on the model of the precedence effect” IEEE Trans. Instrum, and meas., 46(4):842-846, August 1997.

[5] J. Huang, N. Ohnish, and N. Sugie. “A multiple sound source localization system using temporal disparity histograms” J. Robotics Soc. Jpn., 9(1):29-38, February 1991.

[6] H, Li, T. Yoshiara, Q. Zhao, T. Watanabe, J. Huang ,“A spatial sound localization system for mobile robots” IEEE Volume, Issue, Trans. Instrum and Meas., pp:1-6, May 2007.

[7] K. Saberi and D. R. Perrott. “Lateralization threshold obtained under conditions in which the precedence effect is assumed to operate”, J. Acoust. Soc. Am, Vol.87, pp 1732-1737, 1990.