2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

A New Hierarchical Binaural Sound Source Localization Method Based on Interaural Matching Filter

Hong Liu[1,2], Jie Zhang[2] and Zhuo Fu[3]

Abstract— **Binaural sound source localization is an important technique in friendly Human-Robot Interaction (HRI) for its easy-implementation with only two microphones. This paper develops a robust method based on a hierarchical probabilistic model. Reliable frequency sub-bands are used to $PHAT-\rho\gamma$ method in the first layer to obtain a priori crude Interaural Time-Delay (ITD) and the probabilistic distribution of candidate azimuths. The second layer utilizes Interaural Intensity Difference (IID) to reduce matching time and refine candidate azimuths as well as elevations. A novel feature named Interaural Matching Filter (IMF), which can eliminate the difference between ITDs and IIDs, is proposed in the third layer. The probability of sound source location is acquired by computing the similarity between the IMF of received binaural signal and the IMFs in templates. Finally, combined with the former ITDs and IIDs, the similarity matrix is used to make a decision of sound source location based on a Bayes rule. Our innovation lies in adding selecting reliable frequency components into time-delay estimation and foremost taking IMF as a feature of sound source. Compared with several state-of-the-art algorithms, experimental results show our approach has a better performance even in noisy environments without increasing storages, and also has less time complexity.**

I. INTRODUCTION

Binaural sound source localization is an important technique in friendly Human-Robot Interaction (HRI) only by two microphones as the human auditory localization with the capability of pinpointing the sound source swiftly and accurately. There are three important and difficult issues concerning binaural localization: First, how to accurately localize any kind of speech or sound source. Second, how to localize several different sound sources at the same time. Third, how to track one or several moving sound sources [1].

There are two significant binaural (interaural) cues based on differences in time and level of the sound arriving at two ears called interaural time difference (ITDs) and interaural intensity differences (IIDs). Since "Duplex Theory" [2] and cochlear model [3] were proposed, a large amount of binaural localization algorithms have been developed in various experimental environments [4][5].

Most traditional methods are based on ITD or IID and seldom consider the influence on each other [5-7]. Intuitively, with the influence of ITD, the signals received by two ears have different starting points with respect to sound source, which affects the extraction of IID. In addition, these methods usually match new obtained features with templates to assure the direction of sound source, that is, higher resolution needs more templates, and more time will be consumed for overall searching to fulfil practical application.

For these reasons, several new algorithms have been proposed recently. For example, Li et al introduced a Bayes-Rule based three-layer hierarchical system for binaural sound source localization [8]. Along with the similar hierarchical architectures like Finger et al [9], experiments show that hierarchical system can reduce time consumption effectively. Willert et al proposed a biologically inspired and technically implemented binaural sound localization system, where binaural cues are extracted using cochleagrams generated by a cochlear model [10]. Meanwhile, noise and reverberation would degrade the performance of sound localization. Jeub et al proposed an interaural coherence model of room impulse response (RIR) and dual-channel Wiener filter for binaural cues dereverberation preserving [11]. Jacob et al presented a multichannel widely linear Wiener filtering approach to binaural noise reduction using a microphone array [12].

Actually, although most algorithms mentioned above may solve a particular problem well, the time or space complexity will be a vital limitation in the usage for real sound localization systems directly. Willert et al presented two-dimensional frequency versus time-delay representations of binaural cues, so-called activity maps, and we have improved this idea and proposed the concept of Time-Delay Compensation for binaural localization [13]. However, there is a common hypothesis that the noises received by two ears are zero-mean Gaussian, which is not always correct, especially for the noises from fans of robot and burst interferences. Besides, Time-Delay Compensation assumes that the disparity of binaural signals is only reflected in time-delay and linear attenuation of energy.

Accordingly, this work foremost proposes IMF as binaural localization feature in order to avoid considering the type or content of noises and satisfy the requirements of complexity. This idea can be applied to eliminate the influence of time-delay on IID without making any assumptions. The three-layer probabilistic model is presented to reduce the matching

This work is supported by National Natural Science Foundation of China (NSFC, No. 61340046, 60875050, 60675025), National High Technology Researchand Development Program of China (863 Program, No.2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (No. JCYJ20120614152234873, CXC201104210010A, JCYJ20130331144631730, JCYJ20130331144716089).

[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] J. Zhang is with the Engineering Lab on Intelligent Perception for Internet of Things(ELIT), Shenzhen Graduate School of Peking University, Shenzhen, 518055 CHINA. *zhang_jie*827@163.*com*

[3] Z. Fu has graduated from the Engineering Lab on Intelligent Perception for Internet of Things(ELIT), Shenzhen Graduate School of Peking University, Shenzhen, 518055 CHINA.

times and improve accuracy.

In details, a modified spectral weighting function of GCC named PHAT-$\rho\gamma$ [14] based on selecting reliable frequency sub-bands is presented to obtain a priori crude ITDs as well as the probabilistic distribution of azimuths in the first layer, because different frequency bands share different time-delays and reliability for computing ITDs.

In the second layer, IIDs are exploited to reduce the searching space of azimuth and elevation for the prior obtained in the first stage. Theoretically, different elevation of the same azimuth should share same IIDs regarding dual-microphone sound localization, but the regularity of head leads to discrimination existing, that is to say, IIDs can be utilized in this stage to refine candidate azimuth and elevation fortunately [7][15].

In the last layer, the newly-developed IMF feature for sound localization is introduced. Take the signal of left (right) ear as the input of a system and the other as the target signal, we can design a filter to eliminate the disparity between binaural signals and this is what Interaural Matching Filter means. Once the coefficients of IMF in all directions are stored, the process of localization is simplified as calculating similarity between unknown coefficients and templates.

Indeed, IMF also compensates one signal to the other ignoring the cues of lags and energy, but considering the disparity comprehensively. For binaural sound localization, the noises received by the two ears are usually of the same type despite of different contents. Therefore, IMF is more reflective of the relationship between ITD and IID than Time-Delay Compensation as well as noisy tolerance.

The rest of this paper is organized as follows: Section II gives an overlook of this paper. Section III shows the details of our method. Experiments and discussions are shown in section IV. At last, the conclusions are drawn in Section V.

II. OVERALL ARCHITECTURE

An overlook of our hierarchical binaural sound source localization method consisting of signal model and algorithm structure is introduced in this section.

A. Localization Modeling

Let $s(n)$ denote a sound source signal, and the received signals as $x_l(n)$ and $x_r(n)$ on the left and right ears, respectively (see Fig.1(a)):

$$\begin{aligned} x_l(n) &= h_l(\theta,\varphi,r) * s(n) + n_l(n) \\ x_r(n) &= h_r(\theta,\varphi,r) * s(n) + n_r(n) \end{aligned} \tag{1}$$

where $h_l(\theta,\varphi,r)$ and $h_r(\theta,\varphi,r)$ are the Head Related Transfer Functions (HRTFs) of the direct paths from source to the two ears, which heavily rely on azimuth, elevation and distance. In Eq.(1), $n_l(n)$ and $n_r(n)$ are the interferences, and θ,φ,r are the azimuth, elevation and distance, respectively.

As with HRI systems, azimuth and elevation represent the direction of sound source and the distance is neglected. Besides in CIPIC database of head-related impulse responses (HRIRs), all sounds used share the same distance of $1m$ [16]. Therefore, the task of this work is to obtain the azimuth θ and elevation φ from $x_l(n)$ and $x_r(n)$, which are measured in a head-centered interaural-polar coordinate system (see Fig.1(b)). Then Eq.(1) can be simplified as follows:

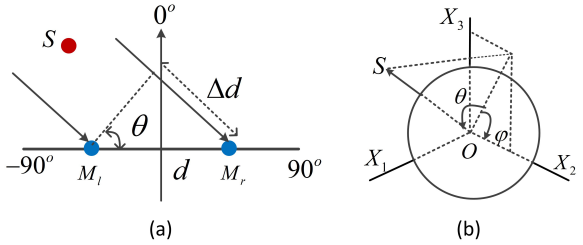

Fig. 1. (a) Signal model of binaural sound localization. For far field, the propagation path of sound source to two eras are thought to be parallel. (b) The interaural-polar coordinate system. The azimuth is the angle between a vector to the sound source and the midsaggital or vertical median plane, and varies from -90^o to $+90^o$. The elevation is the angle from the horizontal plane to the projection of the source into the midsaggital plane, and varies from -90^o to $+270^o$.

$$\begin{aligned} x_l(n) &= h_l(\theta,\varphi) * s(n) + n_l(n) \\ x_r(n) &= h_r(\theta,\varphi) * s(n) + n_r(n) \end{aligned} \tag{2}$$

In the frequency domain, it can be easily obtained:

$$\begin{aligned} X_l(\omega) &= H_l(\theta,\varphi,\omega)S(\omega) + N_l(\omega) \\ X_r(\omega) &= H_r(\theta,\varphi,\omega)S(\omega) + N_r(\omega) \end{aligned} \tag{3}$$

B. Framework

The hierarchical binaural sound localization method consists of four consecutive parts and the framework is shown in Fig.2 (ξ is the location of source):

- *Interaural Time Difference*: This is the first layer, which mainly computes a priori crude ITDs of input signal $x_l(n)$ and $x_r(n)$. The offline training provides the probabilistic distribution of azimuth $p(\theta_i)$ for localization. Moreover, selecting reliable frequency sub-bands is applied to time-delay estimation. The output of this layer is candidate azimuths.
- *Interaural Intensity Difference*: This part exploits IIDs to reduce the searching space of azimuth and elevation based on the prior obtained in the first layer, which is a refine unit to achieve a more accurate probabilistic distribution of azimuth θ and elevation φ. The input and output are similar to these of former layer.
- *Interaural Matching Filter*: This unit designs a Interaural Matching Filter (IMF). Taking the signal of left (right) ear as the input of IMF and the other as the target signal, with the Minimum Mean Square Error (MMSE) criterion we can obtain the impulse response of IMF. The input of this part is impulse response of all directions and the output is the cosine similarity, that is, the probability of sound source located.
- *Decision − making*: A Bayes rule method is used to make the final decision in this stage and the output is the results of localization.

III. HIERARCHICAL BINAURAL SOUND LOCALIZATION

A. Interaural Time Difference

The average time-delays are illustrated in Fig.3. It can be easily obtained that different elevations with the same

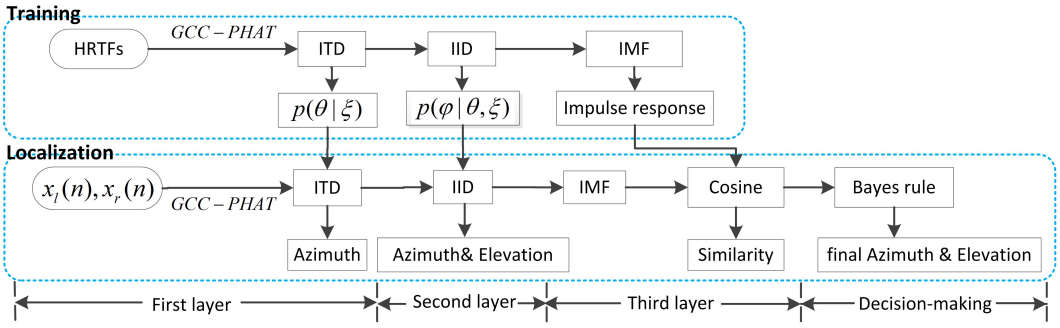

Fig. 2. Block diagram of this framework. The upper training is an offline process providing templates for localization.

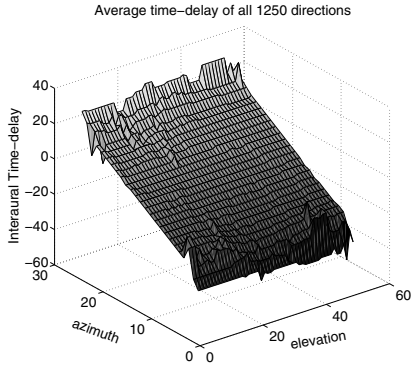

Fig. 3. The average time-delay of all 1250 directions (25 azimuths × 50 elevations) in CIPIC database. The received signals of two ears for computation are the convolution between HRTFs and a certain sound.

azimuth share the same time-delay or lags. In other words, when azimuth is being calculated, the impact of elevation can be ignored. Thereby, in this stage x_l and x_r can be expressed as:

$$\begin{aligned} x_l(n) &= h_l(\theta) * s(n) + n_l(n) \\ x_r(n) &= h_r(\theta) * s(n) + n_r(n) \end{aligned} \tag{4}$$

Assume that binaural signals are counterparts of sound source with time-delay and attenuation, it can be attained:

$$\begin{aligned} x_l(n) &= a_l s(n-\tau_l) + n_l(n) \\ x_r(n) &= a_r s(n-\tau_r) + n_r(n) \end{aligned} \tag{5}$$

where a_l and a_r denote the attenuation factors, τ_l and τ_r are time factors from the sound source to the two acoustic sensors, respectively. Define interaural time-delay $\Delta\tau$ as:

$$\Delta\tau = \tau_r - \tau_l \tag{6}$$

With respect to calculating time-delay $\Delta\tau$, the most widely used method is Generalized Cross Correlation (GCC) [17]. Its improved version, a weighting function of GCC named $PHAT-\rho\gamma$, is used in this paper. The cross-correlation between the signals received by the two ears can be represented as follows:

$$\begin{aligned} R_{x_l,x_r}(n) &= \int_{-\pi}^{\pi} W(\omega) X_l(\omega) X_r^*(\omega) e^{-j\omega n} d\omega \\ W(\omega) &= \frac{1}{\mid G(\omega) \mid^{\rho} + \mid \gamma^2(\omega) \mid} \\ G(\omega) &= X_l(\omega) X_r^*(\omega) \end{aligned} \tag{7}$$

where $W(\omega)$ denotes the weighting function used to sharpen the peak of GCC function, and $G(\omega)$ is the cross power spectrum. In Eq.(7), ρ is reverberation factor determined by signal-to-noise ratio (SNR) in environment and γ is a coherence function. Then the time-delay can be obtained by maximizing cross-correlation function $R_{x_l,x_r}(n)$,

$$\Delta\tau = arg\max_n R_{x_l,x_r}(n) \tag{8}$$

In fact, GCC-PHAT does not consider the reliable frequency channels of a certain sound source, but calculates the time-delay over time instead. However, the signals received by two ears are not always narrow-band and there is a fact that low-frequency sounds travel more easily around the head and the time differences or lags are not affected by φ.

Therefore, this paper proposes a method for selecting reliable frequency channels to improve interaural time-delay estimation. The time-delays of elevation $\varphi = -45^o$ in all azimuths and frequency sub-bands are illustrated in Fig.4, from which it can be seen that there are only several lower frequency channels available to compute time-delay.

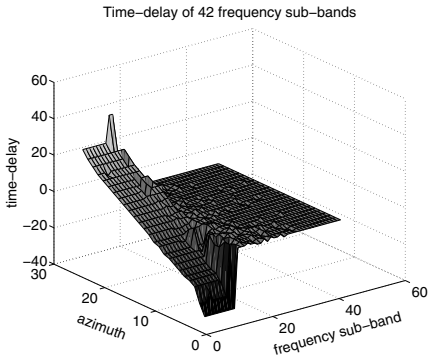

Fig. 4. The time-delay of frequency sub-band in direction of elevation $\varphi = -45^o$. The original signal is decomposed into 42 frequency channels.

As a consequence, a cochlea model is used to decompose the original signals received by two ears into K frequency channels with respective frequency centered by f_c and bandwidth b_c. The filterbank achievement is based on equivalent rectangular bandwidth (ERB)-filters, where f_c and b_c can be computed by using the Glasberg and Moore parameters [17]. Therefore, Eq.(8) can be rewritten as:

$$\Delta\tau_m = arg\max_n R^m_{x_l,x_r}(n), \quad m = 1,2,\cdots,K \tag{9}$$

where $\Delta\tau_m$ represents the time-delay of m frequency channel. Then the final time-delay of a certain direction can be calculated by weighted averaging as:

$$\Delta\tau = \frac{1}{k}\sum_{m=1}^{K} b_m \Delta\tau_m \tag{10}$$

$$b_m = \begin{cases} 0 & \text{if channel } m \text{ is unreliable,} \\ 1 & \text{if channel } m \text{ is reliable.} \end{cases} \tag{11}$$

where k is the number of selected frequency channels used to evaluate time-delay, and b_m is a binary mask as the sign of whether the frequency channel m is reliable and selected.

Then, considering the geometrical relationship between the time-delay and azimuth (see Fig.1(a)), the azimuth θ is easily evaluated from τ as Eq.(12) shows:

$$\Theta = \sin^{-1}\left(\frac{\Delta d}{d}\right) = \sin^{-1}\left(\frac{\Delta\tau c}{d f_s}\right) \tag{12}$$

where c is the speed of sound in the air (344 m/s), Δd is the distance difference from the sound source to two ears, d is the distance between the two ears, and f_s is the sampling frequency. As with CIPIC database, the azimuth is divided into 25 intervals centered by -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 noises in the environment and the possibility of error in calculation, azimuth localization inaccuracy may happen [13].

Since each time-delay corresponds to an only azimuth θ_i, so the probability of azimuth θ_i by $P(\theta_i|\Theta)$ can be trained and stored before localization. Therefore, when the candidate azimuth θ_i is obtained in this layer, all possible τ are tested for each φ. Then the mean value $\overline{\tau_i}$ and standard deviation σ_i are obtained. The probability of azimuth θ_i and available interval are shown as:

$$\begin{aligned} P(\theta_i|\Theta) &= P(\tau_i|\Delta\tau) \sim N(\overline{\tau_i}, \sigma_i^2) \\ \Delta\tau &\subseteq (-3\sigma_i + \overline{\tau_i}, 3\sigma_i + \overline{\tau_i}) \quad when\ \theta = \theta_i \end{aligned} \tag{13}$$

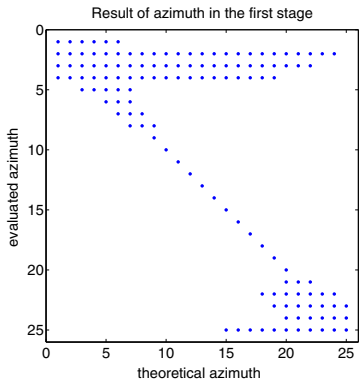

Fig. 5. The evaluated result of azimuth in the first stage. The solid dots represent the possible azimuths, for example, if the actual azimuth Θ is -80^o, the evaluated azimuths θ are $-80^o, -65^o, -55^o, -45^o, -40^o, -35^o$ with different possibilities, and the possibility at $\theta = -80^o$ is maximum..

Accordingly, a crude result in this stage can be obtained and shown in Fig.5, which is the so-called candidate azimuths for the following operations.

B. Interaural Intensity Difference

Under the ideal condition and ignoring the background and reverberation noises, the energy spectrums of the received signals at two ears are:

$$\begin{aligned} E_l(\omega) &= X_l(\omega)^2 = S(\omega)|H_l(\omega)|^2 \\ E_r(\omega) &= X_r(\omega)^2 = S(\omega)|H_r(\omega)|^2 \end{aligned} \tag{14}$$

From an engineering point of view and in intensity domain, the relation between the sound source and HRTFs in Eq.(14) will be simplified as addition [8]. Hereby the logarithmic energy intensities are formulated as:

$$\begin{aligned} I_l(\omega) &= 10\log E_l(\omega) = 10\log S(\omega) + 20\log|H_l(\omega)| \\ I_r(\omega) &= 10\log E_r(\omega) = 10\log S(\omega) + 20\log|H_r(\omega)| \end{aligned} \tag{15}$$

Then, interaural intensity difference can be defined as:

$$\begin{aligned} \Delta I(\omega) &= I_l(\omega) - I_r(\omega) \\ &= 20\log|H_l(\omega)| - 20\log|H_r(\omega)| \\ &= 20\log\frac{|H_l(\omega)|}{|H_r(\omega)|} \end{aligned} \tag{16}$$

Hence, sound source is irrelevant to IIDs, which only relys on HRTFs instead. For a certain subject, the distribution of IIDs is stationary that can be trained before localization.

As for dual-microphone localization, according to the so-called inverse-square-law [7], the IIDs vary from radius. But reflection, diffraction and the regularity of head make IIDs vary from azimuths and elevations, which means IIDs can be utilized to refine the candidate azimuth and elevation. The IIDs of all 1250 directions are illustrated in Fig.6 and they are stored in templates. It can be seen that different directions share different IIDs in the gross.

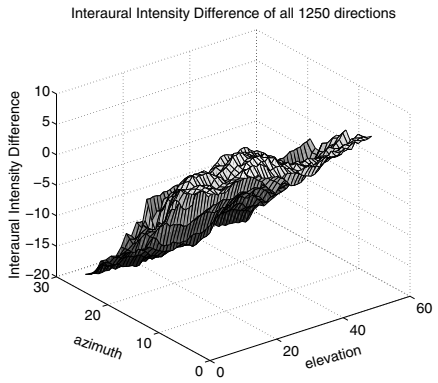

Fig. 6. Interaural Intensity Difference of all 1250 directions. Here IIDs denote the feature of original signals without frequency division.

Here based on the result of first layer, candidate elevations can be assured by matching the IID of undefined location with IIDs within templates using Bayes rule expressed as:

$$P(\varphi|\theta_i) = \frac{P(\theta_i, \varphi)}{P(\theta_i)} \tag{17}$$

where $p(\theta_i)$ is obtained in the previous layer. Similarly, let $\overline{iid_j}$ and δ_j denote the mean value and standard deviation respectively when the candidate azimuth is θ_i. Then the

probability of elevation φ and available interval of *iid* are shown as follow:

$$\begin{aligned} P(\theta_i,\varphi) &= P(\tau_i, iid) \sim N(\overline{iid}_j, \delta_j^2) \\ iid &\subseteq (-3\delta_j + \overline{iid}_j, 3\delta_j + \overline{iid}_j) \ when \ (\theta,\varphi) = (\theta_i,\varphi_j) \end{aligned} \tag{18}$$

C. Interaural Matching Filter

In order to eliminate the disparity between binaural signal $x_l(n), x_r(n)$, it is easy to compose an Interaural Matching Filter (IMF) [19][20] shown in Fig.7, whose task is to press $y(n)$ on towards $x_r(n)$:

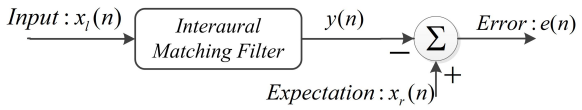

Fig. 7. Linear discrete time Interaural Matching Filter, which includes delay and attenuation. Taking $x_l(n)$ as the input of IMF and $x_r(n)$ as the expectation is equivalent to the contrast situation in theory.

Let $\boldsymbol{w} = [w_0, w_1, \ldots, w_{M-1}]$ be the impulse response of IMF, and the frame length of $x_l(n), x_r(n)$ is M, the output of IMF is expressed as:

$$y(n) = \sum_{i=0}^{M-1} w_i^* x_l(n-i), \quad n = 0,1,\ldots,M \tag{19}$$

where $*$ denotes conjugate. Since the estimation of expectation response $x_r(n)$ always has error, thus define the error function as:

$$e(n) = x_r(n) - y(n) \tag{20}$$

At the same time, the cost function is defined as follow:

$$J(n) = E\{|e(n)|^2\} = E\{e(n)e^*(n)\} \tag{21}$$

where E is the expectation operator. Here considering the MMSE criterion to solve the vector $\boldsymbol{w}$, we can obtain the Wiener-Hopf equation:

$$\sum_{i=0}^{\infty} w_i R_{x_l,x_l}(i-k) = R_{x_l,x_r}(-k), \quad k = 0,1,\ldots,M-1 \tag{22}$$

where R_{x_l,x_l} is the autocorrelation of x_l and R_{x_l,x_r} is the cross-correlation function calculated in the first layer already. If the signal received by left ear is set as:

$$\boldsymbol{x}_l(n) = [x_l(n), x_l(n-1), \ldots, x_l(n-M+1)]^T \tag{23}$$

Then the autocorrelation matrix of $\mathbf{x}_l(n)$ is:

$$\begin{aligned} \boldsymbol{R} &= \{\boldsymbol{x}_l(n)\boldsymbol{x}_l^H(n)\} \\ &= \begin{bmatrix} R_{x_l,x_l}(0) & R_{x_l,x_l}(1) & \ldots & R_{x_l,x_l}(M-1) \\ R_{x_l,x_l}^*(1) & R_{x_l,x_l}(0) & \ldots & R_{x_l,x_l}(M-2) \\ \vdots & \vdots & \vdots & \vdots \\ R_{x_l,x_l}^*(M-1) & R_{x_l,x_l}^*(M-2) & \ldots & R_{x_l,x_l}(0) \end{bmatrix} \end{aligned} \tag{24}$$

Similarly, the cross-correlation vector between input $\boldsymbol{x}_l$ and expectation response $x_r(n)$ is calculated as:

$$\begin{aligned} \boldsymbol{r} &= E\{\boldsymbol{x}_l(n)x_r^*(n)\} \\ &= [R_{x_l,x_r}(0), R_{x_l,x_r}(-1), \ldots, R_{x_l,x_r}(-M+1)] \end{aligned} \tag{25}$$

Therefore, the vector of IMF coefficients can be formulated as:

$$\boldsymbol{w} = \boldsymbol{R}^{-1}\boldsymbol{r} \tag{26}$$

So far, the impulse response $\boldsymbol{w}$ in all directions have been obtained, which includes the information of ITDs and IIDs, and our next goal is to use it for the final localization. Here use cosine similarity between two IMFs to describe the probability of sound source location like:

$$\beta_{\boldsymbol{w}_1\boldsymbol{w}_2} = \frac{<\boldsymbol{w}_1, \boldsymbol{w}_2>}{\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\|} \tag{27}$$

where $<,>, \|\cdot\|$ denote the inner product of vectors and 2nd order norm. If $\boldsymbol{w}_1$ is the coefficients from templates and $\boldsymbol{w}_2$ is the coefficients of received sound, $\beta_{\boldsymbol{w}_1\boldsymbol{w}_2}$ can bewrite the similarity between the two. The mean value and standard deviation of estimation error $e(n)$ in all 1250 directions can be quantized and shown in Fig.8.

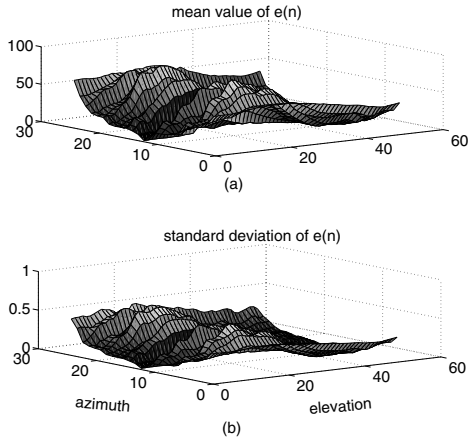

Fig. 8. The mean value and standard deviation of estimation error of IMF (The two sub-figures (a) and (b) depict the mean value and standard deviation, respectively.).

A similarity matrix $\beta_{\boldsymbol{w}_1\boldsymbol{w}_2}$ is illustrated in Fig.9, when the sound source is located in the direction of $\theta = -45^o, \varphi = 5.625^o$. It can be seen the coordination of the brightest dot is $(10,5)$ enclosed by a circle, where the cosine similarity is largest.

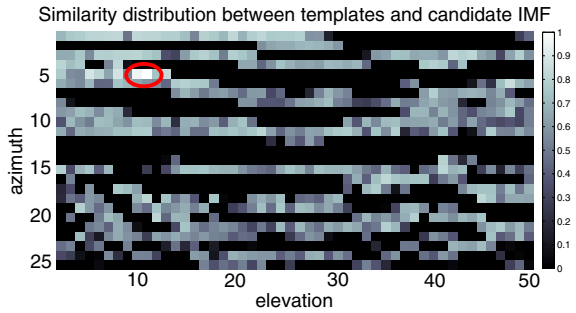

Fig. 9. The similarity matrix when the sound source is located in the direction of $\theta = -45^o, \varphi = 5.625^o$. The more deeper of the pixel, the bigger the similarity is. It can be seen the coordination of the brightest dot is $(10,5)$ enclosed by a circle. The dots of similarity less than zero are set black.

Therefore, the cosine similarity can be used to represent the probability of sound source location. Based on the results of the previous two layers, what we need to do is to calculate

the similarity between the IMFs of candidate directions in templates with the IMF of received signals as the following equation shows:

$$P(\beta_{ij}|\theta_i,\varphi_j) = \frac{P(\theta_i,\varphi_j,\beta_{ij})}{P(\theta_i,\varphi_j)} = arg\max_{\beta_{ij}} \beta_{\boldsymbol{w}_{ij}\boldsymbol{w}_{template}}|_{(\theta_i,\varphi_j)} \tag{28}$$

D. Decision-Making

Once templates including ITDs, IIDs and IMFs are stored, the process of localization is simplified as matching candidate features with templates and Algorithm 1 illustrates the procedures of localization. Here, the Bayes rule is used to make the final decision expressed mathematically as (Let ξ denotes the sound source location):

$$\begin{aligned}\xi &= arg\max_{\xi} P(\xi|\theta,\varphi,\beta) \\ &= arg\max_{\xi} P(\theta|\xi)P(\varphi|\theta,\xi)P(\beta|\theta,\varphi,\xi)P(\xi)\end{aligned} \tag{29}$$

where $P(\theta|\xi), P(\varphi|\theta,\xi), P(\beta|\theta,\varphi,\xi)$ are obtained in the previous layers, and $P(\xi)$ is the prior information same to all locations.

Algorithm 1: Hierarchical Binaural Sound Localization

Input: $x_l(n)$, $x_r(n)$
Output: azimuth θ, elevation φ
Templates: $\overline{\tau_i}$, σ_i, $\overline{iid_j}$, δ_j, $IMFs$;
while $x_l(n)$, $x_r(n)$ *available* **do**
  decompose $x_l(n), x_r(n)$ into K frequency bands;
  for $m \leftarrow 1$ *to* K **do**
    $\Delta\tau_m \leftarrow arg\max_n R^m_{x_l,x_r}(n)$;
    if *frequency band m is reliable* **then**
      $b_m \leftarrow 1$;
    end
  end
  $\Delta\tau \leftarrow \frac{1}{k}\sum_{m=1}^{K} b_m \Delta\tau_m$;
  if $\Delta\tau \subseteq (-3\sigma_i + \overline{\tau_i}, 3\sigma_i + \overline{\tau_i})$ **then**
    $\theta_i \leftarrow arc\sin(\frac{\overline{\tau_i} c}{d f_s})$;
    candidate azimuths $\leftarrow \theta_i$;
    $P(\theta_i) \leftarrow N(\overline{\tau_i}, \sigma_i^2)|_{\Delta\tau}$;
  end
  while θ_i *exists* **do**
    $iid \leftarrow I_l(\omega) - I_r(\omega)$;
    if $iid \subseteq (-3\delta_j + \overline{iid_j}, 3\delta_j + \overline{iid_j})$ **then**
      transform $\overline{iid_j}$ into elevation φ_j;
      candidate elevations $\leftarrow \varphi_j$;
      $P(\varphi_j|\theta_i) \leftarrow N(\overline{iid_j}, \delta_j^2)|_{(\Delta\tau, iid)}$;
    end
    while (θ_i, φ_j) *exist* **do**
      $\boldsymbol{w}_{ij} \leftarrow \boldsymbol{R}^{-1}_{x_l,x_l} \boldsymbol{r}_{x_l,x_r}$;
      $\beta_{ij} \leftarrow \frac{<\boldsymbol{w}_{ij}, \boldsymbol{IMF}_{template}>}{\|\boldsymbol{w}_{ij}\|\|\boldsymbol{IMF}_{template}\|}$;
    end
  end
  $(\theta,\varphi) \leftarrow arg\max_{(\theta,\varphi)} P(\theta)P(\varphi|\theta)P(TDC|\theta,\varphi)$;
  return (θ,φ)
end

IV. Experiments and Discussions

The CIPIC database used in our experiments is measured by the U.C.Davis CIPIC Interface Laboratory, which includes head-related impulse responses (HRIRs) for 45 different subjects (including 27 males, 16 females, and KENAR with large and small pinna). The HRIRs are tested at 1m distance with 25 different azimuths, 50 different elevations, totally 1250 directions for each subject. The sound source is musical signal. The period of each sound for training and localization is 2 seconds and the sampling frequency is 44.1kHz. The result will be compared with Hierarchical System [8], Online Calibration [9] and Time-Delay Compensation (TDC) [13], and this method is IMF for short.

In this paper, the results for test sets are based on different signal parts at $45 \times 1250 \times 100 \times 128 \times 5$, which means 45 subjects, 1250 directions, 100 sound signals and 5 sound activities processed over 128 sample points, which is also the length of window. This algorithm of IMF is validated in different SNRs and with several different sound activities.

A. Azimuth

The accuracy of azimuth θ is shown in TABLE I. It can be seen that in noisy environment our method reaches the highest performance among the four methods clearly.

Specific say, the performances among these four algorithms have small gaps with each other in the experimental environment without noise. All the accuracies are over 89% with the error tolerance 0^o, and over 99% with the error tolerance 10^o, that is, these four algorithms have satisfied the real requirement in quite environments. 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, for example Hierarchical System utilizes the spectral differential cues.

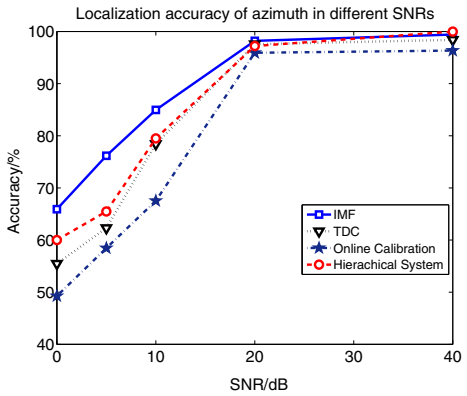

Fig. 10. Localization accuracy of azimuth in different SNRs with 5^o tolerance.

A comparison of the four algorithms with 5^o tolerance in different SNRs is illustrated in Fig.10. It can be acquired our algorithm has a considerable superiority to the other three in noisy environments. With the effect of noise, our algorithm can extract more accurate time-delay because selecting reliable frequency sub-bands is used for $PHAT-\rho\gamma$, which can reject unreliable frequency components to obtain a more robust time-delay.

TABLE I

THE ACCURACY OF θ IN DIFFERENT SNRS

SNR	Environment without noise			20dB			10dB		
Tolerance	$=0^o$	$\leq 5^o$	$\leq 10^o$	$=0^o$	$\leq 5^o$	$\leq 10^o$	$=0^o$	$\leq 5^o$	$\leq 10^o$
ICTDC	93.16%	98.52%	99.56%	88.32%	98.20%	99.13%	70.24%	84.96%	92.72%
TDC	90.28%	98.48%	99.84%	87.56%	97.56%	98.96%	62.64%	78.43%	83.04%
Online Calibration	89.12%	96.76%	99.24%	84.26%	95.92%	98.24%	58.94%	67.52%	75.23%
Hierarchical System	93.90%	98.70%	99.87%	85.64%	97.21%	98.72%	63.64%	79.50%	84.13%

TABLE II

THE ACCURACY OF φ IN DIFFERENT SNRS

SNR	Environment without noise			20dB			10dB		
Tolerance	$=0^o$	$\leq 5.625^o$	$\leq 11.25^o$	$=0^o$	$\leq 5.625^o$	$\leq 11.25^o$	$=0^o$	$\leq 5.625^o$	$\leq 11.25^o$
ICTDC	83.48%	91.32%	94.80%	52.56%	71.44%	78.08%	28.88%	46.96%	55.44%
TDC	70.48%	92.13%	94.65%	20.38%	45.96%	61.05%	10.56%	26.21%	35.98%
Online Calibration	63.61%	90.25%	95.82%	17.22%	39.49%	56.92%	9.53%	20.84%	29.16%
Hierarchical System	64.77%	92.47%	95.23%	20.34%	43.62%	60.92%	10.73%	25.46%	32.10%

B. Elevation

In our experiments, the localization space is divided into 50 elevations ranging from -45^o to $+230.625^o$ in steps of 5.625^o. The localization accuracy of elevation φ is shown in TABLE II. Obviously, *IMF* has acquired the best result compared to the other three algorithms.

In the environments without noise, though the accuracy of *IMF* is highest, the others do not lag far behind. If the tolerance is more than 5.625^o, all methods can achieve more than 90% satisfying the fundamental system roughly. However, in noisy environments the accuracy of other three algorithms drops rapidly and our method has an obvious advantage. While with no tolerance, the accuracy of *IMF* has even achieved more than two times of other three's.

The main reason lies in that ITD offers little help for elevation localization, which depends crucially on I-ID algorithm. Among these four algorithms, *IMF*, *TDC*, and *Hierarchical System* have considered the influence of IID, so they have obtained a more better result than *Online Calibration*. *TDC* has considered the relation between ITD and IID, which makes it more robust than *Hierarchical System*. Compared with *TDC*, *IMF* has not assumed that the disparity of binaural signals is only reflected in time-delay and linear attenuation of energy, but considering overall, which makes it have more universality.

The performance of the four algorithms in different SNRs with 5.625^o tolerance is illustrated in Fig.11. It can be seen that in different noisy environments, *TDC* and *Hierarchical System* have little disparity in accuracy, but *IMF* has achieved a visible localization accuracy, because apart from the layer of *Interaural Intensity Difference*, *IMF* can be effectively extracted for localization whatever the type of noises, which is important to the estimation of elevation. Accordingly, as localization feature IMF is more robust than spectral cues proposed in *Hierarchical System*.

C. Sound Activity Localization

In order to verify the effectiveness and universality for real sound localization system, five different sound activities are expected to evaluate the IMF method in this paper. The five

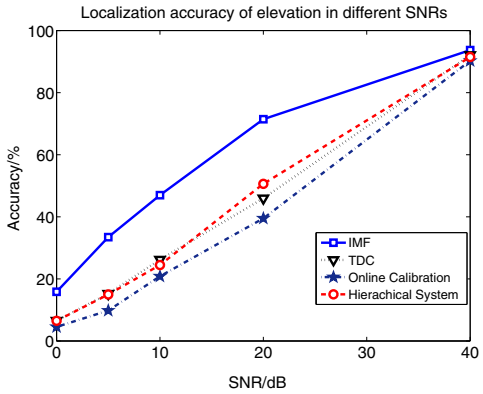

Fig. 11. Localization accuracy of elevation in different SNRs with 5.625^o tolerance.

TABLE III

THE LOCALIZATION ACCURACY OF DIFFERENT SOUND ACTIVITIES

direction	Azimuth		Elevation	
Tolerance	$=0^o$	$\leq 10^o$	$=0^o$	$\leq 11.25^o$
clapping	86.24%	94.24%	80.81%	93.24%
knocking	93.12%	97.44%	84.88%	94.16%
telephone	90.72%	98.40%	79.44%	96.08%
screaming	88.04%	97.44%	73.36%	94.88%
smashing	85.60%	96.08%	66.32%	89.92%

sound activities are all very common in life, including clapping hands, knocking on a door, telephone ringing, screaming and glass smashing, which are recorded in quite official environment. Table III shows the localization accuracy. It can be seen that these activities are well localized for both azimuth and elevation, for example, when the tolerance is 0^o, the accuracy of azimuth can achieve more over 85.6%.

However, it's worth noting that the accuracies of glass smashing are slightly lower than the others'. This phenomenon greatly depends on the sounding principle, because glass smashing has blank sounding time slot and the energy mainly distributes in high frequency channels, which make selecting reliable frequency channels more difficult, but the localization results are fully suitable for HRI systems.

D. Complexity Analysis

Let N_a, N_e and N_c denote the number of azimuth, elevation and the channels of filterbank (approximately equivalent to

TABLE IV

THE SPACE COMPLEXITY OF THE FOUR ALGORITHMS

space	*storage*	*order*
IMF	$N_aN_eN_c+2N_aN_e$	O($N_aN_eN_c$)
TDC	$N_aN_eN_c+N_aN_e$	O($N_aN_eN_c$)
Online Calibration	$N_aN_eN_c+N_aN_e$	O($N_aN_eN_c$)
Hierarchical System	$N_aN_eN_c+2N_aN_e$	O($N_aN_eN_c$)

TABLE V

THE TIME COMPLEXITY OF THE FOUR ALGORITHMS

time	*times of comparison*	*order*
IMF	$N_a+N_a(N_e+N_c)$	O(N_aN_e)
TDC	$N_a+N_aN_eN_c$	O($N_aN_eN_c$)
Online Calibration	$N_a+N_aN_e+N_aN_eN_c$	O($N_aN_eN_c$)
Hierarchical System	$N_a+N_a(N_e+N_c)$	O(N_aN_e)

the order of filter), respectively. Considering the algorithms mentioned above definitely concluding the process of training, and the templates of ITD, IID should be stored before localization.

The space complexity of the four algorithms is shown in TABLE IV. It can be observed that the storage of both *IMF* and *Hierarchical System* is N_aN_e bigger than *TDC* or *Online Calibration*, for *IMF* needs to store the information of IIDs in all directions, so does *Hierarchical System*. But all the orders of storage of the four algorithms are $O(N_aN_eN_c)$, which means the space complexity is acceptable.

When taking the comparison as the basic operation, the time complexity of the four algorithms are shown as TABLE V. It is clearly to be seen that *IMF* and *Hierarchical System* have achieved the lowest time complexity than the other two algorithms, because these two methods have the similar matching mechanism, with which the previous layer provides candidates for the following layers. Compared with *IDC*, *IMF* adds a layer of IID, which can effectively reduce the matching times of elevation. That's why *IMF* and *Hierarchical System* achieve the best time complexity.

V. CONCLUSIONS

In this paper, a new Hierarchical binaural sound localization method based on Interaural Matching Filter has been proposed. Our algorithm utilizes ITD, IID and IMF as the features of sound source. As for time-delay estimation, binaural signals possess prodigious sparsity and selecting reliable frequency components is necessary to improve accuracy. Compared with ITD, IID has little regularity with directions, which means it can be used to reduce searching space and it's successfully used in this work. The newly-proposed IMF has not made any assumptions of the type of noise and relationship between binaural signals reflected in simple time-delay and attenuation, and experiments show that IMF is more robust than traditional features. At last, we would like to emphasize that our method has achieved a favourable localization accuracy for both azimuth and elevation, especially in the noisy environments, and needs less time complexity without requiring more storages for templates. So to speak, we have provided a more appropriate choice for real HRI systems. Our future work may be fastened on studying the factors influencing the accuracy of IMF.

REFERENCES

[1] 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.

[2] L. A. Jeffress, "A place theory of sound localization", Journal of comparative and physiological psychology, vol.61, pp.468-486, 1948.

[3] 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.

[4] H. Liu and X. F. Li, "Time Delay Estimation for Speech Signal Based on FOC-Spectrum", in Proceedings of International Conference on INTERSPEECH, Portland, USA, pp.1732-1735, 2012.

[5] 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 Proceedings of IEEE/RSJ International Conference on Robotics and Systems (IROS), San Francisco, California, USA, pp.2879-2886, 2011.

[6] M. D. Gillette and H. F. Silverman, "A linear closed-form algorithm for source localization from time-differences of arrival", IEEE Signal Processing Letters, vol.15, pp.1-4, 2008.

[7] W. Cui, Z. Cao and J. Wei, "Dual-microphone source location method in 2-D space", in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06), vol.4, pp.845-848, May. 2006.

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

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

[10] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, "A probabilistic model for binaural sound localization", IEEE Transaction on Systems, Man, and Cybernetics, Part B: vol.36, no.5, pp.982-994, Oct. 2006.

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

[12] J. Benesty and J. D. Chen, "A Multichannel Widely Linear Approach to Binaural Noise Reduction Using an Array of Microphones", in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.313-316, 2012.

[13] H. Liu, Z. Fu and X. F. Li, "A Two-Layer Probabilistic Model Based on Time-Delay Compensation Binaural Sound Localization", in Proceedings of IEEE International Conference on Robotics and Automation (ICRA), pp.2690-2697, May. 2013.

[14] H. Liu and M. Shen, "Continuous Sound Source Localization based on Microphone Array for Mobile Robots", in Proceedings of IEEE/RSJ International Conference on Intelligent Robotics and Systems (IROS), pp. 4332-4339, Oct. 2010.

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

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

[17] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay", IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP), vol.24(4), pp.320-327, 1976.

[18] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched noise data", Hearing research, vol.47(1), pp.103-138, 1990.

[19] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications", Neural networks, vol.13(4), pp.411-430, 2000.

[20] S. S. Haykin, "Adaptive Filter Theory", 4/e[M], Pearson Education India, 2005.