2013 IEEE International Conference on Robotics and Automation (ICRA)
Karlsruhe, Germany, May 6-10, 2013

A Two-Layer Probabilistic Model Based on Time-Delay Compensation for Binaural Sound Localization

Hong Liu[1], Zhuo Fu[2] and Xiaofei Li[3]

Abstract— **Interaural Intensity Difference (IID) and Interaural Time Difference (ITD) are two improtant cues for robot acoustic localization both in Artificial Intelligence (AI) and Human-Robot Interaction (HRI) areas. However, it is a challenge job to localize a sound source accurately and swiftly only by two acoustic sensors. In this paper, a time-delay compensation based two-layer probabilistic model is presented for binaural sound source localization. In the first layer, a weighting function of Generalized Cross Correlation (GCC) named PHAT-$\rho\gamma$ is used in low-frequency to obtain the prior time-delay. And in this layer a crude estimate of azimuth can also be acquired. At the same time, the probability of all possible time-delay lags can be achieved from the training data. In the Second layer, a new improved algorithm of IID based on time-delay compensation(named $IID^{\sigma\tau}$) is introduced to refine the probability of the azimuth and the elevation. Lastly, localization result is obtained by Bayes-Rule method. Comparing with three state-of-art algorithms, experimental results show that the proposed method has higher accuracy and costs less time for sound source localization.**

I. INTRODUCTION

As an important part of AI and HRI, auditory system introduces a new area of research for robot perception technologies. In order to realize the capability of accurate sound localization, algorithms of geometric localization using more than two acoustic sensors [1][2] are often used. However, it is more difficult for robots to recognize sound signals accurately and swiftly only by two acoustic sensors, just as human and other mammals do. Many fields can benefit from the capability of pinpointing the sound source swiftly and accurately. For instance, a predator can precisely locate its prey in the wild [3] and many other applications such as [4][5].

There are three important and difficult issues concerning binaural localization of a robot: Firstly, how to accurately localize any kind of sound source; Secondly, how to localize serval different sound sources at the same time; Thirdly, how to track one or serval moving sound sources. Aiming at dealing with these problems, the research on sound source localization by only two acoustic sensors has already been studied for several decades.

As early as one hundred years ago, “Duplex Theory”[6], shows that ITD could be used to localize sound in low frequency, while IID could be applied to the localization of sound in high frequency. Most of existing binaural cues used by ITD were all based on the coincident model proposed by Jeffress[7]. However, “Duplex Theory” neither can distinguish whether the sound comes from front or rear, nor can localize the elevation. To solve this problem, researchers pay more attention to the effects of torso, shoulders, head, and outer ears or pinnae of human body for sound source localization. The cochlear model, developed in 1988 by Lyon [8-10], made it possible to analyze ITD and IID in different sub-bands. Based on cochlear model, a number of monaural and binaural cues for evaluating azimuth, elevation and distance were researched [11-13], which were named Head-Related Transfer Functions (HRTFs).

Cohen [14] elaborated the different representations of auditory space in the midbrain and forebrain. A brain-like neural network for periodic analysis was introduced by Voutsas [15]. From then on, the computational strategy introduced in [14,15] has been widely used for binaural sound source localization. A variety of localization cues were measured, such as ITD and IID, and they would be trained to be templates. Those cues obtained from the sound source to be localized could be grouped together and then be matched with the templates[16-18], and finally the location of sound source would be found from the matching templates. An overview of majority of the binaural localization models could be found in [19].

However, there are two problems in traditional methods. Firstly, it is time-consuming to match the characteristic vector of sound source to all the templates that characterize from each direction. Secondly, ITD and IID are acquired without considering the influence between each other. However, with the influence of ITD, the signals received by two ears have different starting point on sound source. This difference will effect the extraction of IID. To address this issue, several new algorithms were put forward. For instance, Li [20] introduced a Bayes-Rule based three-layer hierarchical system for binaural sound source localization. In their work, ITD, IID, and Spectral Cues were used in three different layers to acquire the information of the direction. Lower layer provided upper layer with candidates, and Bayes-Rule based approach was employed to make the final decision, which could reduce time consumption effectively. Willert introduced a probabilistic model [21] for binaural sound

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), Scientific and Technical Innovation Commission of Shenzhen Municipality (No.JC201005280682A, CXC201104210010A).

[1] H. Liu is with Faculty of Key Laboratory of Machine perception and Intelligence, Peking University, Shenzhen Graduate School, Beijing, 100871 CHINA. hongliu@pku.edu.cn

[2] Z. Fu is with the Engineering Lab on Intelligent Perception for Internet of Things(ELIT), Shenzhen Graduate School of Peking University, Shenzhen, 518055 CHINA. gradyfu@hotmail.com

[3] X. F. Li is with the Engineering Lab on Intelligent Perception for Internet of Things(ELIT), Shenzhen Graduate School of Peking University, Shenzhen, 518055 CHINA. lixiaofei0111@sohu.com

978-1-4673-5642-8/13/$31.00 ©2013 IEEE

localization in the frontal azimuthal half-plane. In his paper the effect of time delay was considered in IID acquisition. Analysis of signals in short time makes it possible to reduce computation. Meanwhile, Finger [22] introduced the approaches and databases for online calibration of binaural sound localization for robotic heads by two-layer model. The result of the first layer gives a prior to the second layer. However, noise would degrade the performance of sound source localization. Jeub proposed a model [23] for binaural cues dereverberation preserving. Actually, each of algorithms above can only solve one of two problems mentioned in the pervious paragraph. This work will resolve both of them.

A two-layer probabilistic model will be presented in this paper to reduce the matching time between new features and templates. Then a new IID based on time-delay compensation, named $IID^{\sigma\tau}$, will be applied to eliminate the influence of time-delay to IID.

In the first layer, a weighting function of GCC named PHAT-$\rho\gamma$ [24] is adopted to calculate the crude value of azimuth which is though as a prior value. However, there are measurement errors between the prior value with the real azimuth of source. Therefore, all the possible azimuths will be achieved by training with a certain probability. The time-delay for each azimuth can also be obtained through training in advance. Once a prior value of azimuth is acquired, all the possible candidate azimuths and time delays can be found in the database. They will be prepared for the second layer as time-delay compensation.

In the second layer, the information of time delays and their possible azimuths will contribute to obtaining IID to offset the influence of time delay, which is named $IID^{\sigma\tau}$. After obtaining all possible time delays and their standard deviations of each candidate direction, $IID^{\sigma\tau}$ of the signals received by two ears will be calculated. As a result, the new feature vector will be extracted. Since ITD offsets the influence of time delays, it will be more effective than traditional IID. Then the probability can be achieved when $IID^{\sigma\tau}_{real}$ matches the template ($IID^{\sigma\tau}_{mod}$). At last Bayes-Rule method is used to make the final decision. The result will be compared to other state-of-art sound source localization models[20-22].

The rest of this paper is organized as follows: Section II gives a real knockdown to the model and the algorithm used in this paper. Experimental results and analysis are shown in section III. At last, the conclusions are drawn in Section IV.

II. TWO-LAYER BASED PROBABILISTIC MODEL

Denote the source signal as $s(n)$, and the received signals as $x_l(n)$ and $x_r(n)$ on the left and the right ears respectively. Then it can be described as follows:

$$x_l(n) = h_l(\theta, \varphi, r) * s(n) + \eta_l \tag{1}$$

$$x_r(n) = h_r(\theta, \varphi, r) * s(n) + \eta_r \tag{2}$$

where $h_l(\theta, \varphi, r)$ and $h_r(\theta, \varphi, r)$ are the transfer functions of the direct paths from source to the two ears, which rely heavily on azimuth, elevation and distance. η_l and η_r are the noises received by left ear and right ear. θ, φ, r are the azimuth, elevation and distance respectively.

With the effect of the head shadow, it was proved that IID cues on frequency makes it possible to estimate the distance to the sound-source for the range going from 1 to 2 meters from the listener [25]. However, when r is lager than $2m$, $h_l(\theta, \varphi, r)$ and $h_r(\theta, \varphi, r)$ almost alter indistinctly. In this paper, the CIPIC database is used, so all sounds used in this paper share a same distance of 1m. Therefore, there are good grounds for ignoring the effect of r in this paper. Without considering r, the task of this work is to obtain the θ and φ from $x_l(n)$ and $x_r(n)$. Then (1) and (2) can be simplified as follows:

$$x_l(n) = h_l(\theta, \varphi) * s(n) + \eta_l \tag{3}$$

$$x_r(n) = h_r(\theta, \varphi) * s(n) + \eta_r \tag{4}$$

A. General Structure

The two-layer based probabilistic model for binaural sound localization presented herewith is composed of three distinct and consecutive processors :

- Candidate azimuth extracting unit: This unit is used to process a pair of original signals $x_l(n)$ and $x_r(n)$ received by two ears of robots and to extract the probability of candidate azimuths θ_i. This work is done in the first layer and prepares for the next unit;
- Direction refining unit: Its task is to calculate the probability of the sound source which is located on a certain elevation φ by matching real $IID^{\sigma\tau}$ with template $IID^{\sigma\tau}_{mod}$ in database. This unit works in the second layer;
- Decision-making unit: In this unit a Bayesian sensor model is employed to decide the direction where the sound source is. In the next three subsections, the three units will be presented in detail.

B. Candidate Azimuth Extracting Unit

The aim of this section is to provide the second layer with the candidate azimuths θ_i and their probabilities. In order to get the candidate azimuth θ_i, it is necessary to remove the effect of φ. In Fig.1, average lags of time delays for each 1250 directions (25 azimuths × 50 elevations) are displayed. There are two obvious phenomenons as following: First, different elevations with the same azimuth share the same time delay or lags; Second, the lags of ITDs vary systematically with the angle of incidence of the sound wave relative to the interaural axis.

Theoretical evidence shows that the lags are independent of frequency in previous works [11-19], which is according to the fact that low-frequency sounds travel more easily around the head and the time differences or lags are nearly unaffected by φ. Therefore, the lag in each frequency channel bank is regarded to be equal to the lag in low frequency channel ($<$ 1500 Hz) approximatively. Accordingly (3) and (4) can be simplified as follows:

$$x_l(n) = h_l(\theta) * s(n) + \eta_l \tag{5}$$

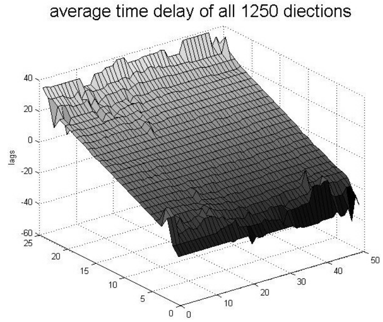

Fig. 1. Average time delay of 25 different azimuths and 50 different elevations in 3D space.

$$x_r(n) = h_r(\theta) * s(n) + \eta_r \tag{6}$$

In this paper, instead of calculating ITD, Time Delay Of Arrival (TDOA) is employed to get a priori azimuth Θ. Then the geometry model becomes very simple. (5) and (6) can also be described as follows:

$$x_l\left(n\right) = a_l s\left(n - \tau_l\right) + \eta_l \tag{7}$$

$$x_r\left(n\right) = a_r s\left(n - \tau_r\right) + \eta_r \tag{8}$$

where a_l and a_r denote the attenuation factors from the sound source to the two acoustic sensors. And $\Delta\tau$ in (9) is the lag that needs to be found.

$$\Delta\tau = \tau_r - \tau_l \tag{9}$$

In order to obtain $\Delta\tau$, a weighting function of Generalized Cross Correlation (GCC) named PHAT-$\rho\gamma$ [24] based on Time Delay Of Arrival (TDOA) is employed in this unit. If it is assumed that (7) and (8) can be approximated by a pure time delay, the cross-correlation between the signals received by the two ears can be represented as follows:

$$R\left(n\right) = \int_{-\pi}^{\pi} W(\omega) X_l\left(\omega\right) X_r{}^*\left(\omega\right) e^{-j\omega n} d\omega \tag{10}$$

where $R\left(n\right)$ is the correlation function of signals between two ears. $W(\omega)$ is the weighting function for sharping the peak of GCC function. $X_l\left(\omega\right)$ and $X_r\left(\omega\right)$ are the power spectrum function of signals between the two ears.

With high signal to noise ratio, phase transform (PHAT) weighting factor is always used as $W(\omega)$ for accurate sound source localization.

$$W\left(\omega\right) = \frac{1}{\mid G(\omega) \mid} = \frac{1}{\mid X_l\left(\omega\right) X_r{}^*\left(\omega\right) \mid} \tag{11}$$

By using the weighting function $W\left(\omega\right)$, the cross-power spectrum of sound source signal can get a robust time delay result. However, When the signal energy is small, the denominator of the weighting function will tend to 0 and the error will increase. In [24], the improvement of $PHAT$ was proposed. A new parameter ρ is introduced into $PHAT$, whose value is determined by the SNR in actual environment. The novel method gives the denominator of the weighting function a coherent factor. Not only this factor could reduce the error, but also it could avoid impacting the cross-power spectrum. More details about PHAT-$\rho\gamma$ can be found in [24]. As a result, the weighting function has been replaced by:

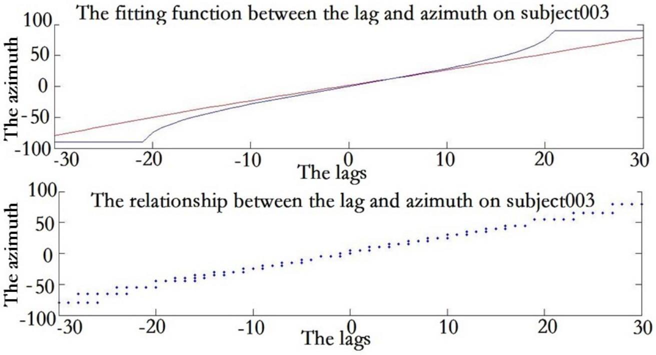

Fig. 2. The upper one show the fitting function between the time-delay lags with the real azimuths, which the red one stands for the polynomial function we used and the blue one represents the arcsine function used in traditional methods. The lower one show the relationship between the time-delay lags with the real azimuths in actual condition.

$$W\left(\omega\right) = \frac{1}{\mid G(\omega) \mid^{\rho} + \mid \gamma^2\left(\omega\right) \mid} \quad 0 \leq \rho \leq 1 \tag{12}$$

Then the lags between two ears can be obtained as:

$$\Delta\tau = arg \max_{n} R\left(n\right) \tag{13}$$

Considering the geometrical relationship of time delay with azimuth, the relationship between $\Delta\tau$ and θ can be described as follows:

$$\Theta = sin^{-1}\left(\frac{\Delta d}{d}\right) = sin^{-1}\left(\frac{\Delta\tau c}{d f_s}\right) \tag{14}$$

where c is the speed of sound in air (344 m/s), Δd is the distance difference between the sound source and two ears. Obviously when $\frac{\Delta\tau c}{d f_s}$ is near or over 1, an error result may be suffered. Therefore, (14) will be rejected and a quartic polynomial by curve fitting will be employed in this work. The polynomial function will be different according to different subjects. In Fig.2, the red curve represents the polynomial by curve fitting, and the blue one denotes arcsine function. The lower picture displays the relationship between the lags and azimuth in each elevations in the ideal state. The result proves that the polynomial by curve fitting is more effective to simulate the relationship between the lags and the azimuths than the arcsine function.

In this paper, azimuth is divided into 25 intervals at first, with center azimuths being -80^o, -65^o, $-55^o, -45^o : 5^o : 45^o$, 55^o, 65^o, 80^o. When a new source appears, its azimuth will be calculated and denoted the nearest center azimuth Θ. Due to the existence of noise in the environment and the possibility of error in calculation, azimuth localization inaccuracy may happen. All possible time delay will be trained when azimuth is calculated as Θ. Because each lag of time delay corresponds a only azimuth θ_i, so the probability

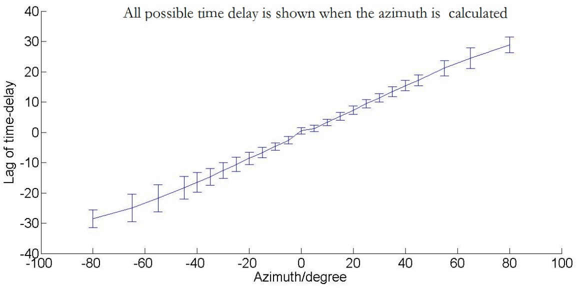

Fig. 3. All possible time delay is shown when the azimuth is calculated as Θ

of the azimuth θ_i by $p(\theta_i|\Theta)$ can be trained and stored before testing with number of different sound signals. In Fig.3, all possible lags of time delay are shown when the azimuth is calculated as Θ. For example, Θ is -80^o, which in fact the azimuth of the sound may come from -80^o or -65^o.

C. The Direction Refining Unit

Because the candidate azimuths are refined in the direction refining unit, therefore, the performance of a sound source localization relies heavily on the accuracy of the extraction of Interaural Intensity Difference. The difficulty in this context is that binaural intensity differences strongly vary over frequency and cannot be unambiguously assigned to the sound source position. The common solution is to decompose the binaural signals into different frequency channels, and to obtain IID in each channel respectively. In this paper, the information of time delay is introduced to compensate IID, and $IID^{\sigma\tau}$ will be acquired, in which τ means the time delay that needs to be compensated, while σ represents the deciding factor of the range of τ.

The first stage of the direction refining unit consists of cochlear and auditory periphery processing, which produces an auditory image model [26]. The AIM processor implements a functional model of a cochlea that decomposes the original signals received by two ears into k frequency channels with respective frequency centre f_c^m and bandwidth b_c^m, $m = 1\ 2 \cdots k$, where k is the number of the cochlea filterbank. The filterbank achievement is based on equivalent rectangular bandwidth (ERB)-filters. The f_c^m and b_c^m can be computed by using the Glasberg and Moore parameters [27].

In the second stage, a new algorithm named $IID^{\sigma\tau}$ is introduced. It is an improvement of the algorithms introduced in [21]. The characteristic $IID^{\sigma\tau}$ is extracted on all possible time-delay lags in all frequency channels.

In order to compensate IID, it is needed to obtain all possible time delays or lags, and the range of τ should be determined first. In this research, when the candidate azimuth θ_i is obtained in the first layer, all possible τ are tested for each φ. Then the mean value $\overline{\tau_i}$ and standard deviation σ_i are obtained. The available interval is shown as follow:

$$\tau \subseteq (-3\sigma_i + \overline{\tau_i}, 3\sigma_i + \overline{\tau_i}) \quad when\ \theta = \theta_i \tag{15}$$

where $\overline{\tau_i}$ is the average time lag between two ears on azimuth θ_i. $\overline{\tau_i}$ and σ_i will change in compliant with θ_i. In Fig.4 and Fig.5, $\overline{\tau_i}$ and σ_i in different azimuths are shown. The upper-lower limit of the range of τ_i will be trained and stored in the database for each candidate θ_i respectively.

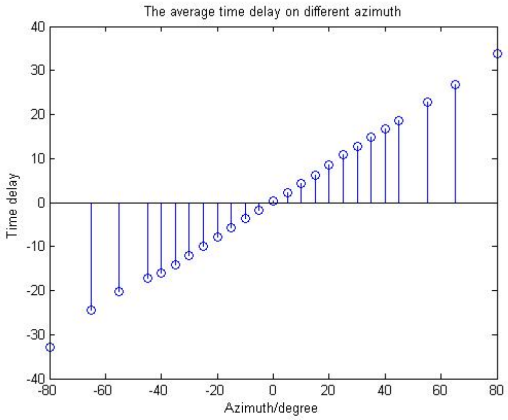

Fig. 4. The average time delay on different azimuth

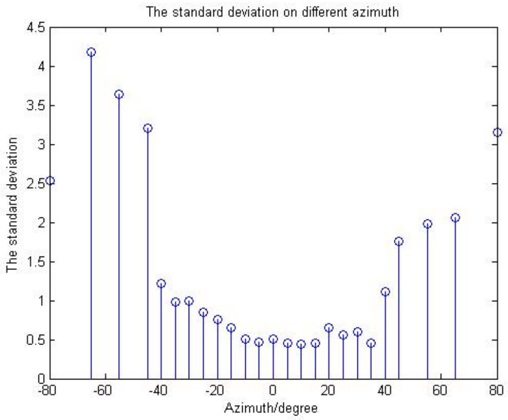

Fig. 5. The standard deviation on different azimuth

Then the $IID^{\sigma\tau}$ is extracted in each frequency bank. In this section, a square window W is used to limit the length of signal, with the length of 128.

In ideal conditions, if the Interaural Intensity Difference is ignored and it is assumed that ITD is $\Delta\tau$ in the m^{th} channel, after time delay compensation, an equation can be achieved as follows:

$$W \odot x_l^m(n - \Delta\tau) = W \odot x_r^m(n) \tag{16}$$

where $W \odot x(n)$ denotes a element-wise multiplication of W and $x(n)$. $x_l^m(n)$ and $x_r^m(n)$ denote the signals that are decomposed into the m^{th} frequency channels.

In fact, the level difference of the amplitudes can not be ignored in binaural sound localization. Noise always has a strong effect on the signals received by two different ears. With this consideration (16) can be transformed into:

$$W \odot x_l^m(n - \Delta\tau) = \lambda_m W \odot x_r^m(n) + \Delta\eta \tag{17}$$

where λ_m denotes the real $IID^{\sigma\tau}$ in m^{th} channel. And $\Delta\eta$ represents the difference between the noises received by two ears. (17) can be rewritten as follows:

$$\begin{aligned}\Delta\eta &= \eta_r - \eta_l \\ &= W \odot x_l^m(n-\Delta\tau) - \lambda_m W \odot x_r^m(n)\end{aligned} \tag{18}$$

where η_r and η_l both are zero-mean Gaussian noises, and $\Delta\eta$ means the disparity between η_r and η_l. Therefore, $\Delta\eta$ is also a zero-mean Gaussian noise, and its standard deviation can be described as $\sigma_{\Delta\eta}$. Since $\Delta\tau$ compensates the influence of ITD effectively, $\Delta\eta$ will achieve minimum value. Therefore, the parameter λ_m in model can be estimated by maximum likelihood estimate method as follows:

$$\widehat{\lambda_m} := \min_{\lambda_m} ||W \odot x_l^m(n-\Delta\tau) - \lambda_m W \odot x_r^m(n)||^2 \tag{19}$$

$\widehat{\lambda_m}$ can be obtained by partially differentiating (19) with respect to λ_m , setting these partial derivatives to zero and analytically solving the resulting equations:

$$\begin{aligned}0 &= \frac{\partial}{\partial\lambda_m}||W \odot x_l^m(n-\Delta\tau) - \lambda_m W \odot x_r^m(n)||^2 \\ &= \frac{\partial}{\partial\lambda_m}\sum_N (x_l^m(n-\Delta\tau) - \lambda_m x_r^m(n))^2 \\ &= \frac{\partial}{\partial\lambda_m}\sum_N (x_l^m(n-\Delta\tau)^2 + (\lambda_m x_r^m(n))^2 - 2\lambda_m x_r^m(n) x_l^m(n-\Delta\tau)) \\ &= \sum_N 2\lambda_m x_r^m(n)^2 - \sum_N 2x_r^m(n) x_l^m(n-\Delta\tau) \\ &= \lambda_m \sum_N 2x_r^m(n)^2 - \sum_N 2x_r^m(n) x_l^m(n-\Delta\tau)\end{aligned} \tag{20}$$

The result is shown as follows:

$$\widehat{\lambda_m} = \frac{\sum_N x_r^m(n) x_l^m(n-\Delta\tau)}{\sum_N x_r^m(n)^2} \tag{21}$$

where N denotes the length of the window.

In order to make the ratio values in a proper range, the logarithmic ratio is calculated as final result of the $IID^{\sigma\tau}$. Then $IID^{\sigma\tau}$ can be described as follows when the candidate azimuth is θ_i gotten in first layer.

$$\begin{aligned}IID^{\sigma\tau}(\Delta\tau, f_m)|_{\theta_i} =&\ 20log_{10}(\widehat{\lambda_m})|_{\theta_i} \\ &= 20log_{10}\frac{\sum_N x_r^m(n) x_l^m(n-\Delta\tau)}{\sum_N x_r^m(n)^2}|_{\theta_i} \\ &\Delta\tau \in (-3\sigma_i + \overline{\tau}_i, 3\sigma_i + \overline{\tau}_i) \\ &m \subseteq (1,k)\end{aligned} \tag{22}$$

where k denotes the frequency channels number of the cochlear filter.

Test on subject 003 in the CIPIC database, the $IID^{\sigma\tau}$ with azimuth -30^o, elevation 0^o and 61.875^o are shown in Fig.6. It can be found that

$$\{\theta_i, \Delta\overline{\tau}_i, \sigma_i\} = \{-30^o, -11, 1\} \tag{23}$$

without considering σ_i, although IID of elevation 0^o and IID of elevation 61.875^o are different, there is still a great similarity. Actually with the help of σ_i, more information can be extracted effectively to distinguish different elevations.

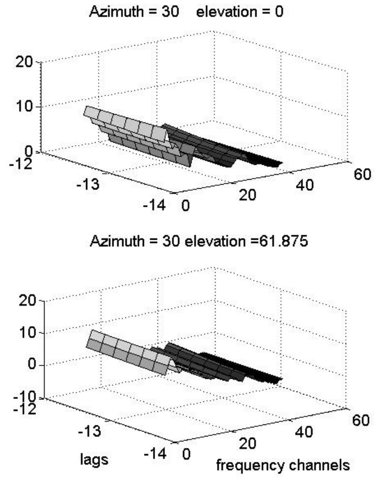

Fig. 6. The $IID^{\sigma\tau}$ with $(\theta,\varphi) \in \{(-30,0)\ and\ (-30,61.875)\}$

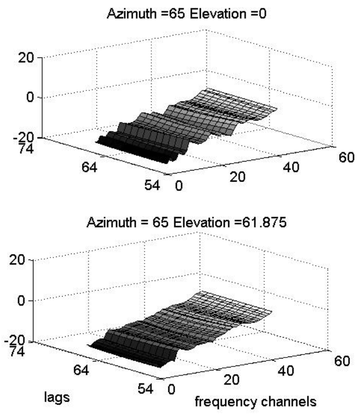

Fig. 7. $IID^{\sigma\tau}$ with $(\theta,\varphi) \in \{(65,0)\ and\ (65,61.875)\}$

The $IID^{\sigma\tau}$ of azimuth 65^o elevation 0^o and 61.875^o are displayed in Fig.7. However, the $3\sigma_i$ = 9, which is much bigger at azimuth 65^o than that at azimuth -30^o. The reason of that lies in the sin(t) changes slowly around -90^o /90^o , which leads to error increasing.

D. Decision-Making Unit

In this section, a Bayes-rule based decision-making approach is employed in this work. A priori azimuth Θ is calculated by GCC+PHAT-$\rho\gamma$. Firstly, and the candidate azimuths θ_i and their probability are obtained. In turn , these candidate azimuths θ_i and their probability serve as a priori for the second step. Based on the information provided by the first layer, the range of the time delay used to be compensated will be obtained. Then the location probability is refined by $IID^{\sigma\tau}$. The process also can be found from **Algorithm 1**, and Fig.8. Mathematically, the decision procedure can be

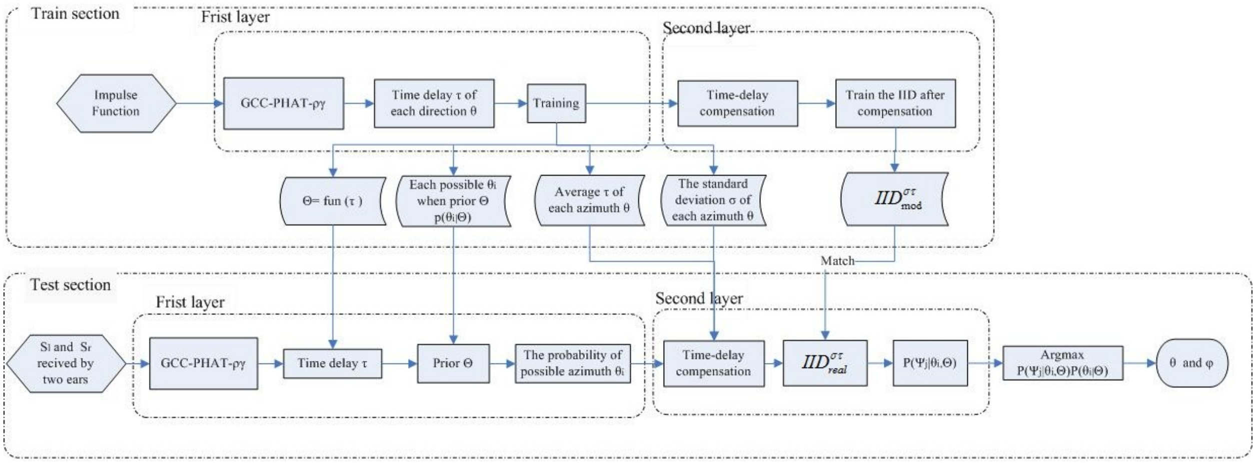

Fig. 8. Flow chart of the algorithm

expressed as:

$$p(\theta,\varphi|O)=\frac{p(\varphi,\theta,O)}{p(O)} = p(\theta|O)p(\varphi|\theta,O) \qquad (24)$$
$$O=(\Theta,IID^{\sigma\tau})$$

where O is the characteristic extracted from original signals received by two ears. Θ denotes the priori azimuth calculated by GCC+PHAT-$\rho\gamma$. While $IID^{\sigma\tau}$ denotes the interaural level difference after time delay compensation between two ear signals in different frequency channels. And $p(\theta,\varphi|O)$ is the probability of the source located at (θ,φ) when characteristic O is observed. $p(\theta|O)$ has been stored after training database.

When a new sound source appears in the environment, the $IID^{\sigma\tau}_{real}$ can be calculated by the algorithm described in section B. In an ideal condition, it is expected to match with a certain template $IID^{\sigma\tau}_{mod}|_{\theta_i,\varphi_j}$, which represents the direction of the sound source. However, with the effect of noise, the relation between $IID^{\sigma\tau}_{real}$ and $IID^{\sigma\tau}_{mod}|_{\theta_i,\varphi_j}$ can be described as

$$IID^{\sigma\tau}_{real}=IID^{\sigma\tau}_{mod}|_{\theta_i,\varphi_j}+\eta'_{\theta_i,\varphi_j} \qquad (25)$$

In this paper, experiment with Gaussian noise is test. So $\eta'_{\theta_i,\varphi_j}$ is Gaussian noise with standard deviation $\sigma_{\eta'_{\theta_i,\varphi_j}}$. When the environment is stable, $\eta'_{\theta_i\varphi_j}$ and $\sigma_{\eta'_{\theta_i,\varphi_j}}$ can be obtained by training data. And they may change with azimuth and elevate changing.

The probability for the direction of θ_i and φ_j can be obtained as follows:

$$p(\varphi_j|\theta_i,O)\sim\frac{1}{\sigma_{\eta'_{\theta_i,\varphi_j}}\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2_{\eta'_{\theta_i,\varphi_j}}}||IID^{\sigma\tau}_{real}-IID^{\sigma\tau}_{mod}|_{\theta_i,\varphi_j}||^2} \qquad (26)$$

The decision rule therefore is

$$(\theta,\varphi)=arg\max_{\theta_i,\varphi_j}p(\varphi_j,\theta_i|O)=arg\max_{\theta_i,\varphi_j}p(\theta_i|O)p(\varphi_j|\theta_i,O) \qquad (27)$$

After finding the max probability of all the possible direction, the result(θ , φ) will be obtained accurately.

Algorithm 1: PSEUDO CODE OF ALGORITHM

```
Input: $x_l(n)$ and $x_r(n)$
Output: azimuth:$\theta$ and elevate:$\psi$
Require: cochlear filter, probability $p(\theta_i|\Theta)$ , $\Delta\overline{\tau}_i$ , $\sigma_i$
    and $IID^{\sigma\tau}_{mod}$ $\eta'_{\theta_i,\varphi_j}$ , $\sigma_{\eta'_{\theta_i,\varphi_j}}$;
if $x_l(n)$ and $x_r(n)$ available then
    Get azimuth $\Theta$ by GCC+PHAT-$\rho\gamma$;
    Return $\theta_i$ and $p(\theta_i|\Theta)$ ;
end
if $\theta_i$ exists then
    find $\Delta\overline{\tau}_i$ , $\sigma_i$ from database;
    for $\Delta\tau\subseteq(-3\sigma_i+\Delta\overline{\tau}_i,3\sigma_i+\Delta\overline{\tau}_i)$;
        get $IID^{\sigma\tau}_{real}$ Match $IID^{\sigma\tau}_{real}$ with $IID^{\sigma\tau}_{mod}$ ;
        return $p(\varphi_j|\theta_i,O)$;
end
$(\theta,\varphi)=arg\max_{\theta_i,\varphi_j}p(\theta_i|O)p(\varphi_j|\theta_i,O)$;
return $(\theta,\varphi)$
```

III. EXPERIMENTS AND DISCUSSIONS

The CIPIC database of head-related impulse responses (HRIRs) [28] is used in our experiments. The database is measured by the U.C.Davis CIPIC Interface Laboratory, which includes head-relate impulse responses for 45 different subjects (including 27 males, 16 females, and KENAR with large and small pinna). And the HRIRs is tested in 1m distance with 25 different azimuths, 50 different elevations, totally 1250 directions for each subject. It is a commonly used database in sound source localization experiments.

In this work, 20 groups of sounds(10 for human voice and 10 for music) is used to train the standard templates. The parameter $p(\theta_i|\Theta)$, σ_i , $\overline{\tau}_i$ can be get after statistics of azimuth results. And $IID^{\sigma\tau}$ is obtained after time delay compensated for each sound. The template $IID^{\sigma\tau}$ will be acquired by EM algorithm for each direction. And 100 groups of real sound signals (50 for human voice and 50 for music) are considered for each direction. The duration of

TABLE I

THE ACCURACY OF θ LOCALIZATION IN TWO DIFFERENT EXPERIMENTAL ENVIRONMENTS

	Environment without noise			Environment with 20DB noise		
	$= 0^o$	$\leq 5^o$	$\leq 10^o$	$= 0^o$	$\leq 5^o$	$\leq 10^o$
Probabilistic Model	92.72%	98.64%	100%	75.94%	83.96%	87.74%
TDC	90.28%	98.48%	99.84%	87.56%	97.56%	98.96%
OnlineCalibration	89.12%	96.76%	99.24%	84.26%	95.92%	98.24%
Hierarchical System	93.90%	98.70%	100%	85.64%	97.21%	98.72%

each signal is 2 seconds. There will be $100 \times 45 \times 25 \times 50 = 5625000$ groups of signals used in the experiment. The sampling frequency is 44100Hz. The result will be compared with the methods of Probabilistic Model [21], Hierarchical System [20] and Online Calibration [22]. The method based on Time Delay Compensation proposed in this paper is short for TDC.

A. Results of experiments

In this paper, the results for test sets are based on different signal parts at $45 \times 1250 \times 100 \times 128$, which means 45 subjects, 1250 directions, 100 sound signals are processed over 128 sample points which is also the length of the window. Two different experimental environments are considered, one for ideal condition without noise, the other for condition with SNR 20 dB (the additive white Gaussian noise environment). The result is obtained with different error tolerance, $(0^o,\ 5^o\ and\ 10^o)$ for θ and $(0^o,\ 5.625^o,\ 11.25^o, and\ 25^o\ )$ for φ.

There are three tables shown as follows. They display the result of θ and φ localization in two different experimental environments. TABLE I shows the accuracy of θ localization in the environment both without noise and 20 DB white Gaussian noise by different algorithms. While TABLE II shows the accuracy of φ localization in the two experimental environments. The space complexity and the time complexity needed in four algorithms are shown in TABLE III and TABLE IV.

B. Analysis of results

As shown in TABLE I, for θ localization, the performances between these four algorithms have small gaps with each other in the experimental environment without noise. All the accuracies of θ localization are over 89% with the error tolerance 0^o, and over 99% with the error tolerance 10^o. It can be found that Hierarchical System has the best performance while Online Calibration has the worst performance, which is probably due to the different cues used in different algorithms. In Probabilistic Model, ITD and IID are processed as a whole, which will reduce the effect from each other; In our method, TDOA is used in first layer; and $IID^{\sigma\tau}$ is used in second layer. ITD and IID are used in two different layers in Online Calibration model; Except for ITD and IID, spectral cues are also used as the third layer in Hierarchical System. Our algorithm is an improvement of IID used in Probabilistic Model. The redundant information is discarded which decreases the computing complexity heavily with a little accuracy dropping.

With the effect of noise, the performances of all these four algorithms drop rapidly for azimuth localization. However, it can also be found that our algorithm is more robust than other algorithms. The reason is that the PHAT-$\rho\gamma$ used in the first layer has been proved more robust. Without an accurate prior information, Probabilistic Model gets a worse result than our method. Comparing with $IID^{\sigma\tau}$ of our algorithm, IID has lager measurement error for Hierarchical System duo to the effect of time delay, in addition, the noise affect spectral cues heavily. Online Calibration has a similar first and second layer with Hierarchical System. The difference is that ITDs are considered in each frequency channel, which has weak improvement on localization.

TABLE II shows that the accuracies of elevation localization is lower than that of azimuth localization. ITD/ TDOA offers little help for elevation localization. Therefore, the accuracies of elevation localization will depend crucially on IID algorithm. Time delay is introduced to improve IID in Probabilistic Model and our method, which is suitable for any kind of sound. Because all possible lags were taken into account in Probabilistic Model, it got the best performance. Spectral cues make Hierarchical System work better than Online Calibration system.

It can also be found in TABLE II that the performance of elevation localization decline dramatically with the effect of noise. There are two reasons for this phenomenon. First, cumulative error has a certain effect on the accuracy of localization, which means the elevation location performance depend on the localization accuracy of azimuth. Second, elevation is so sensitive to the error of IID, that the elevation localization error will be lager than azimuth. In multi-layer structural system, the lower layer can provide candidates for the upper layer, which can reduce the influence of neighbor azimuth on a same elevation. Although three-layer model is used in Hierarchical System, there is an insignificant increase compared to Online Calibration system, which is due to the fact that spectral cue is not steady in noise environment. Our method can get the most robust result in noise environment.

TABLE III shows that the space complexity and the time complexity used for matching with templates of four algorithms. Here $n_a, n_e\ and\ n_c$ denote the number of azimuth, elevation and the channels of the filterbank. In this work, $IID^{\sigma\tau}$ is needed for each direction It is concluded that the algorithm used in this paper can get a better performance with less storage space and computing cost.

TABLE II

THE ACCURACY OF φ LOCALIZATION IN TWO DIFFERENT EXPERIMENTAL ENVIRONMENTS

	Environment without noise				Environment with 20DB noise			
	$= 0^o$	$\leq 5.625^o$	$\leq 11.25^o$	$\leq 22.5^o$	$= 0^o$	$\leq 5.625^o$	$\leq 11.25^o$	$\leq 22.5^o$
Probabilistic Model	73.28%	95.15%	97.67%	98.68%	12.39%	25.28%	48.57%	64.37%
TDC	70.48%	93.13%	96.65%	98.08%	20.38%	45.96%	61.05%	77.07%
OnlineCalibration	63.61%	90.25%	95.82%	97.63%	17.22%	39.49%	56.92%	70.53%
Hierarchical System	64.77%	92.47%	96.23%	97.79%	20.34%	43.62%	60.92%	75.73%

TABLE III

THE SPACE COMPLEXITY AND THE TIME COMPLEXITY NEEDED IN FOUR ALGORITHMS

	storage		*time*	
Probabilistic Model	O($n_a^2 n_e n_c$)	bad	O($n_a n_e$)	bad
TDC	O($n_a n_e n_c$)	good	O(n_e)	good
OnlineCalibration	O($n_a n_e n_c$)	good	O(n_e)	good
Hierarchical System	O($n_a n_e n_c$)	good	O($n_a n_e$)	bad

IV. CONCLUSIONS

In this paper, a time-delay compensation based two-layer probabilistic model for binaural sound localization is presented. At first, a weighting function of GCC named PHAT-$\rho\gamma$ is used in first layer for the candidate azimuth θ_i. With the information of candidate azimuth θ_i, the range of time delay τ needed to be compensated will be obtained. In the second layer, a new Interaural Intensity Difference cue after time-delay compensation named $IID^{\sigma\tau}$ is introduced to refine the information of azimuth θ and elevation φ. Comparation is made with Probabilistic Model(one-layer model), Online Calibration (two-layer model) and Hierarchical System (three-layer model). The experimental results show that the proposed method has higher accuracy and needs less processing time for sound source localization.

REFERENCES

[1] H. Liu and X.F. Li, Time Delay Estimation for Speech Signal Based on FOC-Spectrum, in International Conference on INTERSPEECH, Portland, USA, 2012.

[2] X.F. Li, H. Liu and X.S. Yang, Sound Source Localization for Mobile Robot Based on Time Difference Feature and Space Grid Matching, in IEEE/RSJ International Conference on IROS, San Francisco, California, USA, pp. 2879-2886, 2011.

[3] A. J. King, J. W. Schnupp, and T. P. Doubell, The shape of ears to come: dynamic coding of auditory space, TRENDS in Cognitive Sciences, vol. 5, no. 6, pp. 261-270, 2001.

[4] S. B. Andersson, A. A. Handzel, V. Shah, and P. Krishnaprasad, Robot Phonotaxis with Dynamic Sound-source Localization, in IEEE International Conference on Robotics and Automation (ICRA'04), 2004.

[5] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda, Soundlocalization for humanoid robots - building audio-motor maps based on the HRTF, in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06), Beijing, China, October 2006.

[6] W. J. Strutt, On our perception of sound direction, Philos. Mag., vol. 13, pp. 214-232, 1907.

[7] L. A. Jeffress, A place theory of sound localization, J. Comp. Physiol. Psych., Vol. 61, pp. 468-486, 1948.

[8] R. F. Lyon, A computational model of filtering, detection, and compression in the cochlea, In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'82), pp. 1282-1285, May. 1982.

[9] R. F. Lyon, A computational model of binaural localization and separation, in Proc.IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'83) , pp. 1148-1151,1983.

[10] R. F. Lyon, and C. Mead, An analog electronic cochlea, IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP, vol. 36, pp. 1119-1134, 1988.

[11] Z. M. Fuzessery , Speculations on the role of frequency in sound localization, Brain Behav. Evol., vol. 28, 1986.

[12] J. Blauert, Spatial Hearing (MIT Press), Cambridge, MA, 1983.

[13] Middlebrooks, J. C., and D. M. Green, Sound localization by human listeners, Annu. Rev. Psyc h,vol. 42, pp. 135-159, 1991.

[14] Y. E. Cohen and E. Knudsen, Maps versus clusters: Different representations of auditory space in the midbrain and forebrain, Trends Neurosci., vol. 22, no. 3, pp. 128-135, Mar. 1999.

[15] K. Voutsas, G. Langner, J. Adamy, and M. Ochse, A brain-like neural network for periodicity analysis, IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 35, no. 1, pp. 12-22, Feb. 2004.

[16] N. Roman and D. Wang, Binaural tracking of multiple moving sources, IEEE Transactions on Audio, Speech, and Language Processing(TASL), vol. 16, no. 4, pp. 728-739, May. 2008.

[17] M. Raspaud, H. Viste, G. Evangelista, Binaural Source Localization by Joint Estimation of IID and ITD, IEEE Transactions on Audio, Speech and Language Processing(TASL), vol.18, pp.68-77, Jan. 2010.

[18] C. Lim, R. O. Duda, Estimating the Azimuth and Elevation of a Sound Source from the Output of a Cochlear Model, Conference Record of the Twenty-Eighth Asilomar Conference on Signals, Systems and Computers(ACSSC), Vol. 1, pp.399-403, Oct. 1994.

[19] R. M. Stern, D.Wang, and G. J. Brown, Computational Auditory Scene Analysis. Piscataway, NJ: Wiley/IEEE Press, 2006.

[20] D. Li and S. E. Levinson, A bayes-rule based hierarchical system for binaural sound source localization, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03) , vol. 5, pp. 521-524, Apr. 2003.

[21] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, A probabilistic model for binaural sound localization, IEEE Trans. Syst, Man, Cybern, B, vol. 36, no. 5, pp. 982-994, Oct. 2006.

[22] H. Finger, Paul Ruvolo, Approaches and Databases for Online Calibration of Binaural Sound Localization for Robotic Heads, IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS'10), pp. 4340-4345, Oct. 2010.

[23] M. Jeub, M. Schafer, T. Esch, P. Vary, Model-Based Dereverberation Preserving Binaural Cues, IEEE Transactions on Audio, Speech, and Language Processing (TASL),Vol.18, pp. 1732-1745 Sept 2010.

[24] H Liu, Miao Shen, Continuous Sound Source Localization based on Microphone Array for Mobile Robots IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4332-4339, Oct. 2010.

[25] B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco, Tori of confusion:Binaural localization cues for sources within reach of a listener, Journal of the Acoustical Society of America, vol. 107, no. 3, pp. 1627-636, March 2000.

[26] R. D. Patterson, M. H. Allerhand, and C. Gigure, Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform, J. AcoustS. Soc. Am, pp. 1890-1894, 1995.

[27] B. Glasberg and B. Moore, Derivation of auditory filter shapes from notched noise data, Hear. Res., vol. 47, no. 1/2, pp. 103-138, Aug. 1990.

[28] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, New York, pp. 99-102, Oct. 2001.