2015 IEEE International Conference on Robotics and Automation (ICRA)
Washington State Convention Center
Seattle, Washington, May 26-30, 2015

Binaural Sound Source Localization Based on Generalized Parametric Model and Two-Layer Matching Strategy in Complex Environments

Hong Liu, Cheng Pang and Jie Zhang

Abstract— **Binaural sound source localization is an important technique involving Human-Robot Interaction (HRI), video conference, speech enhancement, etc. In many real application scenarios, especially for closed environments, the affect of reverberation and noise would degrade the precision of position estimations. Therefore, a new binaural sound source localization method based on generalized parametric model and two-layer matching strategy is proposed in this paper for complex environments. Firstly, cepstral prefiltering is utilized for dereverberation of binaural signals. Then, two binaural cues computed from a dual-channel frequency representation, are combined to estimate the azimuths of sources. Additionally, the generalized parametric model is presented to describe the relationship between the azimuth and binaural cues through finding the optimal scaling factors from training data. At last, a two-layer matching strategy based on Bayesian rule is used to make the final decision, which can effectively decrease the computation complexity. Experiments have validated the proposed approach and show that it achieves favorably better results compared with several available methods without extra spacial burden.**

I. INTRODUCTION

Binaural sound source localization is an essential part to achieve a friendly Human-Robot Interaction (HRI), because it is equipped with two microphones as the human auditory localization with capability of locating sound source accurately and swiftly. Hence, it is widely applied in acoustic communication, such as intelligent video conference and speech enhancement [1][2]. As to binaural localization, there exists three difficult but important problems: (1) how to accurately localize any kind of speech or sound source; (2) how to localize several different sound sources at the same time; (3) how to track the moving sound sources [3].

Interural Time Difference (ITD) and Interural Intensity Difference (IID) are two significant binaural cues based on differences in time and level of binaural signals. After "Duplex Theory" [4] and cochlear model [5] were proposed, a great number of binaural sound source localization systems have been developed in multiple aspects [6-10]. These algorithms can be loosely divided into three categories: (1) those based upon maximizing the output power of a steered beamformer; (2) approaches employing only time-difference of arrival (TDOA) information by running short-time cross-correlation function; (3) techniques adopting the measured head related transfer function (HRTF).

However, the performance of most methods for source localization in noisy or reverberant environments degrades rapidly and even cannot work [11-13]. Under real scenarios, the noise generated by a robot itself or from environment like air-conditioner disturbs the localization process. Furthermore, the sound waves reflections from walls and furniture in addition to direct path impinge on the ears. Early reflections can have amplitudes similar to that of direct signal, but different directions of arrival interfere with the localization of the real sound source [14]. Moreover, computational complexity is also a limitation to real-time localization systems.

Due to the above reasons, a number of novel algorithms have been proposed recently. For instance, Li et al. proposed a three-layer hierarchical binaural sound source localization system based on Bayes-Rule [15]. Along with the similar hierarchical architectures like Finger et al. [16], experiments showed that hierarchical system could reduce time consumption effectively. Willert et al. introduced a probabilistic model for binaural sound source localization by extracting binaural cues from cochleagrams generated by a cochlear model [17]. As to reverberation, one idea to remove the negative effects is to pass the reverberant signal through a second filter that inverts the reverberation process and recovers the original signal. Jeub et al. presented a novel two-stage binaural dereverberation algorithm which consists of the model of the room impulse response (RIR) and a dual-channel Wiener filter to preserve the binaural cues [18]. Benesty et al. provided a multichannel widely linear approach to deal with the noise reduction of binaural signal [19].

Besides, in some previous works, the binaural signals are decomposed into perceptual bands and the interaural cues are extracted from these bands. When several sources at different locations have significant energy within a given perceptual band, the resulting estimations of azimuth for that band will generally not correspond to any of the actual azimuth of the sources. Raspaud et al. made their study on the Fourier analysis of binaural signals in order to compute the ITD and IID [20].

Accordingly, a robust localization method is proposed to deal with actual environments in this paper. Before local-

This work is supported by National Natural Science Foundation of China (NSFC, No. 60875050, 60675025, 61340046), National High Technology Research and Development Program of China (863 Program, No. 2006AA04Z247), Science and Technology Innovation Commission of Shenzhen Municipality (No. 201005280682A, No. JCYJ20120614152234873), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130001110011).

H. Liu is with Faculty of Key Laboratory of Machine perception (Ministry of Education), Peking University, Shenzhen Graduate School, Beijing, 100871 CHINA. *hongliu@pku.edu.cn*

C. Pang is with the Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Shenzhen Graduate School of Peking University, Shenzhen, 518055 CHINA. *chengpang@sz.pku.edu.cn*

J. Zhang is with the Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Shenzhen Graduate School of Peking University, Shenzhen, 518055 CHINA. *zhangjie827@sz.pku.edu.cn*

ization, there are two procedures need to be conducted, which include cepstral prefiltering for dereverberation of binaural signals and training templates for ITD and IID. Although joint estimation based on ITD and IID has been studied for many years and Parisi et al. dereverberated binaural signals for binaural source localization by cepstral prefiltering [21], they have not considered how to find the more generalized scaling factors which are used to describe the relationship between azimuth and binaural cues. Therefore, the two binaural cues computed from a two channel frequency representation, are combined to estimate the azimuths of sources in recordings. Then, in order to improve the precision of localization and generalization of scaling factors, a generalized parametric model which is evaluated by minimum mean square criterion and do not vary from different subjects is proposed. At last, a two-layer matching strategy based on a Bayesian rule is utilized to make the final decision of azimuth, in which ITD is used to select candidate azimuths in the first layer and IID to refine the results. In our method, generalized parametric model can decrease the space complexity and two-layer strategy costs less time complexity. The experiments contain the simulations based on CIPIC database [22] and several different sound activities localization. In terms of the noisy and reverberant environments, our method can achieve a favourable localization performance.

The rest of this paper is organized as follows: Section II introduces the problem of binaural localization and an algorithm of cepstral prefiltering. Section III shows the details of proposed method including templates establishment, generalized parametric model and two-layer matching strategy. Experiments and discussions are shown in section IV. Finally, conclusions are given in Section V.

II. CEPSTRAL PREFILTERING

A. Localization Description

Let $s[n]$ denote the sound signal emitted by the source in the discrete-time domain, the received binaural signals in a reverberant environment can be modeled as

$$x_i[n] = h_i[n] * s[n] + v_i[n], \ \forall i = l, r, \tag{1}$$

where $h_i[n]$ is the impulse response between the source and ears, and $v_i[n]$ represents the corresponding interference term, which is usually regarded as an uncorrelated, zero-mean, stationary Gaussian random process, l and r mean the left and right channels respectively. The impulse response $h_i[n]$ involves two independent effects which consist of the acoustic property of room (i.e. reverberation) and the Head Related Transfer Functions (HRTFs). The modeling scheme in reverberant environments is illustrated in Fig.1. It shows that the propagation paths from sound source to a receiver include direct path and a series of reflections. The HRTFs are derived from the direct path and the reflections contain the effect of reverberation. As with source, azimuth θ and elevation φ are used to denote direction, which are implicit in HRTFs, then Eq.(1) can be rewritten as

$$x_i[n] = h_i[\theta, \varphi, n] * s[n] + v_i[n], \ \forall i = l, r. \tag{2}$$

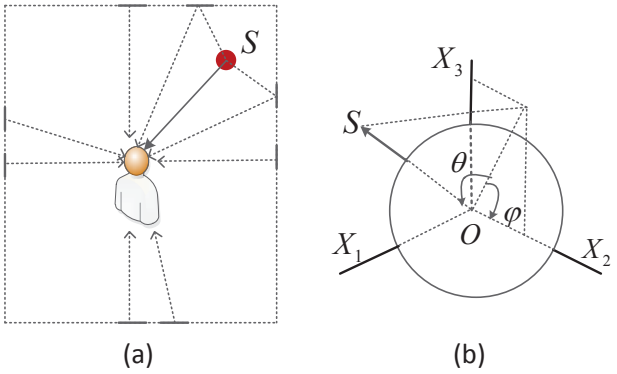

Fig. 1. (a) Signal model of binaural sound localization in reverberant environment. (b) The interaural-polar coordinate system. The azimuth is the angle between a vector to the sound source and the midsaggital or vertical median plane, and varies from -90^o to $+90^o$. The elevation is the angle from the horizontal plane to the projection of the source into the midsaggital plane, and varies from -90^o to $+270^o$.

As to HRI systems and video conference, the azimuth is more important than the elevation in general, thus the former is the main focus in the context.

B. Cepstral Prefiltering

Cepstral prefiltering has been shown to effectively reduce the influence of reverberation on sound source localization [21][23][24]. The complex cepstrum of binaural signal is defined as the following

$$\hat{x}_i[k] = \mathscr{F}^{-1}\{\log\{X_i(\omega)\}\}, \ \forall i = l, r, \tag{3}$$

where $X_i(\omega)$ is the Fourier transform of $x_i[n]$, ω is angular frequency, $\mathscr{F}^{-1}\{\cdot\}$ represents the inverse Fourier transform, $\log\{\cdot\}$ is the complex logarithm and the variable k represents quefrency. Then the cepstral transformation of Eq.(1) can be obtained as

$$\hat{x}_i[k] = \hat{h}_i[k] + \hat{s}[k] + \hat{v}_i[k], \ \forall i = l, r, \tag{4}$$

where $\hat{h}_i[k]$ and $\hat{s}[k]$ are the cepstrum of the room impulse response and source signal respectively. The term $\hat{v}_i[k]$ which denotes the cepstrum of interference is given by

$$\hat{v}_i[k] = \mathscr{F}^{-1}\left\{\log\left(1 + \frac{V_i(\omega)}{H_i(\omega)S(\omega)}\right)\right\}, \ \forall i = l, r, \tag{5}$$

where $V_i(\omega)$, $H_i(\omega)$ and $S(\omega)$ denote the Fourier transforms of $v_i[n]$, $h_i[n]$ and $s[n]$, respectively. In most real applications, there is a common assumption that the interference level is low enough so that $v_i[n]$ and its cepstrum $\hat{v}_i[k]$ are negligible. Since the spectrum can be decomposed into a cascaded system which consists of a *minimum phase component* (MPC) and an *all pass component* (APC), Eq.(4) is equivalent to

$$\hat{x}_i[k] = \hat{h}_{i,mpc}[k] + \hat{h}_{i,apc}[k] + \hat{s}[k] + \hat{v}_i[k], \ \forall i = l, r. \tag{6}$$

Our goal is to reduce the effect of reverberation entirely characterized by $\hat{h}_i[k]$, which is an additive component of binaural signal cepstrum. Specifically, it is reasonable to subtract the part of $\hat{h}_i[k]$ due to reverberation from $\hat{x}_i[k]$. The subtraction can effectively reduce the SRR (signal to

reverberation ratio) in binaural signals. Special attention must be taken in the cepstral filtering to avoid introducing phase distortions, which may cause an inaccurate time delay estimation. $\hat{h}_{i,apc}[k]$ contains important information about the time delay between the binaural signals. Therefore, modification to $\hat{h}_{i,apc}[k]$ is susceptible to lead the serious bias of the final time delay estimation. However, Stéphenne et al. have demonstrated the time delay is relatively insensitive to small modification on $\hat{h}_{i,mpc}[k]$ [23]. These observations drive us to develop a dereverberation algorithm by estimating and subtracting $\hat{h}_{i,mpc}[k]$ from $\hat{x}_i[k]$, then the subtracted cepstrum is transformed back to the time domain through inverse Fourier transform. Based on the above theory, the result generated from the dereverberation algorithm by cepstral prefiltering is shown in the Fig. 2 with reverberation time $T_R = 0.5s$.

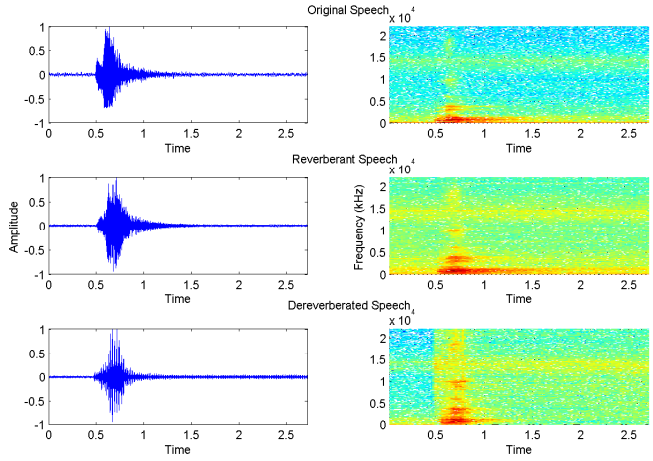

Fig. 2. Dereverbration with cepstral prefiltering ($T_R = 0.5s$)

As shown in Fig. 2, the speech spectrum of reverberant speech becomes fuzzy under the effect of the reverberation compared with the original speech, which would lead to uncorrect extraction of binaural cues for localization. After dereverberation by cepstral prefiltering, the dereverberated speech becomes clear in waveform and the tail in the spectrogram is shorten obviously. Meanwhile, it has enhanced peaks which cause a little distortion, yet the acoustic perceptional intelligence is acceptable by listening test.

III. BINAURAL SOUND SOURCE LOCALIZATION

A. Binaural Cues Extraction

Two different frequently-used observations of acoustic environments are provided by binaural sound recordings after cepstral prefiltering, that are ITD and IID. The two physical cues used in this paper are based on the sliding short-time Fourier transform (STFT) spectra of the binaural signals. As to each frame of binaural signal, the IID (in dB) can be calculated from

$$\Delta I(\omega) = 20\log_{10}\left|\frac{X_r(\omega)}{X_l(\omega)}\right|, \tag{7}$$

where $X_r(\omega)$ and $X_l(\omega)$ are the STFTs of the right and left channel of the binaural signals, respectively. While one or both of the $|X_i(\omega)|$ is null, the interaural differences are regarded as invalid and discarded. Besides, according to the spectra of binaural signal, the ITD is gotten by

$$\Delta T_p(\omega) = \frac{1}{\omega}\left(\angle\frac{X_r(\omega)}{X_l(\omega)} + 2\pi p\right), \tag{8}$$

where p is called the phase unwrapping factor which is a priori unknown integer. The factor is necessary for the fact that the angle which the spectral ratio corresponds to is calculated modulo 2π. However, p makes the phase become ambiguous above a certain frequency, which is mainly dependent on the size and shape of the head. The parameter p indexes these positions, with $p = 0$ corresponding to the source position closest to zero azimuth ($\theta = 0$). A negative p corresponds to a position on the left side ($\theta < 0$). Positive p corresponds to positions on the right side. In this case, possible values of p depend on the physical layout of the sensors and sources. The frequency, which equals twice the largest possible delay between the two ears, corresponds to the highest frequency, because the phase can be estimated without ambiguity. Below this frequency only $p = 0$ is physically realizable. For an average head size the phase ambiguity occur for frequencies above approximately 1500 Hz [20].

B. Templates Establishment

In order to retrieve the azimuth from a given frequency bin of the STFT pair, the IID and ITD measurements of that bin are matched to the measured IID and ITD from the HRTF of the subject. Since the HRTFs are assumed to be time-invariant, they are dependent on the azimuth angle θ instead of time index n. In this case, the templates which consist of intensity difference $\Delta I_s(\theta,\omega)$ and time difference $\Delta T_s(\theta,\omega)$ for subject s are established for localization. As to Eq.(7) and (8), they can be rewritten as

$$\Delta I_s(\theta,\omega) = 20\log_{10}\left|\frac{HRTF_r^s(\theta,\omega)}{HRTF_l^s(\theta,\omega)}\right|, \tag{9}$$

$$\Delta T_s(\theta,\omega) = \frac{1}{\omega}\left(\angle\frac{HRTF_r^s(\theta,\omega)}{HRTF_l^s(\theta,\omega)} + 2\pi p\right), \tag{10}$$

where $HRTF_r^s$ and $HRTF_l^s$ means the HRTFs on the right and left ears respectively.

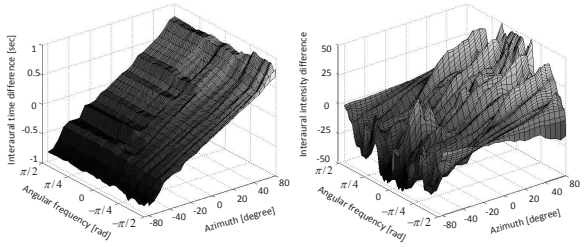

Fig. 3. Interaural time and intensity difference for CIPIC subject #21 versus azimuth and angular frequency.

In Eq.(10), the time difference also depends on the phase unwrapping factor p. The ambiguity is eliminated through unwrapping the modulo 2π phase difference of the HRTFs along frequency. The actual phase difference of the HRTFs

is assumed to be a continuous function versus frequency. Besides, p is supposed to be 0 at $\theta = 0$, where the phase difference ought to be very small. The ITD and IID as functions of azimuth and frequency for one particular head are shown in Fig.3.

C. Generalized Parametric Model

In the previous subsections, a crude method has been presented to extract the binaural cues and discussed how to establish binaural templates in order to lookup azimuths. Here, a precise generalized parametric model for azimuth estimate is proposed.

According to the basic geometric relationship, Raspaud et al. introduced a scaling factor $\alpha(\omega)$ which depends on frequency (and subject)[20] and it is expressed as

$$\Delta T_s(\theta,\omega) = \alpha_s(\omega) r \frac{\sin\theta + \theta}{c}, \tag{11}$$

where r means the "head radius" and c is the sound propagation speed in air (set to be 344 m/s). The formulation between IIDs and azimuth does not obey monotonic function but a more complex function. Based on the study of HRTFs in the CIPIC database, they proposed the following model

$$\Delta I_s(\theta,\omega) = \beta_s(\omega)\sin\theta. \tag{12}$$

The IID and ITD models in the two previous formulations can be optimized for a given head by finding the scaling factors $\alpha_s(\omega)$ and $\beta_s(\omega)$ that give the closest match to the smoothed HRTF data. Since ITD and IID templates have been established by training, the best matching parameters can be resolved by Maximum Likelihood estimation through

$$\begin{aligned} \Delta ET_s(\omega) &= \Sigma_\theta \Big(\alpha_s(\omega) r \frac{\sin\theta + \theta}{c} - \Delta T^T_{s,p}(\theta,\omega) \Big)^2, \\ \Delta EI_s(\omega) &= \Sigma_\theta \Big(\beta_s(\omega)\sin\theta - \Delta I^T_s(\theta,\omega) \Big)^2. \end{aligned} \tag{13}$$

When setting the parameters $\alpha_s(\omega)$ and $\beta_s(\omega)$, the derivative of $\Delta ET_s(\omega)$ ($\Delta EI_s(\omega)$) versus $\alpha_s(\omega)$ ($\beta_s(\omega)$) is set to be zero. Therefore, the $\alpha_s(\omega)$ and $\beta_s(\omega)$ can be given as

$$\begin{aligned} \alpha_s(\omega) &= \frac{c\sum_\theta(\sin\theta+\theta)\cdot\Delta T^T_{s,p}(\theta,\omega)}{r\sum_\theta(\sin\theta+\theta)^2}, \\ \beta_s(\omega) &= \frac{\sum_\theta \Delta I^T_s(\theta,\omega)\cdot\sin\theta}{\sum_\theta \sin^2\theta}. \end{aligned} \tag{14}$$

In practical situations, the average parameters are enough for the azimuth accuracy, i.e. generalized parametric model, which does not differ from subjects so that it can be trained to reduce the complexity of localization. Regarding the number of subjects as N, the generalized $\alpha(\omega)$ and $\beta(\omega)$ can be formulated as

$$\begin{aligned} \alpha(\omega) &= \frac{c\sum_s\sum_\theta(\sin\theta+\theta)\cdot\Delta T^T_{s,p}(\theta,\omega)}{Nr\sum_\theta(\sin\theta+\theta)^2}, \\ \beta(\omega) &= \frac{\sum_s\sum_\theta \Delta I^T_s(\theta,\omega)\cdot\sin\theta}{N\sum_\theta \sin^2\theta}. \end{aligned} \tag{15}$$

Since the $\alpha_s(\omega)$ and $\beta_s(\omega)$ for every subject has been proved to follow the same change trend in [20], the overall change trend can be reflected by the generalized $\alpha(\omega)$ and $\beta(\omega)$. Hence, the generalized parametric model which extracts the generalized scaling factor based on minimum mean square error cannot only improve the generalization ability, but also occupy less storage space.

D. Two-layer Matching Strategy

In order to retrieve the azimuth from the binaural cues by the generalized parametric model, it is necessary to inverse Eq.(11) and (12) based on Eq.(14) or (15) such that

$$\theta_{T,p}(\omega) = g^{-1}\Big(\frac{c}{r\alpha(\omega)}\Delta T_p(\omega)\Big), \tag{16}$$

$$\theta_I(\omega) = arc\sin\frac{\Delta I(\omega)}{\beta(\omega)}, \tag{17}$$

where $\Delta T_p(\omega)$ and $\Delta I(\omega)$ are defined in binaural extraction section, $g^{-1}(\cdot)$ which is the inverse function of $g(\theta) = \sin\theta + \theta$, can not be computed directly. Hence $\hat{g}$ which is a polynomial approximation of g over the interval of interest, is computed by a Chebyshev series, then it can be given as

$$\tilde{g}^{-1}(x) = \frac{x}{2} + \frac{x^3}{96} + \frac{x^5}{1280}. \tag{18}$$

This approximation in Eq.(16) is used in practical localization system.

Since both IID and ITD are a function of the azimuth, they can also be related to each other. The joint evaluation of these quantities proposed in [20] is used in order to improve the azimuth estimate. Specifically, the noisy $\Delta I(\omega)$ (the binaural signals have been dereverberated before) provides a rough estimate of the azimuth for each left/right spectral coefficient pair. Then, this estimate is utilized to choose the "correct" p through selecting the $\Delta T_p(\omega)$ which lies closest because of ambiguity in high frequency bands, thus p can be derived as

$$p = arg\min_p |\theta_{T,p} - \theta_I|. \tag{19}$$

The ITD used in the azimuth estimate here is to make the estimate "more precise", i.e. the standard deviation for a given p of these estimates is smaller. Consequently, in order to reduce the time complexity of localization in realistic scenes, a two-layer matching strategy is designed.

In the first layer, ITD is utilized to make a rough estimate of azimuth because different p leads to ambiguity, and the probability of $P(\theta_{T,p}|ITD)$ represents the normalized distance between $\Delta T_p(\omega)$ and $\Delta T^T_{s,p}(\theta,\omega)$. Then, in the second layer, the correct p is selected and a more precise azimuth can be obtained by combining IID information. The final result is determined based on the Baysian rule through

$$\theta = arg\max_\theta \Big\{ P(\theta_{T,p}|ITD)\cdot P(\theta_I|p,IID) \Big\}, \tag{20}$$

where $P(\theta_I|p,IID)$ denotes the probability of candidate azimuths on the premise of "correct" p and IID, and it is computed by the normalized distance between $\Delta I(\omega)$ and $\Delta I^T_{s,p}(\theta,\omega)$. The detailed process of cepstral prefiltering and the matching strategy is described in Algorithm 1.

Before localization, cepstral prefiltering is performed on $12ms$ time frames using an exponential window as indicated

Algorithm 1: *Binaural sound source localization*

Input: $x_i[n], i = l, r$
Output: azimuth θ
Requirements: ITDs, IIDs, scaling factors
Apply the exponential window $\mu[n] = \alpha^n$ to each frame of $x_i[n], i = l, r$;
Compute the corresponding cepstra $\hat{x}_i[k]$;
Compute the MPC $\hat{x}_{i,mpc}[k]$ of $\hat{x}_i[k]$;
Average $\hat{x}_{i,mpc}[k]$ over successive frames to estimate $\hat{h}_{i,mpc}[k]$;
Cepstrum subtraction: $\tilde{x}_i[n] = \hat{x}_i[k] - \hat{h}_{i,mpc}[k]$;
Transform back to the time domain and apply the inverse exponential window $\mu'[n] = \alpha^{-n}$;
Compute binaural cues $\Delta T_p(\omega)$ and $\Delta I(\omega)$;
Crude $\theta_{T,p}$ estimate using Eq.(16) and count the probability of $P(\theta_{T,p}|ITD)$;
Compute θ_I using Eq.(17);
Find the "correct" p;
Compute the probability $P(\theta_I|p, IID)$;
$\theta = arg\max\{P(\theta_{T,p}|ITD) \cdot P(\theta_I|p, IID)\}$;
return θ

in [23]. What needs to be pointed out is the exponential window $\mu[n] = \alpha^n$, where $0 < \alpha < 1$ and $0 \le n \le K-1$, K being the frame size, which is applied to binaural signals. The function of the exponential window is to move the zeros and poles of the z-transform of binaural signals towards the interior of the unit circle, so as to increase the relative importance of MPC over APC.

Some experimental results are shown in Fig.4 to further illustrate the processing. Darkest regions represent the more likely angles. In order to observe the performance of the proposed method for difference frequencies, a white noise signal is chosen as source signal. A window length of $12ms$ is selected to compute the STFT spectra and binaural cues. The panels in the first row show two-dimensional histograms as function of azimuth and frequency, based on azimuths estimate from ITDs only. It can be seen that above approximately $1-2kHz$, the ITD-based azimuth estimates are ambiguous, which is caused by the different choices of p in Eq.(8). As frequency increases, more values of p are possible, and this figure has not shown the all possible p. The panels in the second row show similar histograms based on azimuths estimate from IIDs. It can be seen that although there is no ambiguity for this case, it has a larger standard deviation than those based on ITD. In addition, IID-based azimuth estimate cannot obtain a visual result in low frequency band. In practice, ITD is used to evaluate some crude azimuth candidates as well as the corresponding probability. Then IID is used to select the "correct" p and refine the azimuth estimate. As a result, the panel in the third row shows the final azimuth by joint estimate of ITD and IID, and the panel in the fourth row shows the probability distribution, from which we can see all the four cases have achieved the correct directions.

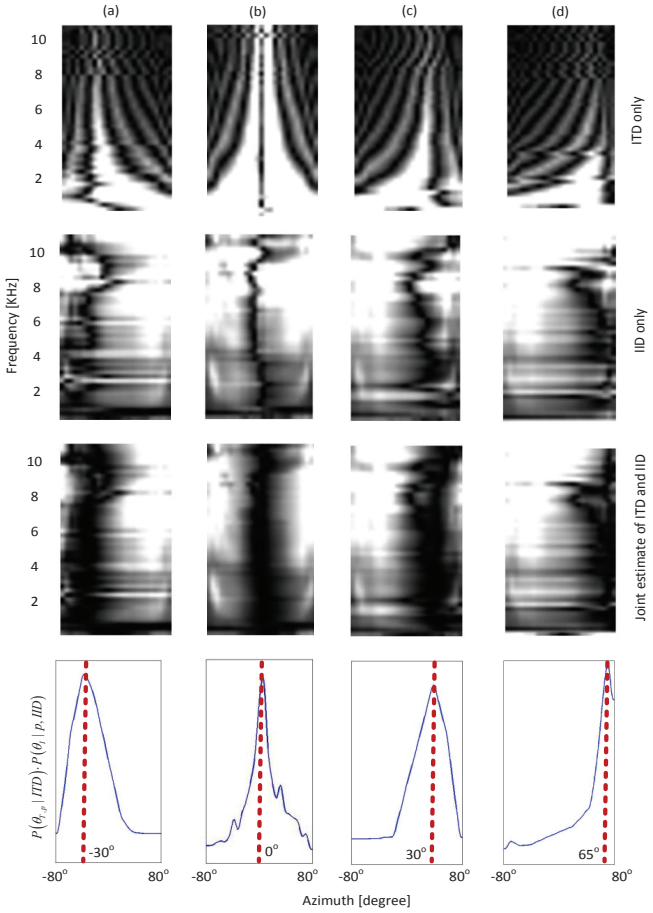

Fig. 4. Histogram of azimuth estimates for four different azimuths angles of -30^o, 0^o, 30^o and 65^o, (a)-(d), respectively. First row: based on ITD only. Second row: based on IID only. Third row: based on joint estimate of ITD and IID using our two-layer matching strategy. Bottom row: marginal probability distributions of localized azimuths.

IV. Experiments and Discussions

The CIPIC database used in this paper is measured by the U. C. Davis CIPIC Interface Laboratory. It contains head-related impulse responses (HRIRs) for 45 different subjects, which include 27 males, 16 females, and KEMAR with large and small pinna. The HRIRs are measured at $1m$ distance with 25 different azimuths, 50 different elevations, totally 1250 directions for each subject [22].

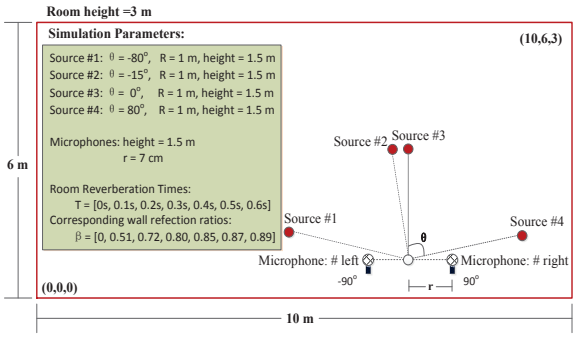

Fig. 5. Simulation scene and parameters of experimental environments. The average radius of heads in CIPIC datasets is $7cm$ approximately.

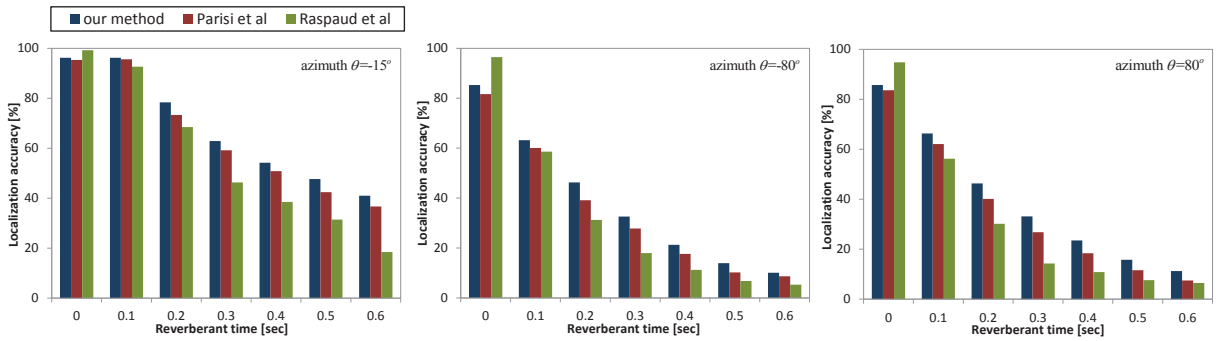

Fig. 8. Localization accuracy at azimuth $\theta = -15^o, -80^o, 80^o$ at different reverberant times.

The experimental environment is a room of $(10 \times 6 \times 3)m$, it is simulated by Roomsim toolbox [25] which is based on the image method [26]. The head is placed at the position $(6 \times 2 \times 1.5)m$. Sound source signal is a musical period, sampled at $44.1kHz$. Since the major focus of this paper is azimuth, the elevation angle is set to 0 and the sound source is positioned at variable horizontal angles with respect to the head. The subject #21 in the CIPIC HRIR database is used as the Kemar head impulse response. The simulation scene and detailed parameters are illustrated in Fig.5.

A. Performance of Cepstral Prefiltering

At first, the performance of cepstral prefiltering verified at different reverberant times is presented in Fig.6. It is obvious that there is a large disparity between the two cases, especially at high reverberant times. Nevertheless, the result without cepstral prefiltering is a little better when $T_R = 0s$, because cepstral prefiltering lead to a little distortion of the dereverberated signal compared with the original signal.

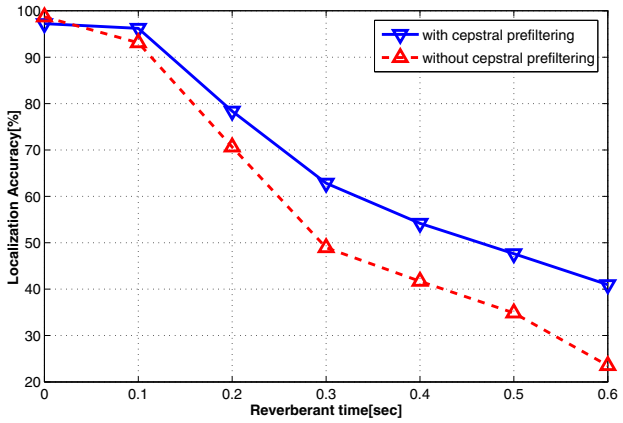

Fig. 6. Comparison of localization accuracy between the method with and without cepstral prefiltering.

When placing a source like speech signal at $\theta = 0^o$ at different reverberant times and testing the proposed method by 100 runs, the average results are shown in Fig.7. It can be seen that though the reverberant time up to $0.6s$, the proposed method achieves the accuracy of 40%, and the false results almost happen right by $\theta = 0^o$. In other words, the cepstral prefiltering is effective to eliminate reverberation.

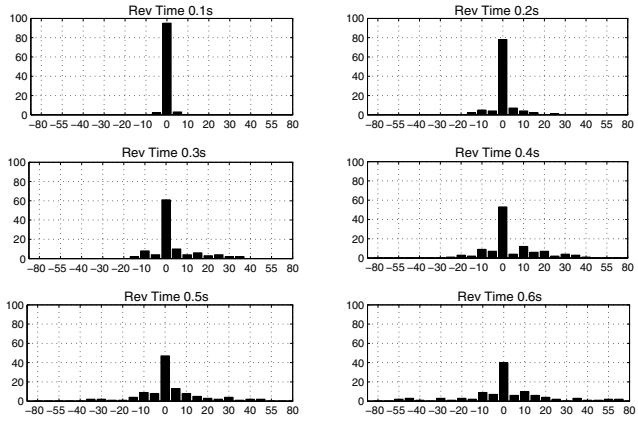

Fig. 7. Histograms of the localization results at azimuth $\theta = 0^o$ at different reverberant times (from $T_R = 0.1s$ to $T_R = 0.6s$).

B. Performance of Generalized Parametric Model

In the next place, the performance of generalized model is further evaluated compared with the method proposed by Raspaud et al. [20] and Parisi et al. [21]. In this experiment, azimuths $(-15^o; -80^o; 80^o)$ are separately estimated by the three methods at different reverberation times from 0 to $0.6s$. It is obvious that our method has obtained the certain superiorities compared with other two methods, especially in strong reverberant time.

From Fig.8, the performance of the Raspaud's method decreases seriously with the increase of the reverberant time, because the extraction of ITD and IID in Raspaud's method is deteriorated by the affect of the reverberation. The Parisi's method achieves some better accuracy than Raspaud's because of cepstral prefiltering. Compared with Parisi's method, our method gets some better result at different reverberant times. The good performance is primarily owing to that the generalized parametric model improve the joint estimate of ITD and IID through finding the optimal scaling factors for any subject, which makes it become more stable and generalized.

C. Comparison with State-of-the-arts

Additionally, some comparisons with several state-of-the-art methods which include Hierarchical System [11], Online Calibration [12] and Interaural Matching Filter (IMF) [6]

TABLE I

THE ACCURACY OF θ IN DIFFERENT SNRS

SNR	Environment without noise			20dB			10dB		
Tolerance	0^o	5^o	10^o	0^o	5^o	10^o	0^o	5^o	10^o
our method	**92.43%**	**98.60%**	**99.81%**	**87.92%**	**97.28%**	**99.02%**	**68.76%**	**82.73%**	**90.34%**
IMF	93.16%	98.52%	99.56%	88.30%	98.20%	99.13%	70.24%	84.96%	92.72%
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%

are carried out in noisy environment without reverberation ($T_R = 0$). From TABLE I, it can be seen that our method achieves the comparable performance with IMF. Specifically, the performances among these four algorithms have small gaps with each other in the 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, they have satisfied the real requirement in quiet 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 such as the spectral differential cues. However, in noisy environments (10/20dB), the proposed method has reached favourable results. This superiority mainly owing to the usage of generalized parametric model, which is only frequency-related based on Maximum Likelihood estimation. Therefore, in variable environments, this method can be of good consistency and robustness compared with the traditional methods. In terms of the localization resolution, we should note that the tolerance 0^o does not mean without error, but $error < 5^o$ instead, and tolerance 5^o means $error < 10^o$.

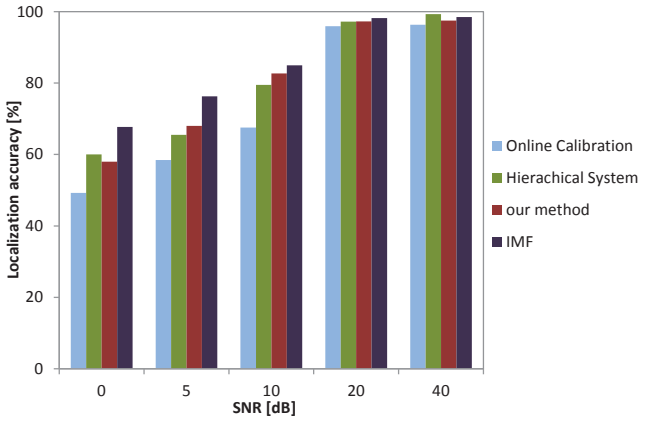

Fig. 9. Localization accuracy of azimuth in different SNRs with 5^o tolerance.

More detailed localization accuracies of the four algorithms with 5^o tolerance in different SNRs are illustrated in Fig.9. It can be obviously acquired that our algorithm works as good as IMF in mild noisy cases and better than the other two. However, in strong noisy environments (i.e. $SNR < 5dB$), the performance of these methods decreases rapidly, because in which case *Hierarchical System* and *Online Calibration* cannot calculate correct ITD for the classical GCC-PHAT cannot extract the notable spectral peak. In addition, strong noise will influence the phase unwrapping, so does azimuth estimate by ITD.

D. Sound Activity Localization

In order to verify the universality for real sound localization system, five different sound activities are used to evaluate the proposed method. All these five sound activities are very common in daily life, including clapping hands, knocking on a door, telephone ringing, screaming and glass smashing, which are recorded in office environment (SNR is 20dB approximately). Fig. 10 shows the azimuth localization accuracies of the five sound activities by the proposed method. It can be seen that these activities are well localized in horizontal direction such that when the tolerance is 0^o, the accuracy of azimuth can achieve more over 85.6%.

Nevertheless, it is worthy to be noted that the accuracies of screaming are slightly lower than the other four sound activities. This phenomenon is mainly caused by the sounding principle, because the intensity of screaming mainly converge to high frequency bands, which leads to the deviation of the correct ITD, yet the localization results are fully suitable for practical application.

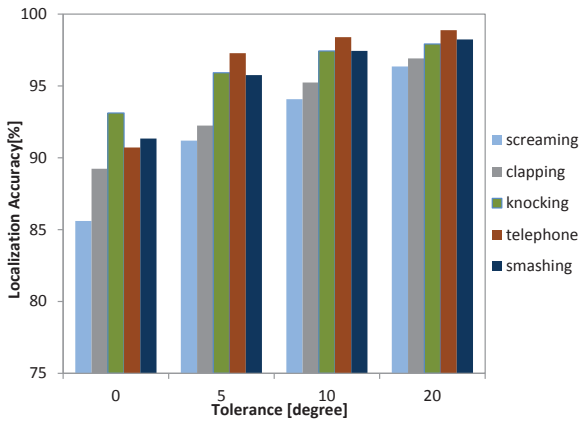

Fig. 10. localization accuracy of different sound activities.

E. Complexity Analysis

Let N_a, N_e, N_c, N_h denote the number of azimuth, elevation, the channels of filterbank and the length of HRTFs in discrete time domain, respectively ($N_h \approx 5N_c$). The algorithms mentioned above all include the process of training, and the templates of ITD, IID should be stored before localization.

TABLE II shows the space complexity of the four algorithms. From TABLE II, the storage of the proposed method is least among the four algorithms, because we only need to store ITDs and IIDs in 25 azimuths versus each angular frequency. However, *IMF*, *Hierarchical System*, and *Online Calibration* must consider all directions ($N_a N_e$) and

N_c frequency channels for binaural templates so that the order of storage of these three algorithms reaches $O(N_a N_e N_c)$.

The time complexity of the four localization algorithms is shown in TABLE III when taking the comparison as the basic operation. It is obvious that our method has achieved the lowest time complexity than the others because of the two-layer matching strategy, with which the previous layer provides candidates for the following layer. Nevertheless, the focus of this paper is the localization of azimuths neglecting the information of elevations, thereby the superiority of computational complexity is inconsequential.

TABLE II

THE SPACE COMPLEXITY OF THE FOUR ALGORITHMS

space	*storage*	*order*
our method	$2N_aN_h$	$O(N_aN_c)$
IMF	$N_aN_eN_c+2N_aN_e$	$O(N_aN_eN_c)$
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 III

THE TIME COMPLEXITY OF THE FOUR ALGORITHMS

time	*times of comparison*	*order*
our method	$2N_a$	$O(N_a)$
IMF	$N_a+N_a(N_e+N_c)$	$O(N_aN_e)$
Online Calibration	$N_a+N_e+N_c$	$O(N_aN_eN_c)$
Hierarchical System	$N_a+N_a(N_e+N_c)$	$O(N_aN_e)$

V. CONCLUSIONS

In this contribution, a novel and effective binaural sound source localization method using generalized parametric model and two-layer matching strategy for reverberant environments is proposed. In practice, the azimuth information is more important than elevation so that the core of this work lies in localizing azimuths. Firstly, cepstral prefiltering is utilized to reduce the influence by reverberation, which could improve the average azimuth accuracy by 3.93%. Then, an generalized parametric model is involved to improve the joint estimation of ITD and IID through extracting the optimal generalized scaling factors, which makes the accuracy of azimuth estimation reach 68.75% in strong noise environment (e.g. SNR=10dB). Finally, a two-layer matching strategy is applied to reduce the computational complexity effectively. In summary, the new method is verified to be efficient and able to adapt to different environments. Our future work will focus on the extraction of robust binaural localization cues in reverberant environment.

REFERENCES

[1] H. Liu, Z. Fu and X. F. Li, "A Two-Layer Probabilistic Model Based on Time-Delay Compensation Binaural Sound Localization", in Proc. IEEE ICRA, pp. 2690-2697, May. 2013.

[2] B. Lee, J. S. Choi, D. J. Kim and M. S. Kim, "Sound Source Localization in Reverberant Environment using Visual information", in Proc. IEEE IROS, pp. 3542-3547, Oct. 2010.

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

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

[5] 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.

[6] H. Viste and G. Evangelista, "Binaural source localization", Proceeding of the 7th Conference on Digital Audio Effects (DAFx'04), pp.145-150, Oct. 2004.

[7] K. Youssef, S. Argentieri and J. Zarader, "A Binaural Sound Source Localization Method Using Auditive Cues and Vision", in Proc. IEEE ICASSP, pp.217-220, 2012.

[8] H. Liu, J. Zhang and Z. Fu, "A New Hierarchical Binaural Sound Source Localization Method Based on Interaural Matching Filter", in Proc. IEEE ICRA, pp.1598-1605, May, 2014.

[9] M. Heckmann, T. Rodemann, F. Joublin, C. Goerick and B. Schölling, "Auditory Inspired Binaural Robust Sound Source Localization in Echoic and Noisy Environments", in Proc. IEEE IROS, pp.368-373, Oct. 2006.

[10] X. F. Li and H. Liu, "Sound Source Localization for HRI Using FOC-based Time Difference Feature and Spatial Grid Matching", IEEE Transactions on Cybernetics, vol.43, no.4, pp.1199-1212, 2013.

[11] H. Liu, J. Zhang, "A Novel Binaural Sound Source Localization Model Based on Time-Delay Compensation and Interaural Coherence", in Proc. IEEE ICASSP, pp.1438-1442, May, 2014.

[12] 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.

[13] 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, no.4, pp.320-327, 1976.

[14] M. Y. Wu and D. L. Wang, "A Two-Stage Algorithm for One-Microphone Reverberant Speech Enhancement", IEEE Transactions on Acoustics, Speech and Language Processing (ASLP), vol.14, no.3, May. 2006.

[15] D. Li and S. E. Levinson, "A bayes-rule based hierarchical system for binaural sound source localization", in Proc. IEEE ICASSP, vol.5, pp.521-524, Apr. 2003.

[16] H. Finger and P. Ruvolo, "Approaches and Databases for Online Calibration of Binaural Sound Localization for Robotic Heads", in Proc. IEEE IROS, pp.4340-4345, Oct. 2010.

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

[18] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant and P. Vary, "Noise Reduction for Dual-Microphone Mobile Phones Exploiting Power Level Differences", in Proc. IEEE ICASSP, pp.1693-1696, Mar. 2012.

[19] J. Benesty and J. D. Chen, "A Multichannel Widely Linear Approach to Binaural Noise Reduction Using an Array of Microphones", in Proc. IEEE ICASSP, pp.313-316, 2012.

[20] M. Raspaud, H. Viste and G. Evangelista, "Binaural Source Localization by Joint Estimation of ILD and ITD", IEEE Transactions on Audio, Speech and Language Processing (ASLP), vol.18, no.1, pp.68-77, Jan. 2010.

[21] R. Parisi, F. Camoes, M. Scarpiniti and A. Uncini, "Cepstrum Prefiltering for Binaural Source Localization in Reverberant Environments", IEEE Signal Processing Letters, vol.19, no.2, pp.99-102, Feb. 2012.

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

[23] A. Stéphenne and B. Champagne, "Cepstral prefiltering for time delay estimation in reverberant conditions, in Proc. IEEE ICASSP, vol.5, pp.3055- 3058, 1995.

[24] R. Parisi, R. Gazzetta, and E. Di Claudio, "Prefiltering approaches for time delay estimation in reverberant environments", in Proc. IEEE ICASSP, vol.3, pp.2997-3000, 2002.

[25] D. R. Campbell, K. Palomäki, and G. Brown, "A matlab simulation of shoebox room acoustics for use in research and teaching", Computer Information Systems, vol.9, no.3, pp.48-51, 2005.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics", The Journal of the Acoustical Society of America, vol.65, pp.943-950, Apr. 1979.