Continuous Sound Source Localization based on Microphone Array for Mobile Robots

Hong Liu and Miao Shen

Abstract— **It is a great challenge to perform a sound source localization system for mobile robots, because noise and reverberation in room pose a severe threat for continuous localization. This paper presents a novel approach named guided spectro-temporal (ST) position localization for mobile robots. Firstly, since generalized cross-correlation (GCC) function based on time delay of arrival (TDOA) can not get accurate peak, a new weighting function of GCC named PHAT-$\rho\gamma$ is proposed to weaken the effect of noise while avoiding intense computational complexity. Secondly, a rough location of sound source is obtained by PHAT-$\rho\gamma$ method and room reverberation is estimated using such location as priori knowledge. Thirdly, ST position weighting functions are used for each cell in voice segment and all correlation functions from all cells are integrated to obtain a more optimistical location of sound source. Also, this paper presents a fast, continuous localization method for mobile robots to determine the locations of a number of sources in real-time. Experiments are performed with four microphones on a mobile robot. 2736 sets of data are collected for testing and more than 2500 sets of data are used to obtain accurate results of localization. Even if the noise and reverberation are serious. The proportion data is 92% with angle error less than 15 degrees. What's more, it takes less than 0.4 seconds to locate the position of sound source for each data.**

I. INTRODUCTION

As an important part of perception, auditory system leads to a new direction of the research on robot perception technologies. To simulate human auditory mechanism, sound localization technology uses acoustic sensor array to receive sound signal and estimate the positions of objects. Firstly, these sound signals collected by the sensors are handled by electronic devices, and then it can be achieved to perform the sound source position detection, identification and target localization plus tracking [1].

The application of sound source localization technologies has broad prospects in the areas of mobile robots and communication technologies. Early in 1995, Irie from MIT [2] installed a simple auditory system in robot, although functions of this auditory system were limited, which guides a prosperous way to the research of robot auditory in future. In 2006, Honda Research Institute in Japan [3] developed a multi-source real-time tracking system through a joint of In-Room Microphone Array (IRMA) and the microphone array embedded in a robot's head. IRMA consisted of 64-channel microphones embedded in a wall, while 8 microphones were embedded in the robot's head. As a result, it needs quite a lot of microphones to implement the system. In 2007, scientists in Canada [4] developed an obstacle-avoidance robot based on beamforming technology using 8-channel microphone array. In this system, the priori knowledge of sound source and ambient noise were needed, therefore, it has high computational complexity and weak real-time efficiency. At the same time, Hara reported that an 8-channel system using a robot embedded microphone array (REMA) had better performance for sound source localization [5]. However, the performance is worse when the robot is in motion, because it is difficult to synchronize signal capturing with motion precisely and to adapt to acoustic environmental changes after a robots motion [3].

Sound source localization technology based on microphone array can be categorized into 3 classes: (1) Directional technology based on high resolution spectral estimation [6]. (2) Controllable beamforming technology based on the biggest output power [7]. (3) technology based on time delay of arrival (TDOA) [8][9]. First method often aims at narrowband signals, but voice signals are broadband signals which need to improve positioning accuracy at cost of increasing the computational complexity. Consequently, it is ineffective in speaker positioning. Second method requires a priori knowledge of sound source and environmental noise and the computational complexity is high. The last TDOA method has low operation and strong real-time. It is suitable for single sound source localization. Through appropriate improvements to overcome the noise and reverberation, it can achieve a better positioning accuracy.

Currently, there are many difficulties in sound source localization using TDOA method for mobile robots as follows: (1) When people speaks continuously, with the robot's movement, the room reverberation [10] changes simultaneously, which not only needs to obtain an accurate reverberation, but also its localization algorithm have a strong need for real-time [3]. What's more, to locate the continuous sound source, the voice segment must be detected accurately. (2) When robot is in motion, the motor leads to the noise increases, strict de-noising algorithm is needed to distinguish voice and noise effectively. (3) Large size microphone array [11][12] installed on robot will lead to move difficultly, so that it is necessary to calculate a precise location of source with appropriate amount of microphones. (4) Most signal sound source locating system need prior knowledge of signal, for a number of sources generated in a short time can not achieve continuous localization. Continuous localization means, for single source system, robots can quickly estimate space

Hong Liu is with the Key Laboratory of Machine Perception and Intelligence, Peking University, Shenzhen Graduate School, Shenzhen, 518055 CHINA. hongliu@pku.edu.cn

Miao Shen is with the Key Laboratory of Machine Perception and Intelligence, Peking University, Beijing, 100871 CHINA. shenmiao@cis.pku.edu.cn

reverberation model, detect endpoints of voice segments and locate a number of sound sources which generating in a short time in real-time. Obviously, the most critical factors of continuous localization for mobile robots are performance of real-time and accuracy. Algorithm must has a low computational complexity for real-time demand and the changes of environmental reverberation must be detected in time to meet accuracy.

To address the issue which the parameters of ST position [17] have weights, this paper produces guidance on ST position method and designs a novel guided ST position algorithm using GCC-PHAT-$\rho\gamma$ method to get priori knowledge, such as the rough location of sound source, changes of environmental reverberation, etc. It is proved that the detection of sound location is in strong real-time for mobile robot using 4-channel microphone array. Our contributions are as follows.

1) A improved weighting function of GCC-PHAT-$\rho\gamma$ method is used to calculate the time difference of microphone pair, and reduce noise.

2) A guided ST position method is proposed combining with PHAT-$\rho\gamma$ method to reduce reverberation and determine accurate location of sound source on mobile robot and a continuous real-time location system is achieved while the robot is in motion.

The rest of this paper is organized as follows: Section II describes the general structure of system and array model. De-noising techniques, PHAT-$\rho\gamma$ and guided ST position method will appear in Section III. In Section IV experimental results and analysis are shown, and conclusions are drawn in Section V.

II. ARRAY MODEL AND GENERAL STRUCTURE

In this section, microphone array model and general structure are introduced. First, a microphone array model is determined, and then, the signal characteristics received by a pair of microphones are used to calculate the time difference. The process of estimating azimuth angle of sound source is shown in Fig. 1.

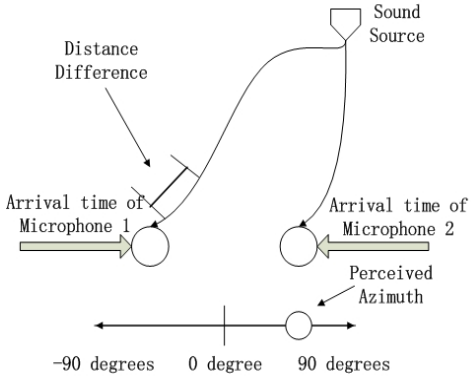


Fig. 1: Estimation of sound source azimuth angle.

A. Array Model and Parameter Design

Design of a microphone array model relates to the following factors on mobile robot, (1) appropriate amount of microphones, (2) the shape of microphone array, (3) the distance of two microphones, (4) the features and properties of robot. Firstly, too many microphones are inconvenient for moving the robot and hence increase the complexity of algorithm. By contrast, if too few microphones are used, the positioning accuracy cannot be guaranteed. Secondly, when the sound source and the two microphones are in a straight line, the error of time difference of linear microphone array increases. At the same time, circular array, such as Yuki Tamai who produced a 32-channel circular microphone array [13], needs a lot of microphones. Finally, the distance of two microphones can not be too small, because the quantization error will increase. For example, if the distance is 8 *cm* and sample rate is 44.1 *kHz*, the maximum offset of signals from the two microphone channels is $(8 \times 44100)/34000$. It is about 10 sampling points. That means the direction of a sound source changes from 0° to 180°, the deviation of sampling points appeared on the microphones is between $\pm 10°$ and the quantization error would amount to nearly 3°. However, the distance of two microphones also can not be too far according to far-field assumption constraints [14][15]. Paper [15] shows that when the microphone distance b and the distance r from sound source to the center of microphone array satisfy the relationship $r/b > 3$, the quantization angle error is less than 0.4°.

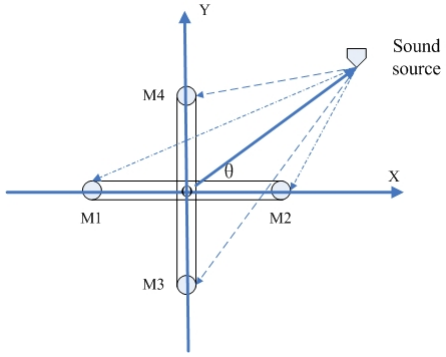


Fig. 2: Schematic of microphone array model.

Our microphone array system is composed of four microphones with a cross-shaped plane. The schematic of microphone array model is shown in Fig. 2. Any one microphone pair in this array can be used to calculate the time difference of sound transmitting from the sound source to these two microphones. The experiments show that four microphones can estimate direction of sound source accurately and more microphones will only increase the computational complexity of algorithm. For the safe relative distance between robot and speakers must be preserved, minimum of r is 100 *cm*. According to the relationship $r/b > 3$ and the robot size, the distance of two microphones is identified as 40 *cm*.

B. General Structure

A flow chart of continuous localization is shown in Fig. 3. De-noising algorithm, GCC-PHAT-$\rho\gamma$ method and guided ST position method are three parts in this structure and will be introduced in next section.

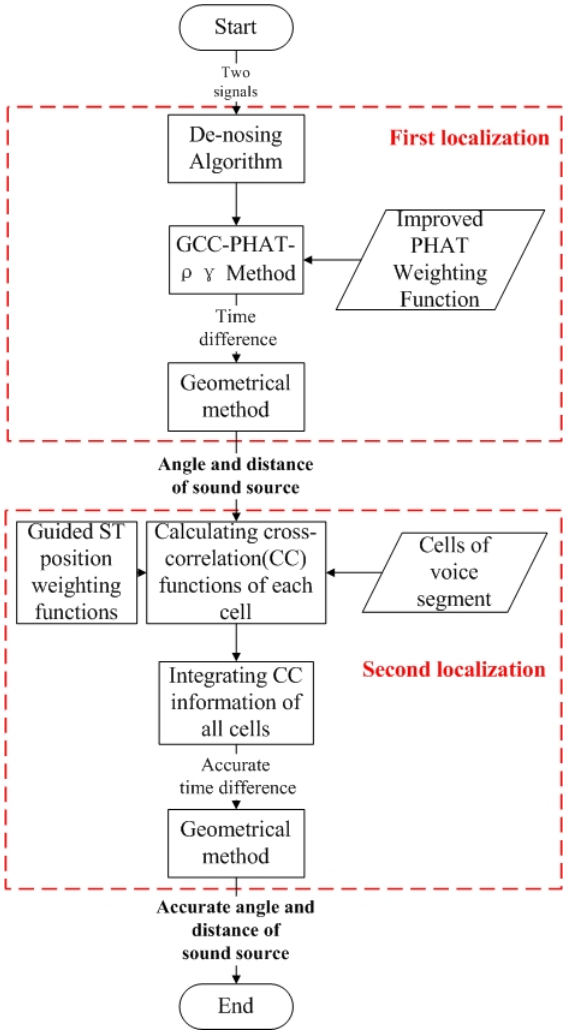


Fig. 3: Flow chart of continuous sound source localization.

III. GUIDED ST POSITION METHOD

The GCC-PHAT-$\rho\gamma$ method above is based on the non-reverberation model. It can not detect the speakers correctly in multi-source or directional interference noise environment. The motor of robot makes noise with fixed frequency while moving and noise sources in room may result in directional interference. Therefore, This section introduces a de-noising algorithm and proposes a method named guided ST position to eliminate noises, room reverberation and calculate the location of sound source, in real-time, continuously.

A. De-noising Algorithm

Signal that microphones receive is voice mixed with noise when there exits noise interference. Comparing to clean speech, the statistical characteristics of noisy speech change according to noise source characteristics, noise statistics law, the noise amplitude, noise interference voice mode, etc. which makes the feature distribution of clean speech change from Gaussian distribution to non-Gaussian distribution. Anyway, model of clean speech fails while the signal is noisy speech. Stationary background noise is mainly processed according to the environment where robots in. Spectral subtraction and cepstral mean normalization method [16] are applied to the algorithm.

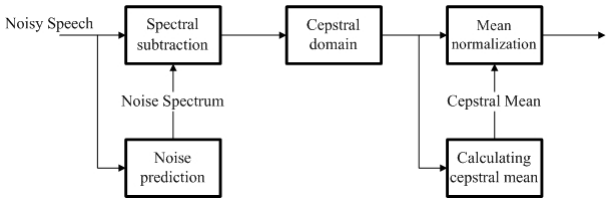


Fig. 4: Block diagram of the de-noising algorithm.

Background noise and speech can be approximated as the sum of the power relations. The principle of spectral subtraction is as follows. By estimating the noise power in each frame and subtracting it from the total power spectrum of the frame, we can get the estimation of pure audio power spectrum, while the phase of clean speech frame is replaced by the phase of noisy speech. Finally the noise can be greatly eliminated. Cepstral mean normalization is also applied to the de-noising algorithm. Its goal is to eliminate bias in the cepstral domain that caused by convolution noise, such as telephone channel, etc. However, it is also strongly affected by dealing with background noise. The block diagram of the de-noising algorithm is shown in Fig. 4.

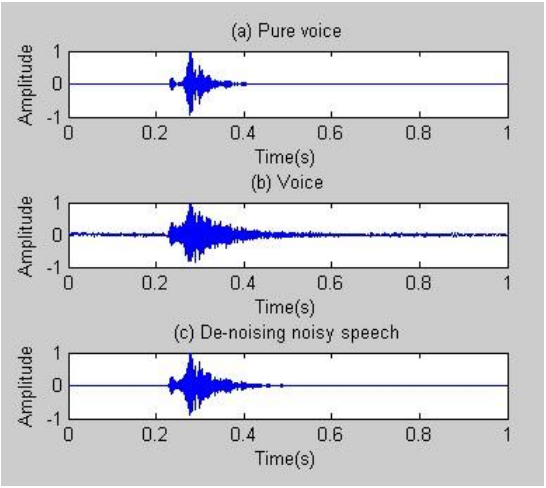


Fig. 5: Time-domain waveform diagram.

Take advantage of this de-noising algorithm, this paper gives the de-noising effect of noisy speech in the time domain. Fig. 5(a), (b) and (c) are a time-domain waveform diagram, describing post-envelope of ups and downs over time of the pure voice, voice and de-noising noisy speech. We can see that waveform of after de-noising speech can be seen clearly. Therefore, the de-noising algorithm can effectively

remove the background noise and has little effect on the signal distortion.

B. GCC-PHAT-$\rho\gamma$ Method

Generalized Cross Correlation-Phase Transform (GCC-PHAT) method through the appropriate improvement can quickly determine the approximate location of sound source. For the classic time difference of arrival algorithm, such as GCC method, the peak position of cross-correlation function is the relative delay between two signals. Firstly, the algorithm calculates the cross-power spectrum and gives a certain weight to curb the impact of noise and reflection in the frequency domain. And then the cross-correlation function is got through transforming the cross-power spectrum to time domain. Suppose the sound source signal is $s(t)$ and it is p-point in space. The position of m-microphone is q_m. The signal that microphone m-microphone received is $x_m(t)$, therefore,

$$x_m(t) = s(t) \otimes h_m(q_m, p, t) + n_m(t) \tag{1}$$

where $h_m(q_m, p, t)$ is the response function between m-microphone and sound source. It includes direct and reverb sound. The additive noise is representative of $n_m(t)$. Assuming that the noise and signal are not related, the above equation can be written as,

$$x_m(t) = \alpha_m s(t - \tau_m) + s(t) \otimes h'_m(q_m, p, t) + n_m(t) \tag{2}$$

The signal $x_m(t)$ microphone received contains direct sound delay and attenuation. The attenuation factor is α_m. $h'_m(q_m, p, t)$ is impulse response between the sound source and the non-direct sound of m-microphone. τ_m is the time delay from sound source to m-microphone. Suppose τ_{mn} is the time difference of sound transmitting from the sound source to m and n-microphone. Therefore, it is useful to get the peak of the following GCC function,

$$R_{mn}(t) = \int_0^{\pi} W_{mn}(\omega) S_m(\omega) S_n^*(\omega) e^{-j\omega t} \tag{3}$$

where $W_{mn}(\omega)$ is the weighting function. For different noise and reverberation conditions, different weighting functions can be used to sharpen the peak of the GCC function. If the cross-power spectrum function of signals $x_m(t)$ and $x_n(t)$ is $G_{mn}(\omega)$, then,

$$\begin{aligned} G_{mn}(\omega) = &\alpha_m \alpha_n S_m(\omega) S_n^*(\omega) e^{-j\omega(\tau_m - \tau_n)} + \alpha_n S_n^*(\omega) N_m(\omega) + \\ &\alpha_m S_m(\omega) e^{-j\omega\tau_m} N_n^*(\omega) + N_m(\omega) N_n^*(\omega) \end{aligned} \tag{4}$$

Phase transform (PHAT) weighting factor can be expressed as,

$$W_{mn}(\omega) = \frac{1}{|G_{mn}(\omega)|} \tag{5}$$

$G_{mn}(\omega)$ and $S_m(\omega)S_n^*(\omega)$ have a greater margin when formula (5) is used in low-SNR situation and the accuracy of location is low. With the decline in SNR, $|S_m(\omega)S_n^*(\omega)|$ in proportion of $G_{mn}(\omega)$ are also gradually declining. Therefore, a parameter ρ is introduced. The value of ρ is determined by the SNR in an actual environments. It is used to make $G_{mn}(\omega)$ which is in the weighting function characterize the cross-power spectrum of sound source signal accurately. When the signal energy is small, the denominator of the weighting function will tend to 0 and the error will be increased. So that the novel method proposed in this paper is to give the denominator of the weighting function a coherent factor. This factor not only reduces the error, but also can not impact the cross-power spectrum. The weighting function has been replaced by:

$$W_{mn}(\omega) = \frac{1}{|G_{mn}(\omega)|^{\rho} + |\gamma_{mn}^2(\omega)|} \quad 0 \le \rho \le 1 \tag{6}$$

It can improve positioning accuracy by using the modified weighting function in this PHAT-$\rho\gamma$ method in small SNR and large reverberation situation.

C. Guided ST Position Algorithm

PHAT-$\rho\gamma$ method described previously is a effective algorithm and it has a strong real-time efficiency. However, the time delay will have more errors when noise and room reverberation more serious. And calculating of angle and distance of sound source depends on the accuracy of time delay. Guided ST position method based on the PHAT-$\rho\gamma$ method can solve this problem. In this method, a local estimate with ST position weighting function is applied to the algorithm. The difference of this method with GCC is that GCC is weighted to the entire voice, while the local estimating method calculates cross-correlation function for every cell of voice segment separately.

Paper [17] proposed a new weighted method which named spectro-temporal position (ST position) and it is a weighting function. As the voice segment has more than one cell, the i-frame and j-band is denoted by $c(i, j)$. Cross-correlation functions of all cells are calculated and weighted using ST position method. Finally, time delay information of each cell is super positioned to form the final time delay.

$$Timedelay = \underset{\tau}{argmax}\left(\frac{1}{\Gamma}\Sigma_{i,j\in P}\Phi_{i,j} \cdot CC_{i,j}(\tau)\right) \tag{7}$$

where P is the set of i, j, and Γ is the number of elements in P. $\Phi_{i,j}$, CC and τ are the weighting function, cross-correlation function and time difference of cell $c(i, j)$, respectively. Q is defined as the cross-correlation coefficient in the sound source localization in every cell.

However, Q has a relationship with frequency f and the distance from starting point of voice segment s. The cell becomes more reliable when its Q-value increases and meanwhile, more resistant to room reverberation. When the cell is getting close to the starting point of the voice segment, the proportion of direct voice in this cell is bigger and the same as its Q-value. The biggest problem of this method is that Q-value is unknown during operation in the algorithm. It needs to statistic using pre-training data and this increases the limitations of algorithm. In this paper, the approximate sound source location has been got using PHAT-$\rho\gamma$ method. Then sequence of voice signal received by different microphones and the reverberation environment of a robot are estimated using the rough sound source location. It is a novel approach

---

Algorithm guided spectro-temporal (ST) position

Required: $i = 0$, $j = 0$, the number of elements (i, j) is Γ, $i \in [o,m]$, $j \in [o,n]$, $sum = 0$ and f is frequency, s is temporal position relative to the start of the voice segment
Preprocessing phase(first localization):
1: Generate sound source location A using PHAT-$\rho\gamma$ method
2: For each A, estimate room reverberation

Query phase(second localization):
1: Define $Q \in [0,1]$
2: Define $c(i, j)$ as the cell in i-frame and j-band
3: For voice segment
4: While ($i < m$ and $j < n$) do
5: Estimate f, s
6: Compute $Q_{ij}(f,s)$
7: Compute cross-correlation function $CC_{ij}(t)$
8: $i = i+1$, $j = j+1$, $sum = sum + Q_{ij}(f,s) * CC_{ij}(t)$
9: End while
10:Compute $sum = sum/\Gamma$, $Timedelay' = \underset{\tau}{argmax}(sum)$
11:End for.

---

to take advantage of this approximate location of sound source to determine the Q-value without advance statistics. The details of this guided ST position algorithm are shown in Algorithm-1. The weighting function is changed as,

$$Timedelay' = \underset{\tau}{argmax}(\frac{1}{\Gamma}\Sigma_{i,j\in P}Q(f,s)\cdot CC_{i,j}(\tau)) \quad (8)$$

The problem of ST position weighting method is solved and the computational complexity of algorithm is not excessively increase through experimental tests. Weight each cell using guided ST position weighting method and integrate all cells to get the peak value of cross-correlation function which is the accurate time delay.

IV. EXPERIMENTS AND ANALYSIS

A. Configuration of Experimental Environment

A sound localization system is designed on a mobile robot using a MARIAN TRACE8 multi-channel audio sample card and a crossing field of four BSWA MPA416 microphones. The processor model of the computer is Q6600 and the computer memory is 4G. Multi-threading programming based on Direct Sound is adopted to ensure the synchronization of the audio signals. Our scene graph of robot and microphone array model is shown in Fig. 6.

Fig. 6: Scene graph of robot and microphone array model.

The experimental environment is a room of $8m \times 8m$. Combining with the robot's size, the diagonal distance of the microphones is designed as 40cm. In addition, the data sampling rate will largely affect the performance of localization. However, higher data sampling rate means that more data points need to be analyzed and calculational complexity will increase. Therefore, this system selects the data sampling rate of 44.1 kHz. As illustrated in Fig. 7, the microphones are placed on the shoulder of the robot with a height of $100cm$. Given that the height of mouth is about 150 cm for a standing adult, the plane of the microphone field is chosen as the standard when evaluating the localizing performance on the horizontal plane and all the localizing results have to be projected on to this standard plane. The robot is placed in the center of the room, with another 72 points in the floor as testing sound sources. The sound sources are evenly distributed in every 15° and three positions in each direction, with a distance of 1, 2, 3 m from the center, respectively. The following three groups of highly-targeted experiments are designed to test the performance of this method precisely.

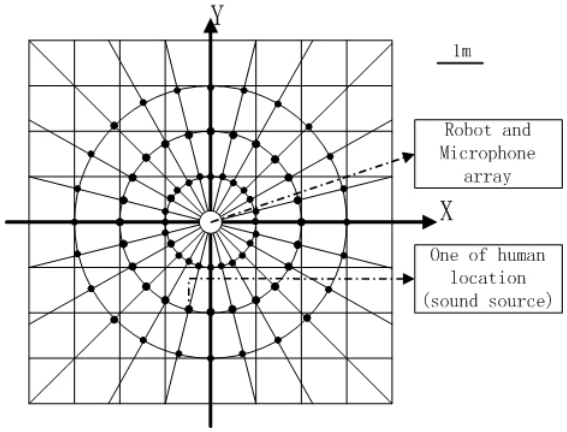


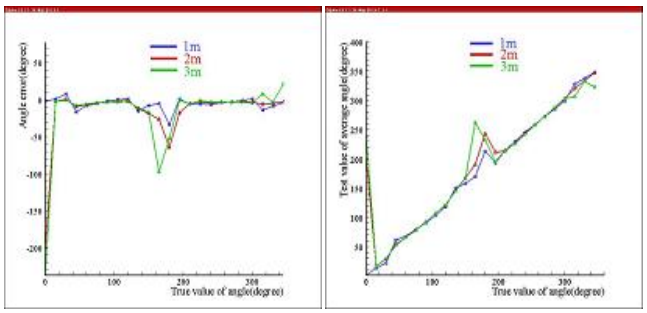
Fig. 7: Layout of robot microphone field and testing points.

B. Continuous Localization on Stationary Robot

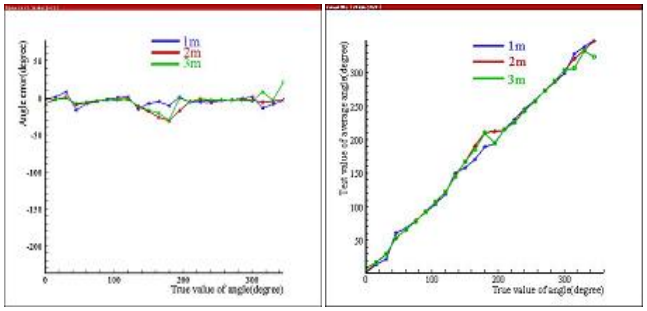
First group is carried out under stationary situation. Although people and robot in this group have fixed position, one speaker can speak many times continuously on one point in a short time, also, different speakers on different points can speak one by one in a short time, continuously. Algorithm is able to get endpoints of voice segments in time and estimate the location of speakers, continuously. Assume that the direction X in Fig. 7 is the direction of 0°, data is performed sampling every 15°. The robot stands in the center of the layout while the speakers are standing at the black points. Each point is sampled 28 times and the content of speech is Chinese words ”dingwei”, ”pengpeng”, ”nihao” and ”qingzhuyi” which mean ”location”, the robot's name, ”hello” and ”pay attention”, respectively. Every word is sampled 7 times on every point and the whole experiment is carried out by 3-5 persons. Finally 2016 groups of data are obtained. The localizing performance is evaluated in three situations with different SNRs: at night with air conditioner off and the SNR is about 40, normal in-door environments with 25 of SNR, music noise and SNR is 10.

This group is aimed at comparing the three localization algorithms under different SNRs and different distances between sound source and microphone fields. From Table I, the results of sound localization for stationary robot show that the guided ST position algorithm gives the best performance for all the tests and does not increase the computational complexity severely. All the calculations can be finished within 0.4 seconds.

Fig. 8(a) and (b) show the results of guided ST position method. The distances of sound source and the robot are 1, 2, 3 *m* which are represented by three lines with blue, red and green color, respectively. This figure shows that the overall volatility of the angle error is bigger when the sound source is farther away from the robot. Good results are obtained at 0° and 180°, however, for the other true values of angles, the positions of 2 and 3 *m* have small angle error. When the sound source is farther away from the robot, the test results of sound source at 0° and 180° are worse. The reason for this result is at 0° and 180° around, the choice of microphone pair caused huge errors on time difference. The improved method is to add another group of microphone pair to calculate the time difference, however, this will increase the computational complexity. Fig. 8(c) and (d) are the improved results.



(a) Angle and error. (b) Angles of true and test value.



(c) Improved angle and error. (d) Improved true and test angle value.

Fig. 8: Angle results of guided ST position method.

C. Continuous Localization on Mobile Robot

This group is second group and carries out the localization on mobile robot. In this group, People participating in the experiment can speak in a short time disorderly and continuously. However, The room reverberation changes while speakers are walking and speaking continuously, so that guided ST position method is used to estimate the reverberation for real-time and determine the accurate location of sound source continuously. Since the purpose of experiment is to calculate the relative location of robot and speakers,

TABLE I: Angle localization results of stationary robot

		One meter	Two meters	Three meters	Average
PHAT	SNR=40	88.10% (592)	89.14% (599)	87.80% (590)	88.34% (594)
	SNR=25	86.62% (582)	87.95% (591)	85.27% (573)	86.62% (582)
	SNR=10	83.78% (563)	85.12% (572)	80.51% (541)	83.14% (559)
$PHAT-\rho\gamma$	SNR=40	92.26% (620)	93.01% (625)	90.03% (605)	91.77% (617)
	SNR=25	89.88% (604)	91.96% (618)	88.69% (596)	90.18% (606)
	SNR=10	86.31% (580)	87.50% (588)	85.27% (573)	86.36% (580)
Guided ST Position	SNR=40	94.49% (635)	96.43% (648)	90.77% (610)	93.90% (631)
	SNR=25	91.07% (612)	95.09% (639)	89.88% (604)	92.01% (618)
	SNR=10	90.47% (608)	92.26% (620)	88.98% (604)	90.57% (609)

TABLE II: Angle localization results of mobile robot

		Correct proportion (Angle error less than 15°)
PHAT	SNR=40	80.8% (194/240)
	SNR=25	77.9% (187/240)
	SNR=10	69.2% (166/240)
$PHAT-\rho\gamma$	SNR=40	89.6% (215/240)
	SNR=25	82.9% (199/240)
	SNR=10	77.1% (185/240)
Guided ST Position	SNR=40	92.5% (222/240)
	SNR=25	90.8% (218/240)
	SNR=10	90% (216/240)

this group does not move the robot in first part. However, the motor of robot works to simulate mobile situation and the speakers walk slowly in the room, so that speakers and robot have a relative motion. The speakers must only speak on the 24 specified directions defined previously is to carry out the location results compared with the true value conveniently. Each angle is sampled 10 times and 720 groups data are obtained. This group also needs 3-5 persons and three different situations with SNRs of 10, 25 and 40 are also calculated. The results of three methods with different SNRs are compared in Table II.

Table II shows that guided ST position method reaches a best result. What's more, unlike other methods, when the noise is serious, the location performance of guided ST position method does not decline seriously. For each angle, Fig. 9 shows the group numbers which the difference of test value and true value is less than 15° by using guided ST position method when SNR is 25. From this figure, it is clear that the sound source at 180° gets the worst result. If more tests are carried out, this phenomenon which also appears in first group will more obvious. Also, it can add another group of microphone pair to improve the method.

The other part of experiment on mobile robot is as follows. The robot moves slowly along the direction X, while the speakers walking and only one speaking every time in this $8m \times 8m$ room. Also, speakers can generate a number of sound source. The main noises come from not only the air conditioning, but also the body noise generated by robot. However, more than 200 groups of tests are carried out and every time the robot can get accurate sound source using guided ST position method. This method can get accurate results of localization even when the robot and people are in continuous motion because it is real-time and all the calculations can be finished in 0.4 seconds.

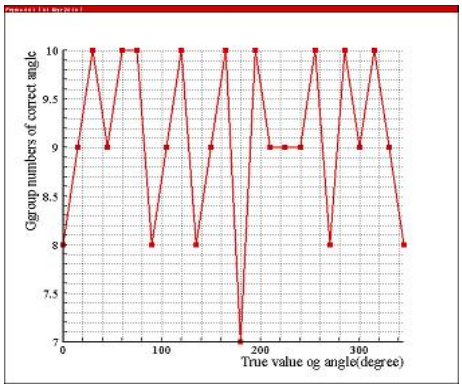


Fig. 9: Angle results using guided ST position method.

D. Hide-and-seek Game

A game named hide-and-seek is designed on this robot for continuous localization which combines with speech recognition and hand detection technology. In the room of $64M^2$, robot locates in the center of the room first and five people or more around it. It can identify orders, such as "localization", "turn left" and so on. When robot moves according to the orders, the people located in the different places put up hands and cry "localization", robot could find the sound direction and turn to the people. When it faces to the people directly, the camera will detect the people who puts up the hands, and move towards him. Fig. 10 shows robot and human hide-and-seek experiment.

Fig. 10: Robot and human hide-and-seek experiment.

Anyway, for the different height of the people, the height of speaker's mouth couldn't keep 150 *cm* accurately. Also, the detector couldn't stand at the accurate place. Therefore, it is inevitable to lead to the mistake. However, from an experimental point of view, these mistakes are permitted in controllable range. What's more, results in above groups of experiments show that 90% data whose angle errors is greater than 15° have angle errors of 25°. Due to camera's detection range in our system is $\pm 23°$, therefore, for future work, it can locate sound source using audio-visual fusion mechanisms and making the visual as a pretreatment immediately and accurately. Meanwhile, multiple sources can also be considered.

V. CONCLUSIONS

A sound localization method is proposed for mobile robot. The heart of the methodology is guided ST position localization based on a microphone array of four microphones.During the implementation, the movements of the robot and human have different SNR, which should be taken into consideration. A comparison between our PHAT-$\rho\gamma$ method and other implementations is also taken in experiments. The results of our experiments demonstrate that this system provides better performance in aspects of localization angel, real-time and audio-visual fusion.

VI. ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China(NSFC, No.60875050, 60675025), National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247), Shenzhen Scientific and Technological Plan and Basic Research program (No.JC200903160369A), Natural Science Foundation of Guangdong (No.9151806001000025).

REFERENCES

[1] J.C. Chen, Kung Yao and R.E. Hudson, Source localization and beamforming,*IEEE Signal Processing Magazine*, vol.19, No.2, March 2002, pp.30-39.

[2] I.E. Robert, "Robust sound localization: An application of an auditory perception system for a humanoid robot", *MIT Department of Electrical Engineering and Computer Science*, 1995.

[3] K. Nakadai, H. Nakajima and M. Murase, "Real-Time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays", *in IEEE/RSJ International Conference on Intelligent Robots and Systems* , Beijing, China, September 2006, pp.852-859.

[4] J.M. Valin, F. Michaud, B. Hadjou, Robust localization and tracking of simulaneous moving sound sources using beamforming and particle filtering, *Robotics and Autonomous Systems*, vol.55(3), March 2007, pp.216-228.

[5] I. Hara, "Robust speech interface based on audio and video information fusion for humanoid HRP-2", *in IEEE/RSJ International Conference on Intelligent Robots and Systems* , 2004, pp2404-2410.

[6] T. Lobos, Z. Leonowicz, J. Rezmer, and P. Schegner. High-resolution spectrum-estimation methods for signal analysis in power systems. *IEEE Trans. on Instrumentation and Measurement*, vol.55(1), 2006, pp.219-225.

[7] J.M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach", *in Proceedings IEEE International Conference on Robotics and Automation*, vol.1, 2004, pp.1033-1038.

[8] C.H.Knapp, G.C. Carter, "The generalized correlation method for estimation of time delay", *in IEEE Transactions on Acoustics, Speech and Signal Processing*, 1976, pp.320-327.

[9] Q. H. Wang, T. Ivanov, and P. Aarabi, "Acoustic robot navigation using distributed microphone arrays," *in Information Fusion*, vol.5, June 2004, pp.131-140.

[10] R. Ratnam, D.L. Jones, and W.D. OBrien, Fast algorithms for blind estimation of reverberation time, *IEEE Signal Processing Letters*, vol.11(6), June 2004, pp.537-540.

[11] P. Aarabi and S. Zaky, "Robust sound localization using multi-source audiovisual information fusion", *in Information Fusion*, 2001, pp.209-223.

[12] J.M. Valin, F. Michaud, J. Rouat, and D. Ltourneau, "Robust sound source localization using a microphone array on a mobile robot", *in IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, pp.1228-1233.

[13] Y. Tamai, S. Kagami, Y. Amemiya and Y. Sasaki, "Circular microphone array for robot's audition", *in Proceedings of IEEE Sensors*, 2003, pp.1100-1105.

[14] H. Liu, Acoustic positioning using multiple microphone arrays, *Journal of the Acoustical Society of America*, vol.117(5), 2005, pp.2772-2782.

[15] J.W. Duan, Y.C. Shi, X.J. Chen, "Study on the directing performance of the linear microphone array", *in China National Computer Conference*, 2007.

[16] J.S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise", *in IEEE Transactions on Acoustics, Speech and Signal Processing*, 1978, pp.471-472.

[17] H. Christensen, N. Ma, S.N. Wriqley and J. Barker, "A speech fragment approach to localising multiple speakers in reverberant environments", *in IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp.4593-4596.