Robust Acoustic Localization Via Time-Delay Compensation and Interaural Matching Filter

Jie Zhang, *Student Member, IEEE*, and Hong Liu, *Member, IEEE*

***Abstract*—Acoustic localization is an essential technique in speech capturing, speech enhancement, video conferencing, and human–robot interaction. However, in practical situations, localization has to be performed in abominable environments, where the presence of reverberation and noise degrades the performance of available position estimates. Besides, the designed systems should be adaptive to locomotion of targets with low computational complexity. In the context, this paper introduces a robust hierarchical acoustic localization method via time-delay compensation (TDC) and interaural matching filter (IMF). Firstly, interaural time-delay (ITD) and interaural level difference (ILD), which are cues involved in first two layers, respectively, are yielded by TDC all at once. Then, a novel feature named IMF, which can eliminate the difference between binaural signals, is proposed in the third layer. The final decision making is based on a Bayesian rule. The relationships among the three layers are that the former layer provides candidate directions for later ones such that the searching space becomes gradually smaller to reduce matching time. Experiments using both a public database and a real scenario verify that TDC and IMF are robust for acoustic localization, and hierarchical system has less consumption time.**

***Index Terms*—Hierarchical acoustic localization, time-delay compensation, interaural matching filter.**

I. Introduction

ACOUSTIC source localization aims at estimating the direction of a sound source by using the collected signals measured from specific acoustic sensors. It has played an important role in various fields such as speech capturing, enhancement, hearing aids, hand-free telephone devices, video conferencing, intelligent human-robot interactions (HRI), etc. Environmental perception and interpersonal communication strongly depend on hearing, where there are three important and difficult issues concerning acoustic localization: 1) How to accurately localize any kind of speech or sound source? 2) How to localize several different sound sources simultaneously, and 3) How to track one or several moving sound sources [1]? For these, Bogaet *et al.* evaluated the influence of three multi-microphone noise reduction algorithms on the ability to localize sound sources by users of hearing aids [2]. Chu *et al.* described a voice source localization algorithm used in the PictureTel automatic camera pointing system for video conference [3]. Wang *et al.* presented a source localization system for robots and navigation based on steered response power-phase transformation algorithm [4]. Ishi *et al.* evaluated a MUSIC-based real-time sound localization system in real noisy environments, which was available for multiple sound sources [5]. Hu *et al.* localized the position of a mobile robot and multiple sound sources simultaneously [6].

There are several kinds of well-known methods for source localization based on a microphone array, including: 1) Directional technologies based on high-resolution spectral estimation; 2) Controllable beamforming technologies by maximizing output power; 3) Approaches based on time difference of arrival (TDOA), which requires lower time consumption. As for HRI or video conferencing scenarios, TDOA-based methods would be more suitable as the azimuthal localization and the rotation of camera should be in real time.

However, for many applications small-sized sensor arrays are required for the localization systems because of a demand for easily-carrying and low computational complexity. As a result, dual-channel acoustic localization has become popular in recent decades. One of the primary abilities of the human auditory system is to localize sound source by two ears. Thus a desirable goal of robotic localization is pinpointing the sound sources swiftly and accurately only by two sensors. Similar to the fact that we cognize sound position by loudness, tone and orientation, the two significant binaural (interaural) cues based on differences in time and level of the sound arriving at two ears should be used. These are interaural time difference (ITD) and interaural level differences (ILD). The ITD, which is caused by the different distances from sound source to sensors, is commonly used in the TDOA-based approaches, and ILD is often brought about by the distrinct attenuation ratios.

Since the "Duplex Theory" [7] and cochlear model [8] were proposed, a large amount of binaural localization algorithms have been developed. Viste *et al.* proposed a method based on the short-time Fourier transform (STFT) spectra of binaural signals for combined evaluation of ITD and ILD for each individual spectral coefficient to localize sources in the horizontal plane [9], [10]. Birchfield *et al.* investigated the possibility of applying ILD to computer-based systems and involved the

Manuscript received August 13, 2014; revised January 29, 2015 and April 14, 2015; accepted May 31, 2015. Date of publication June 19, 2015; date of current version August 13, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Animashree Anandkumar. This work is supported by the National Science and Technology Support Plan (No. 2015BAF15B00), 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). *(Corresponding author: Hong Liu.)*

The authors are with the Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Key Laboratory of Machine Perception (Ministry of Education), Shenzhen Graduate School, Peking University, Beijing 100871, China (e-mail: zhangjie827@sz.pku.edu.cn; hongliu@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2015.2447496

1053-587X © 2015 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

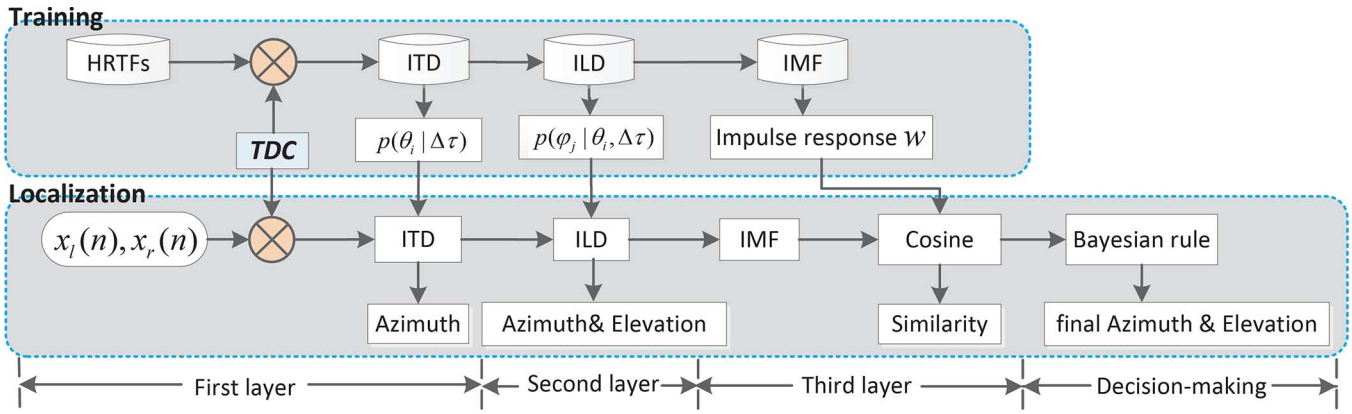


Fig. 1. Block diagram of this framework. The upper training is an offline process providing templates for localization. For convenient drawing, IC before TDC has been ignored and ILD is put after ITD. Indeed, both ITD and ILD are resolved from TDC.

inverse-square-law to localize a sound source by the received energy of two microphones [11]. Cui *et al.* reported a dual-microphone geometrical localization method using time-delay estimate (TDE) and ILD technique [12]. May *et al.* learned the dependency of ITD and ILD on azimuth by azimuth-related Gaussian mixture models (GMMs) and a probabilistic model was presented for localization based on a binaural auditory front-end [13]. Youssef *et al.* addressed a binaural sound source localization method using auditive cues and vision for robotic humanoid context [14].

In addition, among the numerous algorithms, several are related to our work. For example, Li *et al.* introduced a Bayes-Rule based three-layer hierarchical system for binaural sound source localization [15]. Along with the similar hierarchical architectures like Finger *et al.* [16], experiments show that the hierarchical system can reduce time consumption effectively. Willert *et al.* proposed a biologically inspired binaural sound localization system, where binaural cues are extracted using cochleagrams generated by a cochlear model [17]. Jeub *et al.* proposed an interaural coherence model of the room impulse response (RIR) and dual-channel Wiener filter for binaural cues preserving dereverberation [18]. Benesty *et al.* also have made great contributions for binaural noise reduction using a microphone array [19].

These methods are mostly based on the STFT spectra of the input signals, and ILDs and ITDs are estimated for each spectral coefficient. On one hand, ILD-based localization has a relatively large standard deviation, especially at low frequencies. On the other hand, ITD-based localization has smaller standard deviation, but is ambiguous due to phase unwrapping in the Fourier transform. By jointly evaluating these quantities, ILD is used to resolve the ITD ambiguities to effectively improve the azimuth estimates. Furthermore, most of these works decompose the binaural signals into perceptually motivated frequency bands and estimate the interaural cues in these bands. When several sources at different locations have significant energy within a given perceptual band, the resulting azimuth estimates for that band will not, in general, correspond to any of the actual azimuths of the sources. In some cases, therefore, it can be advantageous not to be limited by the frequency resolution of the human auditory system, but rather to estimate azimuths in individual narrow frequency bands. Besides, the traditional binaural cues estimates almost begin with the same standpoint, such as TDE based on generalized cross correlation (GCC), and ILD by logarithmic energy ratio directly. Also as to HRI systems, computational complexity is a vital element that needs to be considered [20].

Accordingly, this paper presents a hierarchical acoustic localization technique using time-delay compensation (TDC) and interaural matching filter (IMF). As with acoustic cues estimates, TDC is used to evaluate ITD and ILD instead of the traditional methods. The proposed TDC foremost yields ITD and ILD concurrently so as to make the realization of binaural cues preservation more convenient. The interaural coherence (IC) [22] is involved to select reliable signal frames to guarantee the validity and stability of ITDs. The TDC and IC mentioned here are a refined and expanded version of a conference proceedings paper [23]. Actually, TDC is an extended version of [25] both in time domain and frequency domain. Then taking 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 the Interaural Matching Filter means. Once the coefficients of IMF in all directions are stored, the process of localization is simplified as calculating the similarity between the unknown coefficients and templates [24]. It is confirmed that IMF implies ITD and ILD, and it can make sure the location of a sound source all by itself in quiet circumstances. After the three cues are obtained, we can first compute the probabilistic distribution of candidate azimuths by the crude ITD, then ILD is used to refine the former candidates including elevations, and the similarity of IMFs is prepared for decision making. The three layers are combined by a Bayesian rule. Fig. 1 shows the block framework of our method including offline training and online localization. Based on head-related transfer functions (HRTFs), the mean value and variance of ITDs and ILDs in each direction can be trained offline as well as for the IMFs. Therefore, the localization space shrinks gradually from the first to third layer referring to templates so that we can achieve lower computational complexity. The experiments are carried out in both simulated enclosures based on CIPIC database [27] and a real hall, where the speech data is collected by two microphones in a 3D space.

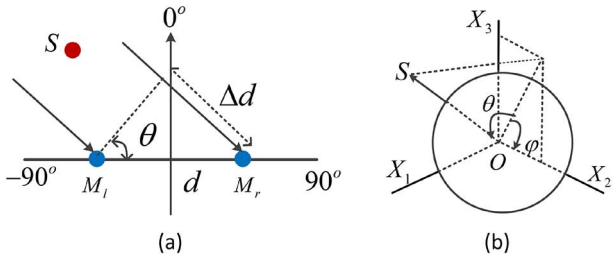


Fig. 2. (a) Signal model of binaural sound localization. For far field, the propagation path of sound source to two eras are thought to be parallel. (b) The interaural-polar coordinate system. The azimuth is the angle between a vector to the sound source and the midsaggital or vertical median plane, and varies from $-90°$ to $+90°$. The elevation is the angle from the horizontal plane to the projection of the source into the midsaggital plane, and varies from $-90°$ to $+270°$.

The rest of this paper is organized as follows: TDC and IMF are introduced in Sections II and III, respectively. The hierarchical localization strategy is described in Section IV. Experiments and analysis are shown in Section V. At last, some conclusions and discussions are drawn in Section VI.

II. Time-Delay Compensation

A. Acoustic Cues Estimates

Considering the far field scenario in Fig. 2, the propagation paths from the sound source to acoustic sensors are roughly parallel. Let $s(n)$ denote a sound source signal, assuming that binaural signals are counterparts of the sound source with time-delay and attenuation to simplify analysis. We then attain

$$\begin{aligned} x_l(n) &= a_l s(n-\tau_l) + v_l(n), \\ x_r(n) &= a_r s(n-\tau_r) + v_r(n), \end{aligned} \tag{1}$$

where a_l and a_r denote the attenuation factors, τ_l and τ_r are the time factors from the sound source to the two acoustic sensors, and $v_l(n)$, $v_r(n)$ are the interferences, respectively. Let us define interaural time-delay $\Delta\tau$ as

$$\Delta\tau = \tau_r - \tau_l. \tag{2}$$

Therefore, taking the idea of time-delay compensation into account, including time alignment and intensity compensation, the relationship between binaural signals will be

$$W \odot x_l(n-\Delta\tau) = \lambda W \odot x_r(n) + \Delta v, \tag{3}$$

where W, λ and Δv denote the window function, attenuation difference and the disparity of noises received by ears, respectively. In fact, Δv is also the error of TDC, and the most amazing task is to make binaural signals without difference. From the standpoint of noises, (3) can be replaced by

$$\Delta v = W \odot x_l(n-\Delta\tau) - \lambda W \odot x_r(n).$$

In an office environment, Δv is usually thought as zero-mean Gaussian noise. Hereby the variance of Δv can be defined as

$$y = \|W \odot x_l(n-\Delta\tau) - \lambda W \odot x_r(n)\|^2. \tag{4}$$

In this context, the parameters λ and $\Delta\tau$ can be estimated by maximum likelihood estimation as follows

$$\frac{\partial y}{\partial \lambda} = \frac{\partial}{\partial \lambda}\|W \odot x_l(n-\Delta\tau) - \lambda W \odot x_r(n)\|^2. \tag{5}$$

After setting this partial derivative to zero, namely interaural level difference (ILD) λ can easily be solved as

$$\widetilde{\lambda} = \frac{\sum_N W^2(n) x_r(n) x_l(n-\Delta\tau)}{\sum_N W^2(n) x_r^2(n)}, \tag{6}$$

where N denotes the length of window. For practical usage, we represent the logarithmic $\widetilde{\lambda}$ as ILD. As with time-delay $\Delta\tau$, it is difficult to compute from $\partial y/\partial\Delta\tau$ directly, but simplifies (4) in the frequency domain instead, that is

$$Y(e^{j\omega}) = \|\boldsymbol{X}_l(e^{j\omega}) e^{-j\omega\Delta\tau} - \lambda \boldsymbol{X}_r(e^{j\omega})\|^2, \tag{7}$$

where $Y(e^{j\omega})$ and $\boldsymbol{X}(e^{j\omega})$ are the Fourier transforms of variance and binaural signals processed by window function, respectively, i.e., $\mathcal{F}\{W \odot x_r(n)\} = \boldsymbol{X}_r(e^{j\omega})$, $\mathcal{F}\{W \odot x_l(n-\Delta\tau)\} = \boldsymbol{X}_l(e^{j\omega}) e^{-j\omega\Delta\tau}$. Therefore, if

$$\boldsymbol{A}(e^{j\omega}) = \boldsymbol{X}_l(e^{j\omega}) e^{-j\omega\Delta\tau} - \lambda \boldsymbol{X}_r(e^{j\omega}),$$

then $\partial Y(e^{j\omega})/\partial\Delta\tau$ can be formulated as

$$\begin{aligned} \frac{\partial Y(e^{j\omega})}{\partial \Delta\tau} &= \frac{\partial}{\partial \Delta\tau}(\boldsymbol{A}^*(e^{j\omega})\boldsymbol{A}(e^{j\omega})) \\ &= \frac{\partial \boldsymbol{A}(e^{j\omega})}{\partial \Delta\tau} \cdot \frac{\partial Y(e^{j\omega})}{\partial \boldsymbol{A}(e^{j\omega})} \\ &= -j2\omega \boldsymbol{X}_l^*(e^{j\omega}) \boldsymbol{A}(e^{j\omega}) e^{-j\omega\Delta\tau}. \end{aligned} \tag{8}$$

Setting $\partial Y(e^{j\omega})/\partial\Delta\tau$ to zero, for $j\omega$ and $e^{-j\omega\Delta\tau}$ are not equal to zero, we obtain

$$\boldsymbol{X}_l^*(e^{j\omega})(\boldsymbol{X}_l(e^{j\omega}) e^{-j\omega\Delta\tau} - \lambda \boldsymbol{X}_r(e^{j\omega})) = 0, \tag{9}$$

where $*$ indicates the complex conjugate. Then, taking (9) back to the time domain using the inverse discrete Fourier transform, it can be shown as

$$\begin{aligned} \delta(n-\Delta\tau) &= R(n) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\lambda \boldsymbol{X}_l^*(e^{j\omega}) \boldsymbol{X}_r(e^{j\omega})}{\boldsymbol{X}_l^*(e^{j\omega}) \boldsymbol{X}_l(e^{j\omega})} \cdot e^{j\omega n} d\omega, \end{aligned} \tag{10}$$

where $R(n)$ is the proposed GCC-TDC function, which rather resembles the Roth weighting [26], [28] based on an optimal filter with $x_l(n)$, $x_r(n)$ as the input and reference signals [29], respectively. Thereout, $\Delta\tau$ can be estimated as

$$\widetilde{\Delta\tau} = \arg\max_n R(n). \tag{11}$$

As a consequence, $\widetilde{\Delta\tau}$ is the optimal time-delay with the meaning of Minimum Mean Square Error (MMSE) criterion.

B. Interaural Coherence

Based on the aforementioned analysis, both ITDs and ILDs can be yielded from the TDC estimator. Combined with (6),

(10), we can draw a conclusion that although there is a mutual relationship between ITD and ILD, λ has an influence on the height of $R(n)$. On the contrary, λ heavily relies on the time-delay, thus the stationary ITDs should be calculated first. Hereby, interaural coherence (IC) is employed into GCC-TDC. The energies of the left and right sensor are evaluated by recursive averaging as

$$E_l(\kappa,\omega) = \alpha \cdot |X_l(\omega)|^2 + (1-\alpha)\cdot E_l(\kappa-1,\omega),$$
$$E_r(\kappa,\omega) = \alpha \cdot |X_r(\omega)|^2 + (1-\alpha)\cdot E_r(\kappa-1,\omega), \quad (12)$$

where κ marks the frame index with each frame of 5.8 ms duration. The smoothing factor α is determined from time constant T and sampling frequency f_s as $\alpha = 1/(T\cdot f_s)$ [30]. Here the IC function can be defined as

$$\gamma(\kappa,\omega) = \frac{E_{lr}(\kappa,\omega)}{\sqrt{E_l(\kappa,\omega)\cdot E_r(\kappa,\omega)}}, \quad (13)$$

where $E_{lr}(\kappa,\omega)$ is the cross-energy spectrum calculated by

$$E_{lr}(\kappa,\omega) = \alpha \cdot X_l(\omega)X_r^*(\omega) + (1-\alpha)\cdot E_{lr}(\kappa-1,\omega). \quad (14)$$

In the following, only cues with $\sum_\omega \gamma(\kappa,\omega)$ above the empirical threshold γ_0 are meaningful, otherwise the frame is thought to be unreliable and abandoned, i.e.,

$$\text{frame} \leftarrow \begin{cases} \text{reliable}, & \text{if } \sum_\omega \gamma(\kappa,\omega) \geq \gamma_0, \\ \text{unreliable}, & \text{otherwise.} \end{cases} \quad (15)$$

Therefore, only when a frame is reliable, we can utilize the TDC to estimate the ITD and ILD. Actually, the IC function acts as a voice activity detector (VAD), which can effectively smooth the binaural cues.

In fact, our TDE is a revised version of generalized cross-correlation (GCC) based on interaural coherence like Roth weighting. Fig. 3 illustrates the comparison of performance between the proposed GCC-TDC and the typical GCC-PHAT. It can be seen that both GCC-PHAT and GCC-TDC achieve relatively accurate ITDs, yet the variance obtained by GCC-TDC is smaller, because GCC-TDC is fundamentally in view of minimizing variance, which brings about more stable ITDs.

III. Interaural Matching Filter

Apart from the classical ITD and ILD as acoustic cues, in this section we will deduce another new binaural feature for localization. In order to eliminate the disparity between the binaural signal $x_l(n), x_r(n)$, we propose a scenario to compose an Interaural Matching Filter (IMF) shown in Fig. 4, whose task is to project $x_l(n)$ onto $x_r(n)$ to make the error $e(n)$ as small as possible [31].

Let $\boldsymbol{w} = [w_0, w_1, \ldots, w_{N-1}]$ be the impulse response of the IMF, and the frame length of $x_l(n), x_r(n)$ is N, the output $y(n)$ of the IMF is obtained from the convolution between $x_l(n)$ and $\boldsymbol{w}$ as

$$y(n) = \sum_{i=0}^{N-1} w_i^* x_l(n-i), \quad n = 0,1,\ldots,N, \quad (16)$$

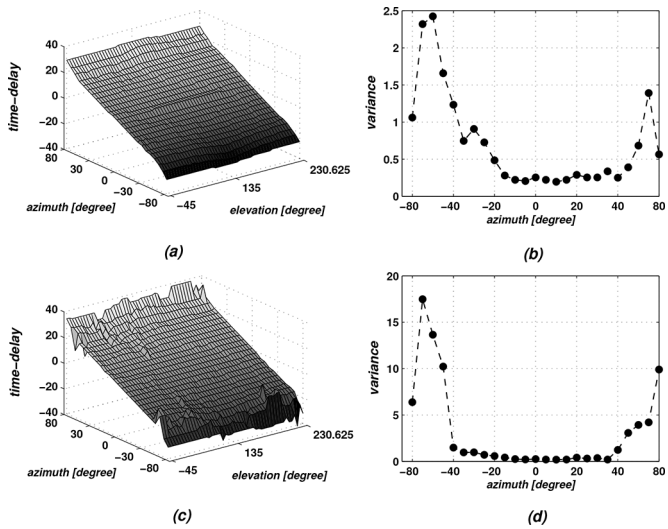


Fig. 3. The comparison of performance between GCC-TDC (upper) and GCC-PHAT (lower).

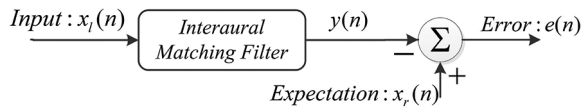


Fig. 4. Linear discrete time Interaural Matching Filter, which implies time-delay and attenuation. Taking $x_l(n)$ as the input of IMF and $x_r(n)$ as the expectation is equivalent to the contrast situation in theory.

where $*$ denotes conjugate, defining the error function as the output of the adder such that

$$e(n) = x_r(n) - y(n).$$

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

$$J(n) = E\{|e(n)|^2\} = E\{e(n)e^*(n)\}, \quad (17)$$

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

$$\sum_{i=0}^{\infty} w_i R_{x_l,x_l}(i-k) = R_{x_l,x_r}(-k), \quad k = 0,1,\ldots,N-1, \quad (18)$$

where R_{x_l,x_l} is the autocorrelation of $x_l(n)$ and R_{x_l,x_r} is the cross-correlation function, which has been calculated in the TDC. If the signal received by left ear is set as

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

then the autocorrelation matrix of $\mathbf{x}_l(n)$ can be expressed as

$$\boldsymbol{R} = E\left\{\boldsymbol{x}_l(n)\boldsymbol{x}_l^H(n)\right\} = \begin{bmatrix} R_{x_l,x_l}(0) & R_{x_l,x_l}(1) & \ldots & R_{x_l,x_l}(N-1) \\ R_{x_l,x_l}^*(1) & R_{x_l,x_l}(0) & \ldots & R_{x_l,x_l}(N-2) \\ \vdots & \vdots & \vdots & \vdots \\ R_{x_l,x_l}^*(N-1) & R_{x_l,x_l}^*(N-2) & \ldots & R_{x_l,x_l}(0) \end{bmatrix}.$$

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

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

Hence, the vector of IMF coefficients can be formulated as

$$\boldsymbol{w} = \boldsymbol{R}^{-1}\boldsymbol{r}. \tag{19}$$

Based on this, we can train offline the impulse response $\boldsymbol{w}$ in all directions using HRTFs or recorded speech, which include the information of TDOA and energetic attenuation. Thereby IMF is capable to represent a specified direction. Furthermore, it is obvious that the similarity between two IMFs can describe the spatial distance of two directions, which inspires that if the IMF of received binaural signals matches certain one in the templates, the sound source is localized. Here a simple but effective cosine-based similarity is adopted as

$$\beta_{\boldsymbol{w}_1\boldsymbol{w}_2} = \frac{\langle \boldsymbol{w}_1, \boldsymbol{w}_2 \rangle}{\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\|}, \tag{20}$$

where $\langle , \rangle$ and $\|\cdot\|$ denote the inner product of vectors and 2nd order norm. If $\boldsymbol{w}_1$ are the coefficients from templates and $\boldsymbol{w}_2$ are the coefficients of received sound, $\beta_{\boldsymbol{w}_1\boldsymbol{w}_2}$ will be the probabilistic distribution of sound source in localization space. In the training process, we can obtain the cost function and the average error function caused by binaural matching as Fig. 5 illustrates based on the CIPIC database including 1250 directions. The subplots (b) and (d) on right panel are the corresponding cost function and the average error function of Fig. 4, and the subplots (a) and (c) depict the results of the contrary scheme, where $x_r(n)$ is taken as the input of IMF and $x_l(n)$ as the expectation. Theoretically, an IMF has the symmetrical property for binaural signals, which means taking $x_l(n)$ as the input of IMF and $x_r(n)$ as the expectation is equivalent to the contrary scheme. Both the two schemes lead to the bigger cost for matching in the directions near to the right ear ($\theta \in [30°, 80°]$), and the bigger error happens in the areas near to the left ear ($\theta \in [-80°, -30°]$). However, apparently the cost and error function of the latter are far severe, which is mainly caused by the measurement error of the database. Additionally, the distinct error will arise when taking advantage of the signals from long distance to design IMFs, because minor signals are always difficult to be captured and processed. Yet to give an overall evaluation and avoid bringing more saltatory cost and error function, the former case will be a better choice. Therefore, we use $x_l(n)$ to input IMFs in this paper.

The impulse responses ($\text{length} = 256$) of IMFs in the time domain are shown in Fig. 6 (When the azimuth $\theta = 0°$, the IMFs of 50 different elevations are embodied in (a). On the other hand, when the elevation $\varphi = 0°$, the IMFs of 25 different azimuths are contained in (b).), from which we can see that the coefficients of IMFs are sensitive to azimuth, because the magnitude of plots in (b) are about 10 times to (a) intuitively. This appearance implicitly reveals a presumption that IMF-based localization is more precise for azimuth than elevation.

The histograms of cosine similarity matrix $\beta_{\boldsymbol{w}_1\boldsymbol{w}_2}$ of several different directions are illustrated in Fig. 7. We utilize a musical

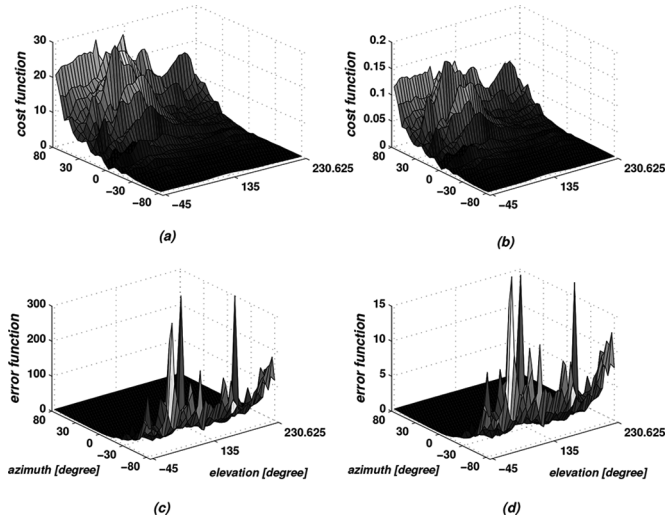


Fig. 5. The right column shows the cost function and the mean of error function in Fig. 4. The left column depicts the response of contrary scheme, where take $x_r(n)$ as the input of IMF and $x_l(n)$ as the expectation.

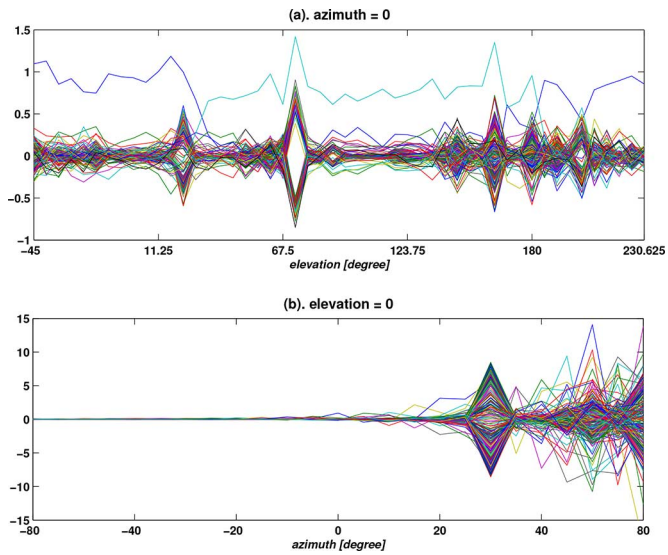


Fig. 6. (a) The impulse response of IMFs when the azimuthal degree is zero. (b) The impulse response of IMFs when the elevation is zero.

period as the sound source. It can be seen $\beta_{\boldsymbol{w}_1\boldsymbol{w}_2}$ generally arrives at maximum value at the direction of sound source located, that is, IMF can be involved for localization to some extend, but not exclude that an ambiguity occurs when $\theta = 0°, \varphi = 0°$. Further more, the peak-like bar is not easily distinguished when $\theta = -40°, \varphi = 0°$ so that it is necessary to combine other cues with IMF as a joint for localization.

IV. Hierarchical Acoustic Localization

For the sound source localization applications, the ITD, ILD and IMF are needed to changed into angels (θ, φ). In the far field, the propagation paths from the source to sensor arrays is thought to be parallel. Considering the geometrical relation in Fig. 1(a), it can be generated

$$\theta = \sin^{-1}(\Delta d/d) = \sin^{-1}(\widetilde{\Delta\tau} c/d f_s), \tag{21}$$

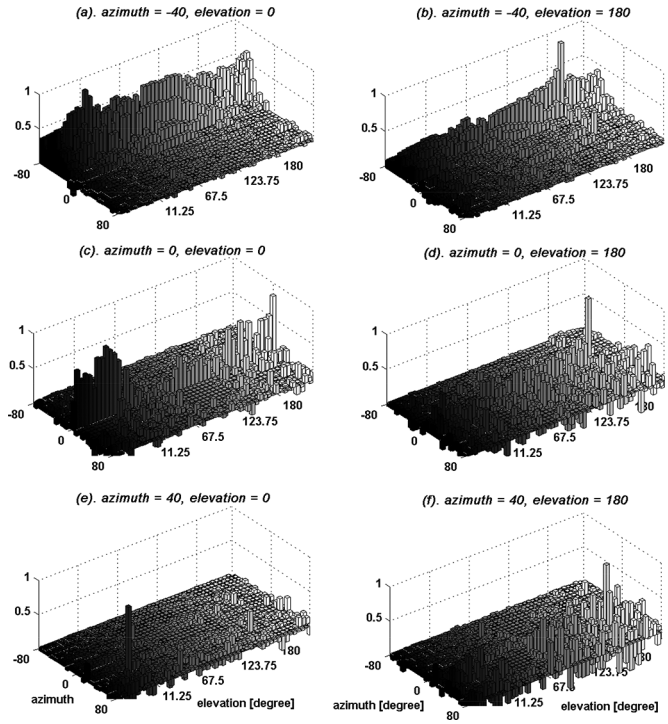


Fig. 7. The histograms of cosine similarity $\beta_{\boldsymbol{w}_1\boldsymbol{w}_2}$ when the sound source is located in different directions.

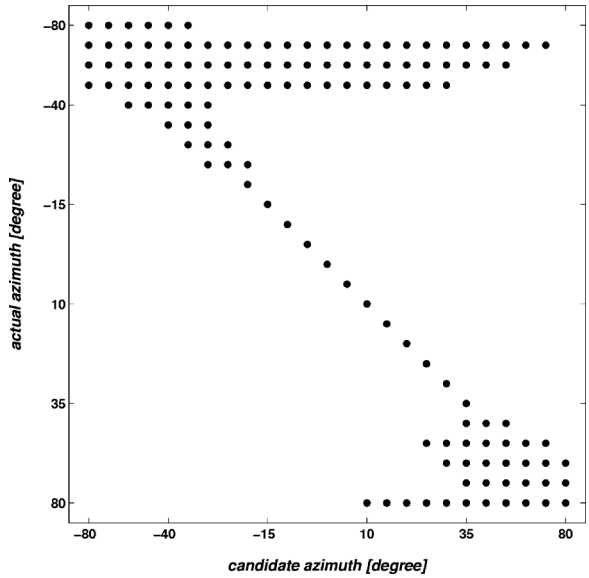


Fig. 8. The evaluated results of azimuth in the first stage. The solid dots represent the possible candidates, for example, if the actual azimuth Θ is $-80°$, the evaluated azimuths θ are $-80°, -65°, -55°, -45°, -40°, -35°$ with different possibilities, and the possibility at $\theta = -80°$ is maximum.

where d is the distance between the two ears (microphones), Δd is interaural distance difference, c is the speed of the sound in air (344 m/s), and f_s is the sampling frequency.

Generally, the acoustic localization is regarded as a pattern classification indeed including two relative progresses, i.e., training and recognition. Training can be based on the HRTFs or prepared auditory dataset, and recognition is online source localization. Firstly, since the azimuth is monotonous to ITD as (21) shows and for a certain azimuth, different elevation shares the same time-delay generally, the mean of time-delay $\overline{\tau_i}$ and the corresponding standard deviation σ_i can be trained for each θ_i versus all elevations. Let N_a denote the number of azimuth such that $i = 1, 2, \ldots, N_a$. Since each time-delay matches one and only θ_i, therefore the probability of θ_i, named $P(\theta_i|\widetilde{\Delta\tau})$, can also be trained before localization. When a new sound source comes, the central azimuth is resolved by (21) and an available azimuthal interval based on ITD is achieved as follows:

$$P(\theta_i|\widetilde{\Delta\tau}) = P(\tau_i|\widetilde{\Delta\tau}) = N\left(\widetilde{\Delta\tau}|\overline{\tau_i}, \sigma_i^2\right),$$
$$\widetilde{\Delta\tau} \subseteq (-3\sigma_i + \overline{\tau_i}, 3\sigma_i + \overline{\tau_i}).$$

Here we save the available interval as candidates with different likelihoods instead of an accurate azimuth, because the areas with larger standard deviation are likely discriminated to the adjacent directions. Thus in order not to omit the potential solutions, ITD is utilized as a coarse azimuthal localization in this step. Visually, the candidates of 25 azimuths in CIPIC are described in Fig. 8.

Then, if we consider the ILD $\widetilde{\lambda}$ similarly, the mean value $\overline{\mu_j}$ and standard deviation δ_j of ILDs can be trained for every direction. Let N_e denote the number of elevation such that $j = 1, 2, \ldots, N_e$. Based on the candidate azimuths in previous stage, the probability of elevation φ_j and available interval of $\widetilde{\lambda}$ are obtained as

$$P(\varphi_j|\theta_i, \widetilde{\lambda}) = P(\widetilde{\lambda}|\widetilde{\Delta\tau}) = N\left(\widetilde{\lambda}|\overline{\mu_j}, \delta_j^2\right),$$
$$\widetilde{\lambda} \subseteq (-3\delta_j + \overline{\mu_j}, 3\delta_j + \overline{\mu_j}).$$

Note that we mainly use ILD to refine elevations for every candidate azimuth. Based on the results of the previous two progresses, what we need to do is to calculate the similarity between the IMFs of candidate directions in templates and the IMF of received signals as

$$P(\beta_{ij}|\theta_i, \varphi_j) = \frac{P(\theta_i, \varphi_j, \beta_{ij})}{P(\theta_i, \varphi_j)} = \arg\max_{\beta_{ij}} \beta_{\boldsymbol{w}_{ij}\boldsymbol{w}_{\text{template}}}|_{(\theta_i,\varphi_j)}. \quad (22)$$

Finally, a Bayesian rule is employed to calculate the probabilistic distribution of the candidate directions to make the final decision expressed mathematically as

$$(\theta, \varphi) = \arg\max_{(\theta_i,\varphi_j)} P(\theta_i, \varphi_j|\widetilde{\Delta\tau}, \widetilde{\lambda}, \boldsymbol{w}_{ij}) = \arg\max_{(\theta_i,\varphi_j)} P(\theta_i|\widetilde{\Delta\tau}) \cdot P(\varphi_j|\theta_i, \widetilde{\lambda}) \cdot P(\beta_{ij}|\theta_i, \varphi_j). \quad (23)$$

Above all, the detailed process is drawn in **Algorithm 1**.

V. Experiments and Discussions

A. Experiments on CIPIC Dataset

The CIPIC database used in our experiments is measured by the U. C. Davis CIPIC Interface Laboratory, which includes HRTFs for 45 different subjects (i.e., 27 males, 16 females, and the KEMAR with large and small pinnas). The database involves 1250 measurements of HRTFs for each subject. These

Algorithm 1: *Hierarchical Acoustic Localization*

Input: ITD $\widetilde{\Delta\tau}$, ILD $\widetilde{\lambda}$, IMF $\boldsymbol{w}_{ij}$
Output: azimuth θ, elevation φ
Templates: ITDs, ILDs, IMFs ;
$\widetilde{\Delta\tau}, \widetilde{\lambda} \leftarrow$ Time-Delay Compensation ;
Design Interaural Matching Filter ;
if $\widetilde{\Delta\tau} \subseteq (-3\sigma_i + \overline{\tau_i}, 3\sigma_i + \overline{\tau_i})$ **then**
  $\theta_i \leftarrow arc\sin(\overline{\tau_i} c/df_s)$;
  $candidate\ azimuths \leftarrow \theta_i$;
  $P(\theta_i|\widetilde{\Delta\tau}) \leftarrow N(\overline{\tau_i}, \sigma_i^2)|_{\widetilde{\Delta\tau}}$;
end
while θ_i *exists* **do**
  if $\widetilde{\lambda} \subseteq (-3\delta_j + \overline{\mu_j}, 3\delta_j + \overline{\mu_j})$ **then**
    transform $\overline{\mu_j}$ into elevation φ_j;
    $candidate\ elevations \leftarrow \varphi_j$;
    $P(\varphi_j|\theta_i, \widetilde{\lambda}) \leftarrow N(\overline{\mu_j}, \delta_j^2)|_{(\widetilde{\Delta\tau}, \widetilde{\lambda})}$;
  end
  while (θ_i, φ_j) *exist* **do**
    $\boldsymbol{w}_{ij} \leftarrow \boldsymbol{R}^{-1}_{x_l,x_l} \boldsymbol{r}_{x_l,x_r}$;
    $\beta_{ij} \leftarrow \frac{<\boldsymbol{w}_{ij}, \boldsymbol{IMF}_{template}>}{\|\boldsymbol{w}_{ij}\|\|\boldsymbol{IMF}_{template}\|}$;
    $P(\beta_{ij}|\theta_i, \varphi_j) \leftarrow \beta_{ij}$;
  end
end
$(\theta, \varphi) \leftarrow arg\max_{(\theta_i,\varphi_j)} P(\theta_i|\widetilde{\Delta\tau}) P(\varphi_j|\theta_i, \widetilde{\lambda}) P(\beta_{ij}|\theta_i, \varphi_j)$;
return (θ, φ)

"standard" measurements were recorded at 25 different interaural-polar azimuths and 50 different interaural-polar elevations at $1m$ distance as Fig. 1 models. Additional "special" measurements of the KEMAR manikin were made for the frontal and horizontal planes [27].

Here we simulate a room-like enclose using the image method [32]. The reverberation time RT60 is fixed to 200 ms, and the corresponding reflection ratios of walls are obtained by the Allen formulation. The head is put in $(6 \times 2 \times 1.5)$ m. A musical period is used as the source signal positioned at different directions. The reference Kemar head impulse response is used, which is subject #21 in the CIPIC HRTF database. The sampling frequency is 44.1 kHz. The simulation scene and detailed parameters are illustrated in Fig. 9.

1) Azimuthal Localization Results: In order to testify the superiority of our hierarchical acoustical localization algorithm, we have compared the azimuthal localization correct rate with different combinations of ITD, ILD and the newly proposed IMF. In addition, several state-of-the-art methods are also compared with this algorithm under different SNR (signal-to-noise ratio) conditions. The necessary parameters used in the experiments are shown in Table I. Note that the frame length of STFT is partially small, because the previous works on one hand have attested that the large frame length is not appropriate for the TDC [25]. On the other hand, the binaural signals with longer frame length would make designing IMF more difficult as well as increasing the computational complexity inevitably. The block length or observation time represents the length of recorded and tested musical signals. Table II shows the localization correct rates of azimuth θ for

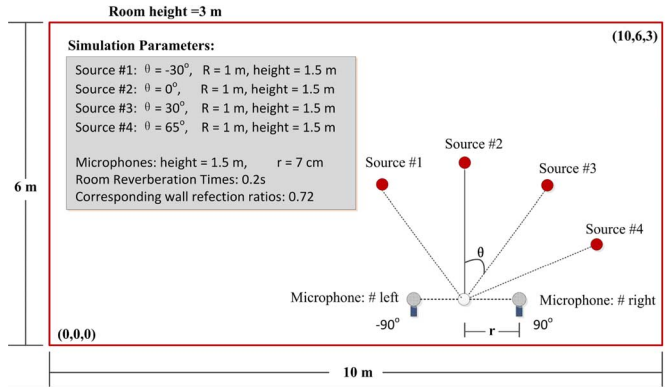


Fig. 9. Simulation scene and parameters of experimental environments.

TABLE I
THE NECESSARY PARAMETERS USED IN EXPERIMENTS

Parameter	Value
Sampling frequency	44.1kHz
Frame length (STFT length)	256 points
Frame shift	128 points
Block length (observation time)	2 s
smoothing factor	0.95
Processor type	i5-2320 @ 3.00GHz

different SNRs using three different combinations of ITD, ILD and IMF. Among ITD, ILD and IMF, there are seven different combinations totally. Above approximately 1–2 kHz, the ITD-based azimuth estimates are ambiguous, which is caused by the different choices in phase unwrapping. Although there is no ambiguity for ILD-based localization, it has a larger standard deviation than those based on ITD. In addition, the ILD-based azimuth estimate cannot obtain a visual result for low frequency bands. In practice, we can use ITD to evaluate some crude azimuth candidates as well as the corresponding probabilities. Then ILD is used to help phase unwrapping and refine the azimuth estimate. Furthermore, IMF is designed from the point of eliminating the disparity between binaural signals so that it concludes some delay and attenuation units, which are reflected in ITD and ILD. In other words, IMF is a complex association of ITD and ILD in fact. That is why we merely select three combinations, i.e., IMF, ITD + ILD and ITD + ILD + IMF, for comparisons.

From Table II, we can see that in general the ITD + ILD-based method achieves more precise solutions than the IMF-based one in severe environments such as when SNR $= 0$ dB. However, when no noise is added, IMF outperforms ITD + ILD obviously, and IMF even achieve 100% correct rate with 10° tolerance. A more detailed comparison is illustrated in Fig. 10(a). By and large, the IMF-based azimuth estimate can work rather effectively in favorable encloses, but the performance degrades rapidly with the SNR decreasing. In contrast, the ITD + ILD-based estimate can obtain better results under low SNR environments. The noise used in our experiments is white Gaussian noise. This is mainly due to that the random noise has corrupted the available binaural signals in the severe environments, which lead to extreme strait for designing IMFs. At the same time, the TDC estimator can considerably extract accurate ITD and ILD. Therefore, combining ITD, ILD and

TABLE II
THE LOCALIZATION CORRECT RATE OF AZIMUTH θ IN DIFFERENT SNRS BY SEVERAL DIFFERENT COMBINATION OF ITD, ILD AND IMF. ITD AND ILD ARE COMPUTED BY TIME-DELAY COMPENSATION

SNR	*No Noise*			*20dB*			0dB		
Tolerance	$=0°$	$\leq 5°$	$\leq 10°$	$=0°$	$\leq 5°$	$\leq 10°$	$=0°$	$\leq 5°$	$\leq 10°$
IMF	98.50%	99.72%	100%	80.40%	82.88%	84.86%	27.76%	34.80%	40.32%
ITD+ILD	90.12%	95.04%	96.83%	76.96%	89.30%	92.61%	62.16%	65.68%	74.00%
ITD+ILD+IMF	99.12%	100 %	100%	84.26%	95.92%	98.24%	65.94%	67.52%	75.23%

TABLE III
THE LOCALIZATION CORRECT RATE OF θ IN DIFFERENT SNRS COMPARING ITD + ILD + IMF-BASED AZIMUTH ESTIMATES WITH OTHER METHODS

SNR	*No Noise*			*20dB*			*0dB*		
Tolerance	$=0°$	$\leq 5°$	$\leq 10°$	$=0°$	$\leq 5°$	$\leq 10°$	$=0°$	$\leq 5°$	$\leq 10°$
ITD+ILD+IMF	99.12%	100 %	100%	84.26%	95.92%	98.24%	65.94%	67.52%	75.23%
TDC−Fu	90.28%	98.48%	99.84%	87.56%	97.56%	98.96%	49.16%	65.68%	74.04%
Probabilistic Model	92.72%	98.64%	100%	75.94%	83.96%	87.74%	40.64%	51.23%	63.20%
Online Calibration	89.12%	96.76%	99.24%	84.26%	95.92%	98.24%	44.23%	56.32%	65.10%
Hierarchical System	93.90%	98.70%	99.87%	85.64%	97.21%	98.72%	49.64%	64.50%	73.30%

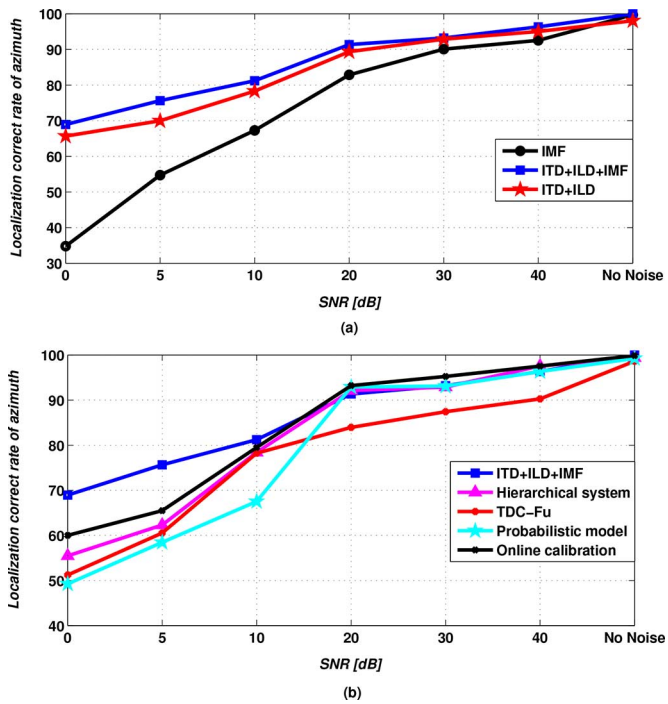


Fig. 10. (a) Comparing the azimuthal localization correct rate with different combinations of ITD, ILD and IMF when the tolerant error is 5°. (b) Comparing the azimuthal localization correct rate using several popular methods when the tolerant error is 5°.

IMF for acoustic localization is reasonably significative under the complex circumjacent conditions.

Table III shows some azimuthal estimate comparisons between ITD + ILD + IMF and several other state-of-the-art methods, such as Hierarchical System [15], Online Calibration [16], Probabilistic Model [17], TDC-Fu [25]. As a matter of fact, both [15] and [16] belong to the hierarchical methods with different binaural cues. In [17], Willert *et al.* proposed a probabilistic map to describe to the relationship between ITD and ILD. The probabilistic map was generalized to be TDC as a localization cue in [25], where [15], [16] and [17] have been compared already. In [23], [24], we extended TDC to estimate ITD and ILD, and used IMF as a new binaural cue. That is why this paper compares these works together. We can conclude that in most cases ITD + ILD + IMF has achieved the best results generally, especially when the enclose is definitely quite ITD + ILD + IMF displays a tremendous superiority. Some similar results have been shown in Fig. 10(b). Specifically saying, the performances among these five algorithms can amount 95% (Tolerance $=5°$), which is quite satisfactory. In details, ITD + ILD + IMF is the best one reaching 100%, which benefits from the effective ability of IMF to denote direction. Besides, Hierarchical System has suboptimal performance while Online Calibration is the worst one, which is probably due to the different cues used. In Hierarchical System, ITD, ILD and differential spatial cues are involved in three layers, respectively, so that it can work better than two-layer systems in principle. As to Probabilistic Model, the activity maps built by the binaural cues are pretty affected by noise so that it localizes well against high SNRs. Based on the Probabilistic Model, TDC-Fu actually revised the activity maps as compensated intensity difference and it also took a two-layer framework subsidiarily, thereby it localize better than the former generally.

However, there are small gaps among them in the noisy environments. Intuitively, except the cases when 40 dB >SNR > 10 dB, Probabilistic Model lags others clearly, which is caused by the inaccurate activity maps affected by noises, and the other four curves twine together. Yet when dealing with low SNRs, ITD + ILD + IMF, TDC-Fu and Hierarchical System can work better than the others obviously. In TDC-Fu, $PHAT-\rho\gamma$ is utilized in the first layer for robust TDE, but Probability Model does not have this trait. Online Calibration has two layers similar with Hierarchical System, but the difference lies in calculating ITDs in frequency subbands, which is not particularly useful for azimuthal localization.

2) Elevation Localization Results: The localization space of experiments conducted in this subsection is divided into 50 elevations ranging from $-45°$ to $+230.625°$ in step of $5.625°$. This increment divides the full 2π angular into 64 equal parts, but only 50 values are used in CIPIC measurement, because the space shielded by the body can be neglected. Similar to the analysis in azimuthal localization, firstly we observe the elevation localization results based on the three different combinations of

TABLE IV
THE LOCALIZATION CORRECT RATE OF ELEVATION φ IN DIFFERENT SNRS BY SEVERAL DIFFERENT COMBINATION OF ITD, ILD AND IMF. ITD AND ILD ARE COMPUTED BY TIME-DELAY COMPENSATION

SNR	*No Noise*			20*dB*			0dB		
Tolerance	$=0^\circ$	$\leq 5.625^\circ$	$\leq 11.25^\circ$	$=0^\circ$	$\leq 5.625^\circ$	$\leq 11.25^\circ$	$=0^\circ$	$\leq 5.625^\circ$	$\leq 11.25^\circ$
IMF	86.21%	90.12%	100%	48.88%	64.88%	84.88%	3.76%	9.12%	15.04%
ITD+ILD	76.68%	92.96%	95.20%	45.92%	65.28%	72.56%	6.56%	14.72%	20.96%
ITD+ILD+IMF	88.35%	93.25%	100%	51.69%	69.54%	86.41%	7.68%	20.47%	25.84%

TABLE V
THE LOCALIZATION CORRECT RATE OF φ IN DIFFERENT SNRS COMPARING ITD + ILD + IMF-BASED ELEVATION ESTIMATES WITH OTHER METHODS

SNR	*No Noise*			20*dB*			0dB		
Tolerance	$=0^\circ$	$\leq 5.625^\circ$	$\leq 11.25^\circ$	$=0^\circ$	$\leq 5.625^\circ$	$\leq 11.25^\circ$	$=0^\circ$	$\leq 5.625^\circ$	$\leq 11.25^\circ$
ITD+ILD+IMF	88.35%	93.25%	100%	51.69%	69.54%	86.41%	7.68%	20.47%	25.84%
TDC − Fu	70.48%	92.13%	94.65%	20.38%	45.96%	61.05%	6.56%	14.72%	20.98%
Probabilistic Model	73.28%	95.15%	97.67%	12.39%	25.28%	48.57%	5.67%	9.48%	16.93%
Online Calibration	63.61%	90.25%	95.82%	17.22%	39.49%	56.92%	4.13%	10.45%	17.62%
Hierarchical System	64.77%	92.47%	95.23%	20.34%	43.62%	60.92%	7.73%	14.25%	19.19%

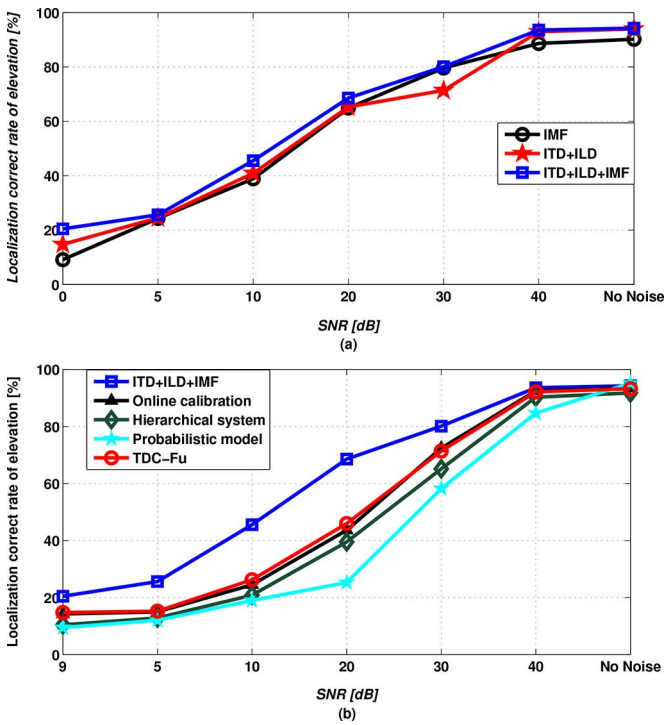


Fig. 11. (a) Comparing the elevation localization correct rate with different combinations of ITD, ILD and IMF. (b) Comparing the elevation localization correct rate using several popular methods (Tolerance $= 5.625^\circ$).

ITD, ILD, and IMF. The elevation localization accuracy for different SNRs is shown in Table IV. Obviously, IMF has acquired better results compared to the ITD+ILD-based method, and ITD + ILD + IMF is the best choice. However, the performance here is particularly worse than that of azimuth, which reveals that IMF is more affected by the azimuth. We have presumed that IMF is more sensitive to the azimuth in Section IV, therefore the phenomenon in Fig. 6 is confirmed. A more detailed comparison is illustrated in Fig. 11(a). It is obvious that ITD + ILD + IMF is most accurate for the elevation localization, thus IMF is helpful as an acoustic cue for localizing.

Table V shows the compared results of elevation localization correct rate for different SNRs between the ITD + ILD + IMF-based estimates and the aforementioned popular localization methods. We can see that ITD + ILD + IMF has presented a more prominent advantage than that in azimuthal localization. Simultaneously, it is can be concluded that the existing state-of-the-art algorithms are less robust to elevation. This is mainly due to that elevation localization largely depends on ILD, but ITD offers little help. Besides, the precise extraction of ILD is crucially influenced by the environmental conditions, accordingly those ILD-related algorithms with noise reduction unit are more likely to achieve better performance in general. Since IMF implies ITD and ILD, and we have use ILD to refine elevations in the second layer, ITD + ILD + IMF has reached the best performance predictably. Like the azimuthal localization, ITD + ILD + IMF obtains excellent results in the environments without noise, such as 100% correct rate when Tolerance $\leq 11.25^\circ$. Since Probabilistic Model can work well against the high SNRs, it gets the proxime accessit in the quiet environments, and then TDC-Fu. Nevertheless, it has less noise immunity so as to lag others far behind against low SNRs, but TDC-Fu, Online Calibration and Hierarchical System would localize much better. Indeed, all the five algorithms are all ILD-related. For example, they all take advantage of time difference and intensity difference in the first and second layer, respectively. As to Hierarchical System, the differential spectral cues used in the third layer do not include any energetic information, yet it aims to distinguish the front-end ambiguity. We can conclude that IMF is more effective than the differential spectral cues for elevation localization.

Some more distinct comparisons of the five algorithms for different SNRs with 5.625° tolerance is illustrated in Fig. 11(b). We can see that in the two extreme environments, they have little disparity in accuracy compared to each other on one hand. Note that we should be more engaged on noise reduction instead of modifying the localization strategy or acoustic cues in the supremely noisy surroundings, such as SNR $\leq$ 0 dB. In addition, in the quite noise-free environments the pre-existings can work accurately. Therefore, we should pay more attention to moderate noise field, such as working offices or meeting rooms, and other acoustic adverse factors for sound localization. However, on the other hand when 5 dB $\leq$ SNR $\leq$ 30 dB,

Fig. 12. Scene graph of experimental environments. Red circles represents the places of omnidirectional microphones.

ITD + ILD + IMF is in the lead clearly. Seeing that elevation is not monotonous with ITD or ILD, hence improving the elevation localization correct rate should begin with using more individual acoustic features to represent directions or formulating elevation versus some parameters. Fortunately, as to practical systems, the azimuth localization is frequently more important than that of elevation.

B. Experiments for Human-Robot Interaction

Our Human-Robot Interaction system is designed on a mobile robot using a MARIAN TRACE8 multi-channel audio sample card and two BSWA MPA416 microphones. Multi-threading programming based on Direct Sound is adopted to ensure the synchronization of audio signals. The scene graph of robot and microphone array are shown in Fig. 12.

Experiments are implemented in a room of 8×8 m^2, where the reverberation time RT60 is about 350 ms approximately. Combining with the robot's size, the linear distance of microphones is set to 40 cm, which are placed on the shoulder of the robot. Given that the height of the mouth is about 150 cm for a standing adult, the plane of microphone field is chosen as the standard when evaluating the localizing performance on the horizontal plane and all the localization results have to be projected onto this standard plane. The robot is placed in the center of the room, with another 96 points on the floor as tested sound sources. The sound sources are evenly distributed in every 15° and four positions at each direction with a distance of 1, 2, 3, 4 m to the center, respectively. At each position, 21 groups of speech data from different people are recorded, the contents of speech are the Chinese words "dingwei", "pengpeng" and "guolai", which mean "location", "robot's name" and "come here", respectively.

Our experiments are conducted at working duration including noises from air conditioner, fans and other silence-breakers, where 10 dB $\leq$ SNR $\leq$ 20 dB. We use the TDC algorithm to estimate ITD and ILD displayed in Fig. 13, where the radius of sound source is 2m. It can be seen that ITD has a very weak

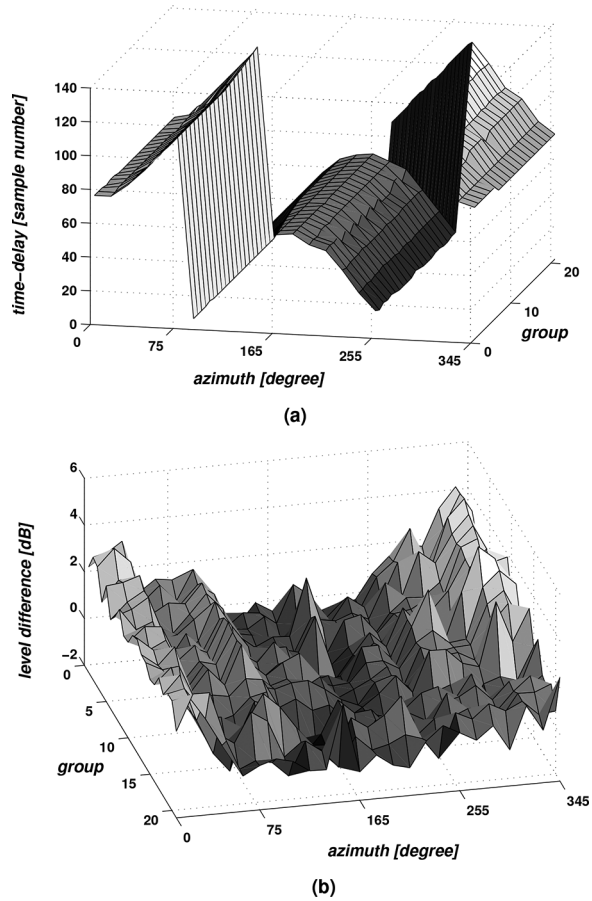


Fig. 13. The acoustic cues (time-delay and level difference) estimates based on time-delay compensation when the radius is 2 m.

TABLE VI
LOCALIZATION ACCURACY OF HORIZONTAL AZIMUTHS BASED ON ITD + ILD + IMF

Error	5°	10°	15°	20°
1*m*	94.25%	97.42%	98.41%	99.63%
2*m*	90.55%	92.35%	94.60%	95.58%
3*m*	87.22%	89.93%	91.36%	92.31%
4*m*	83.13%	85.52%	88.82%	90.19%
Mean	88.79%	91.31%	93.30%	94.43%

fluctuation across different groups in each azimuth, and ILD has an obvious distribution tendency despite irregularity within groups. Further more, here we have not shown the IMFs of this scene like the CIPIC database. Using a musical period as the sound source, the localization accuracy of horizontal azimuths based on ITD + ILD + IMF is presented in Table VI. It can be seen that our method effectively achieves accurate results of 90% almost with reasonable error tolerance under the real conditions. In most real HRI systems, azimuths are enough for a robot to apperceive directions to interact with human. In addition, from Table VI we can also see that the overall angular error is bigger when the sound source is farther away from the robot, i.e., the localization accuracy decreases with distance increasing. Actually, the distance localization is more difficult in the far field and those methods based on large-scale geometrical microphone arrays are more likely to obtain accurate solutions.

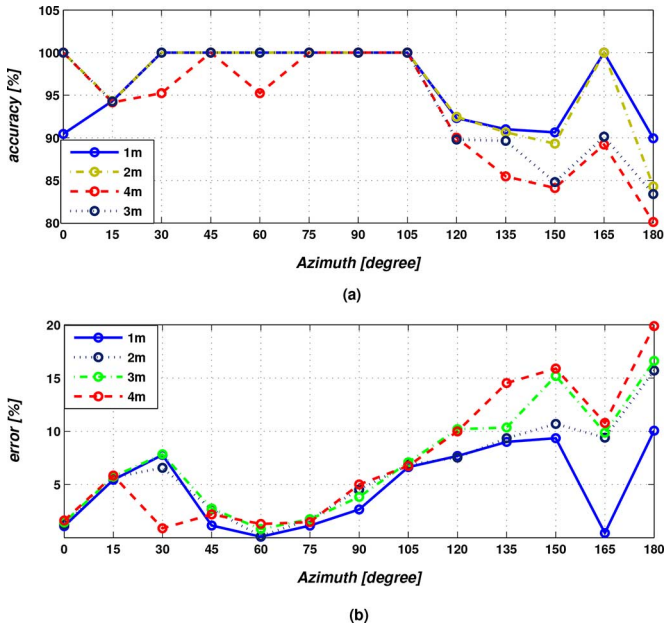


Fig. 14. The average localization accuracy and error in each direction.

TABLE VII
THE LOCALIZATION ACCURACY OF DIFFERENT SOUND ACTIVITIES

Error	0°	5°	10°	20°
clapping	86.24%	92.24%	94.24%	95.92%
knocking	93.12%	95.92%	97.44%	97.92%
telephone	90.72%	97.28%	98.40%	98.88%
screaming	88.04%	95.76%	97.44%	98.24%
smashing	85.60%	93.20%	96.08%	97.36%
Mean	88.75%	94.88%	97.32%	97.66%

In addition, we have counted the average localization accuracy and error for each direction of our HRI system as Fig. 14 shows. We can conclude that the directions in front of the robot are accurately localized and it is more difficult for the two sides (near 0° and 180°). This is probably due to that ITD and IMF have larger variance in those directions around the two sides, which was already conveyed in Figs. 3 and 5, respectively, so that the HRI system is more possible to localize to the abutting directions ambiguously. Besides, when two microphones are used for acoustic localization in the 2-D space free field, two points sharing the same ITD will exist and this phenomena is called front-back confusion, which is neglected and the front space is merely localized. If an artificial pinna for audio microphone is added, this front-back ambiguity will be eliminated, which is also one of our future works.

In order to verify the universality for the realistic sound localization, five different sound activities are expected to evaluate this 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 an office environment (SNR $\approx$ 20 dB). Table VII shows the localization accuracy, from which it can be seen that these activities are well localized in the horizontal plane such that when the tolerance is 5°, and the average accuracy of azimuth can achieve more than 94.8%.

TABLE VIII
THE SPATIAL COMPLEXITY OF THE FIVE ALGORITHMS

Method	*storage*	*order*
$ITD+ILD+IMF$	$2N_aN_e+N_aN_eN_h$	$O(N_aN_eN_c)$
$TDC-Fu$	$N_aN_eN_c+N_aN_e$	$O(N_aN_eN_c)$
Online Calibration	$N_aN_eN_c+N_aN_e$	$O(N_aN_eN_c)$
Hierarchical System	$N_aN_eN_c+2N_aN_e$	$O(N_aN_eN_c)$
Probabilistic Model	$N_a^2N_eN_c$	$N_a^2N_eN_c$

TABLE IX
THE TIME COMPLEXITY OF THE FIVE ALGORITHMS

Method	*times of comparison*	*order*
$ITD+ILD+IMF$	$N_a+N_e+N_h$	$O(N_e)$
$TDC-Fu$	$N_a+N_e+N_c$	$O(N_e)$
Online Calibration	$N_a+N_e+N_c$	$O(N_e)$
Hierarchical System	$N_a+N_a(N_e+N_c)$	$O(N_aN_e)$
Probabilistic Model	N_aN_e	N_aN_e

However, it is worth noting that the accuracies of glass smashing are slightly lower than the others'. This phenomenon greatly depends on the sounding principle, because the glass smashing has blank sounding time-slots and its energy mainly distributes in the high frequency bands, thus performing time-delay compensation overall might not be appropriate but in sub-bands. As we know, ITD has an ambiguity in the high frequency bands because of difference choices of phase unwrapping, and selecting reliable frequency sub-bands is usually difficult. Fortunately, within 10° of erroneous tolerance our system has achieved more than 95% of correct rate such as to satisfy the practical applications.

C. Complexity Analysis

Now we continue to analyze the computational complexity of our algorithm. Let N_a, N_e, N_c, N_h denote the number of azimuth, elevation, the channels of filterbank and the order of IMFs, respectively, and we have $N_h \approx 5N_c$. Considering the algorithms mentioned above definitely concluding the training procedure, the templates of ITD, ILD and IMF should be stored before localization.

The spatial complexity of the five aforementioned algorithms is shown in Table VIII. It can be observed that the storage of ITD + ILD + IMF is not the maximum, because we only need to store ITDs, ILDs and IMFs in N_aN_e directions. However, although TDC-Fu and Online Calibration just need to store two acoustic cues as templates, they must take the divided frequency sub-bands into account instead. Hierarchical System still requires to consider the differential spectral cues in all sub-bands, and Probabilistic Model builds an activity map for N_aN_e directions versus frequency in the frequency domain. Thus the newly proposed IMF has not induce excess storage load for hardware realization.

When taking the comparison as the basic operation, the time complexity of the five algorithms is shown in Table IX. It is clear to see that this method has achieved the lowest time complexity than others because of the hierarchical acoustic localization strategy, with which the previous layer provides candidates for the following layer. We have counted the time consumption of these five algorithms by random tests of 800 times,

and ITD + ILD + IMF successfully reduces the time consumption from 0.5 s of TDC-Fu down to 0.35 s approximately, which greatly relates to:

- ITD + ILD + IMF calculates ITDs and ILDs simultaneously based on time-delay compensation, which decreases the steps to evaluate the binaural cues and makes coding procedure more concise.
- The excellent matching strategy of hierarchical framework can also deflate candidates directions effectively.

Although, computer processors have developed significantly nowadays and computational complexity is no longer a limitation to certain extend, the real-time requirement is still rather crucial to localization systems, especially in some situations like speaker tracking, video conferencing, etc.

VI. Conclusions and Future Works

This paper proposes a robust hierarchical acoustic localization method based on time-delay compensation (TDC) and interaural matching filter (IMF). First, TDC algorithm is foremost utilized to evaluate the common binaural cues, i.e., ITD and ILD in the frequency and time domain, respectively. Although the proposed GCC-TDC function is much alike to the famous Roth weighting, the two mentalities begin with different standpoints. Besides, we take advantage of interaural coherence to select reliable frames for GCC-TDC, which could decrease the variance of ITDs compared to GCC-PHAT.

Actually, the newly designed acoustic cue IMF implies the information of ITD and ILD, and it is even more robust than ITD + ILD. When tested on the CIPIC database, the IMF-based localization is more accurate than the ITD + ILD-based one for both azimuth and elevation in the quiet surroundings, and these results motivate us to use ITD + ILD + IMF features to design our localization system. The weakness of IMF would yet be that noise easily affects its design. When comparing the localization correct rate of our ITD + ILD + IMF with several state-of-the-art methods, the superiority is obviously reflected, especially in the moderately noisy environments, which is mainly benefited from the perfect combination of ITD + ILD and IMF, because IMF can work well in the quiet circumstances and ITD + ILD computed by TDC has noisy robustness to some extend. More importantly, our algorithm has not introduce excess spatial storage, and our practical HRI system can complete an azimuthal localization with 90% accuracy rate approximately within 0.35 s.

Furthermore, this works only take one design of IMF in the experiments since the other scheme leads to the larger error/cost function. Therefore, in the future we will try to adaptively combine the two schemes according to the noisy conditions. We will also pay more attention to the elevation localization, because the experimental solutions show that the elevation localization is much more difficult to determine than azimuth. Exploring a new acoustic feature to represent elevation or modeling based on some off-the-shelf cues maybe make a breakthrough, and doing TDC in frequency sub-bands might overcome the aporia of localizing abnormal sound activities like smashing. In addition, we will try to apply our system to the mobile robots, handheld devices to enhance the quality of communication, speaker tracking applications, etc.

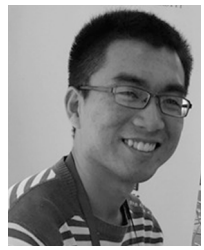
Acknowledgment

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions.

References

[1] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 16, no. 4, pp. 728–739, May 2008.
[2] V. D. Bogaert, T. Doclo, S. Wouters, and J. Moonen, "The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 484–497, 2008.
[3] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE ICASSP*, Apr. 1997, vol. 1, pp. 187–190.
[4] Q. H. Wang, T. Ivanov, and P. Aarabi, "Acoustic robot navigation using distributed microphone arrays," *Inf. Fusion*, vol. 5, no. 2, pp. 131–140, Jun. 2004.
[5] C. T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments," in *Proc. IEEE IROS*, St. Louis, Sep. 2009, pp. 2027–2032.
[6] J. S. Hu, C. Y. Chan, C. K. Wang, and C. C. Wang, "Simultaneous localization of mobile robot and multiple sound sources using microphone array," in *Proc. IEEE ICRA*, Kobe, Japan, 2009, pp. 29–34.
[7] L. A. Jeffress, "A place theory of sound localization," *J. Compar. Physiologic. Psychol.*, vol. 61, pp. 468–486, 1948.
[8] R. F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, pp. 1119–1134, 1988.
[9] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 18, no. 1, pp. 68–77, Jan. 2010.
[10] H. Viste and G. Evangelista, "Binaural source localization," in *Proc. IEEE DAFx*, Naples, Italy, Oct. 2004, pp. 145–150.
[11] S. T. Birchfield and R. Gangishetty, "Acoustic localization by interaural level difference," in *Proc. IEEE ICASSP*, 2005, vol. 4, pp. 1109–1112.
[12] W. Cui, Z. Cao, and J. Wei, "Dual-microphone source location method in 2-D space," in *Proc. IEEE ICASSP*, May 2006, vol. 4, pp. 845–848.
[13] T. May, S. van de Pan, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. on ASLP*, vol. 19, no. 1, pp. 1–13, Jau. 2011.
[14] K. Youssef, S. Argentieri, and J. Zarader, "A binaural sound source localization method using auditive cues and vision," in *Proc. IEEE ICASSP*, 2012, pp. 217–220.
[15] D. Li and S. E. Levinson, "A Bayes-rule based hierarchical system for binaural sound source localization," in *Proc. IEEE ICASSP*, Apr. 2003, vol. 5, pp. 521–524.
[16] H. Finger and P. Ruvolo, "Approaches and databases for online calibration of binaural sound localization for robotic heads," in *Proc. IEEE IROS*, Oct. 2010, pp. 4340–4345.
[17] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, "A probabilistic model for binaural sound localization," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, 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*, Mar. 2012, pp. 1693–1696.
[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*, 2012, pp. 313–316.
[20] X. F. Li and H. Liu, "Sound source localization for HRI using FOC-based time difference feature and spatial grid matching," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 43, no. 4, pp. 1199–1212, Aug. 2013.
[21] M. Y. Wu and D. L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 14, no. 3, pp. 774–784, May 2006.
[22] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, Sep. 2010.
[23] H. Liu and J. Zhang, "A novel binaural sound source localization model based on time-delay compensation and interaural coherence," in *Proc. IEEE ICASSP*, Florence, Italy, May 2014, pp. 1438–1442.

[24] H. Liu, J. Zhang, and Z. Fu, "A new hierarchical binaural sound source localization method based on interaural matching filter," in *Proc. IEEE ICRA*, Hong Kong, China, May 31–Jun. 5, 2014, pp. 1598–1605.
[25] 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*, May 2013, pp. 2690–2697.
[26] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.
[27] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. Workshop IEEE Appl. SPAA*, 2001, pp. 99–102.
[28] P. R. Roth, "Effective measurements using digital signal analysis," *IEEE Spectrum*, vol. 8, pp. 62–70, Apr. 1971.
[29] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4, pp. 411–430, 2000.
[30] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.
[31] S. S. Haykin*, Adaptive Filter Theory*, 4/e[M] ed. : Pearson Education India, 2005.
[32] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Apr. 1979.



Jie Zhang (S'14) was born in Anhui Province, China, in 1990. He received the B.E. degree in electronic information science and technology from Yunnan University, Kunming, China, in 2012. Currently, he is working toward the Master degree at the School of Electronics and Computer Engineering, Shen Zhen Graduate School, Peking University, China.

His current research interests are speech signal processing, speech enhancement, speaker recognition, and sound source localization.

Hong Liu (M'08) received the Ph.D. in mechanical electronics and automation in 1996, and serves as a Full Professor in the School of EE&CS, Peking University (PKU), China. Prof. Liu has been selected as Chinese Innovation Leading Talent supported by "National High-level Talents Special Support Plan" since 2013.

He is also the Director of Open Lab on Human Robot Interaction, PKU, his research fields include computer vision and robotics, image processing, and pattern recognition. Dr. Liu has published more than 150 papers and gained Chinese National Aero-space Award, Wu Wenjun Award on Artificial Intelligence, Excellence Teaching Award, and Candidates of Top Ten Outstanding Professors in PKU. He is an IEEE member, vice president of Chinese Association for Artificial Intelligent (CAAI), and vice chair of Intelligent Robotics Society of CAAI. He has served as keynote speakers, co-chairs, session chairs, or PC members of many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, IEEE SMC and IIHMSP, recently also serves as reviewers for many international journals such as *Pattern Recognition*, IEEE TRANSACTIONS ON SIGNAL PROCESSING, and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.