Binaural Sound Localization Based on Reverberation Weighting and Generalized Parametric Mapping

Pang, Cheng; Liu, Hong; Zhang, Jie; Li, Xiaofei

DOI
10.1109/TASLP.2017.2703650

Publication date
2017

Document Version
Accepted author manuscript

Published in
IEEE - ACM Transactions on Audio, Speech, and Language Processing

Citation (APA)
Pang, C., Liu, H., Zhang, J., & Li, X. (2017). Binaural Sound Localization Based on Reverberation Weighting and Generalized Parametric Mapping. *IEEE - ACM Transactions on Audio, Speech, and Language Processing*, *25*(8), 1618-1632. Article 7926345. https://doi.org/10.1109/TASLP.2017.2703650

Binaural Sound Localization Based on Reverberation Weighting and Generalized Parametric Mapping

Cheng Pang, Hong Liu, Jie Zhang and Xiaofei Li

***Abstract*—Binaural sound source localization is an important technique for speech enhancement, video conferencing and human-robot interaction, etc. However, in realistic scenarios, the reverberation and environmental noise would degrade the precision of sound direction estimation. Therefore, reliable sound localization is essential to practical applications. To deal with these disturbances, this paper presents a novel binaural sound source localization approach based on reverberation weighting and generalized parametric mapping. Firstly, the reverberation weighting as a pre-processing stage, is used to separately suppress the early and late reverberation, while preserving interaural cues. Then, two binaural cues, i.e., interaural time and intensity differences, are extracted from the frequency-domain representations of dereverberated binaural signals for the online localization. Their corresponding templates are established using the training data. Furthermore, the generalized parametric mapping is proposed to build a generalized parametric model for describing relationships between azimuth and binaural cues analytically. Finally, a two-step sound localization process is introduced to refine azimuth estimation based on the generalized parametric model and template matching. Experiments in both simulated and real scenarios validate that the proposed method can achieve better localization performance compared to state-of-the-art methods.**

***Index Terms*—Binaural localization, reverberation weighting, generalized parametric mapping, template matching.**

I. Introduction

BINAURAL sound source localization (SSL) is to determine the spatial direction of a sound source, utilizing the audio recorded by two microphones mounted in the left and right ears. It has wide applications in speech enhancement, speech segregation, hearing aids, human-robot interaction (HRI) and intelligent video conferencing, etc [1]–[4]. Recently, with the advances in array signal processing, SSL has been widely researched. In general, it can be categorized into three classes: 1) techniques based on the high-resolution spectral or beamforming method, e.g., multiple signal classification (MUSIC) [5] and steered response power (SRP) [6]; 2) techniques employing time difference of arrival (TDOA) which is extracted from the cross-correlation function [7]; 3) techniques adopting measured head-related transfer function (HRTF) [8]. Each category has its own advantages and disadvantages.

Binaural SSL based on biological acoustic characteristics has been a prevalent sound localization branch, e.g., hearing aids, humanoid robotics, due to its small-sized sensor array required and easy-equipment. In other words, compared to microphone array based SSL, it is more convenient and friendly. There are three challenging issues concerning binaural SSL: 1) how to accurately localize various types of sound source; 2) how to simultaneously localize several different sound sources; 3) how to track moving sound sources [9]. The first problem mentioned above is considered in this paper.

For binaural SSL, two physical cues consisting of interaural time differences (ITD) and iteraural intensity differences (IID) are frequently used. ITD represents the time-difference of a sound source arriving at two ears, which can be calculated by unwrapping the interaural phase difference (IPD). IID (usually in dB) refers to the intensity difference between binaural signals. In general, the binaural signals are decomposed into perceptual frequency bands, e.g., Gammatone filter bank, or uniform frequency bands, e.g., short-time Fourier transform (STFT), from which the interaural cues are extracted. After "Duplex Theory" [10] and cochlear model [11] were proposed, a large amount of binaural sound source localization systems have been developed based on ITD and IID. For example, Heckmann *et al.* introduced a model of precedence effect to achieve binaural SSL in echoic environments [12]. Youssef *et al.* studied a combination of auditive cues and vision based binaural sound localization in an HRI context [13].

In the presence of noise or reverberation, the performances of most existing approaches degrade significantly [7], [14]–[16]. In realistic environments, the noise generated by surroundings, like air conditioner, is a serious interfering source that cannot be ignored. Furthermore, due to the reverberation, the recorded audio at each ear contains the sound wave coming from the direct path and the sound waves reflected by walls and furniture. Early reflections with different direction-of-arrivals have amplitudes similar to that of direct-path signal, so they lead to negative effects on determination of the true sound source [17]. Besides, high computational complexity is also a limitation to the implementation of real-time sound localization, so the time and storage complexity of the localization methods should be taken into account.

To tackle with above challenges, a number of novel algorithms have been proposed. For instance, Li *et al*. proposed a three-layer binaural SSL system based on the Bayesian rule [18]. Along with the similar hierarchical architecture,

This work is supported by National High Level Talent Special Support Program, National Natural Science Foundation of China (No. 61340046, 61673030, U1613209), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130001110011), Natural Science Foundation of Guangdong Province (No. 2015A030311034).

Cheng Pang and Hong Liu (corresponding author) are with the Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Beijing 100871, China (e-mail: chengpang@sz.pku.edu.cn; hongliu@pku.edu.cn).

Jie Zhang (corresponding author) is with the Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: j.zhang-7@tudelft.nl).

Xiaofei Li is with INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin 38330, France (e-mail: xiaofei.li@inria.fr).

experiments in [19] demonstrated that the hierarchical system can effectively reduce the time consumption of SSL. For acoustic interference, most existing methods focus on the extraction of robust binaural cues, while ignoring their valid combination. Benesty *et al*. provided a multi-channel linear approach to reduce the noise in binaural signals [20]. Willert *et al*. introduced a biologically inspired binaural SSL through extracting binaural cues from cochleagrams generated by a cochlear model [21]. With regard to reverberation, one idea to remove the negative effects by inputting the reverberant signals to a filter that inverts the reverberation process and recovers the original signal. Alternatively, a novel two-stage binaural dereverberation algorithm was proposed by Jeub *et al*., which utilizes a dual-channel Wiener filter to preserve the binaural cues by modelling room impulse response (RIR) [22]. A learning-based approach was also put forward to achieve robust sound localization under reverberant conditions, but its performance is limited by the training conditions [23].

Based on the Fourier analysis for binaural signals, Raspaud *et al.* extracted the ITD and IID in the frequency domain [24]. In order to learn a more comprehensive dependence of ITD and IID on azimuth, probabilistic model, e.g., Gaussian mixture model, is introduced to achieve robust SSL in reverberant environments [25]–[27]. However, the probabilistic model needs to be trained for different signal-to-noise ratios (SNRs) as well as for different reverberation times to keep its environmental adaptivity. Although the previous methods obtain acceptable performance under certain specific conditions, most of them simply focus on estimating the direction of sound source in either noisy or reverberant experimental environments rather than in a realistic environment.

Motivated by the above problems, we propose reverberation weighting and generalized parametric mapping to localize a single sound source in realistic environments including uncorrelated noise and reverberation. Firstly, the reverberation weighting [28] is applied to dereverberate the received binaural signals, which separately suppresses the early and late reverberation meanwhile preserving the direct-path interaural differences. After dereverberation, ITD and IID are extracted from the frequency-domain representations of the dereverberated binaural signals for the online localization. The averaged ITD and IID templates are established using the training data, namely HRTFs. Then, the generalized parametric mapping is proposed to build a generalized parametric model [29] through finding generalized scaling factors. This model describes the mapping relationship from the extracted binaural cues to azimuth estimates. Finally, a two-step sound localization process is used to refine azimuth estimation based on the generalized parametric model and template matching. Rough azimuth estimation is quickly achieved through the generalized parametric model, which is utilised to determine correct phase unwrapping factor to unwrap the online measured ITD. A precise azimuth estimation is obtained by combining the robust estimates of ITD and IID through template matching. Both the normalized distances between the two binaural cues and their averaged templates are calculated, and combined over frequency. The template matching using the averaged ITD and IID templates makes our method achievable under noisy and reverberant conditions. The proposed method is evaluated in a simulated environment based on a public-domain HRTF database. Moreover, the binaural signals collected by an artificial head in a realistic indoor room, are also used to validate the proposed method in practical situations. Note that the methods proposed in this paper are the refined and expanded version of the conference proceedings papers [28] and [29].

Contributions: 1) With regard to the impact of reverberation on SSL, the reverberation weighting is used to suppress the reverberation of binaural signals. Through suppressing the early and late reverberation separately, the proposed reverberation weighting can preserve binaural cues better. Although cepstral prefiltering is used to dereverberate binaural signals for localization [29], [30], it mainly focuses on the dereverberation for time-delay estimation [31].

2) The generalized parametric model is built by the generalized parametric mapping, which is computationally efficient. It can also improve the adaptability of the proposed method to practical applications. The generalized parametric model in [29] is improved by parametric mapping with the optimal generalized scaling factors which are obtained by solving a least squares problem. Although the joint estimation based on ITD and IID has been studied in [24], how to find more generalized scaling factors that are used to describe the mapping relationship from the extracted binaural cues to the objective azimuth, has not been provided. Based on this improved generalized model, the two-step sound source localization is used to refine azimuth estimation. Rough azimuth estimation can be quickly achieved based on the generalized parametric model to unwrap ITD, and a precise azimuth estimation is achieved through the template matching of ITD and IID.

The remainder of this paper is organized as follows. Section II formulates the binaural localization model and presents the reverberation weighting algorithm. Section III details binaural cues extraction and their templates establishment. Section IV presents the generalized parametric mapping and correct phase unwrapping factor estimation. The SSL process based on template matching is detailed in Section V. Experiments and analyses for simulated and realistic environments are shown in Section VI. Finally, Section VII concludes this work.

II. Dereverberation

A. Binaural localization model

In the far field, let $x(n)$ denote the sound signal emitted by a source in the discrete-time domain, the binaural signals received by two ears in the noisy and reverberant environments can be modeled as

$$y_i(n) = h_i(n) \star x(n) + v_i(n),\ i = l, r, \tag{1}$$

where n is the discrete-time index, $\star$ denotes convolution operation, $h_i(n)$ is the impulse response between the source and ears (i.e., *binaural room impulse response*, BRIR), $v_i(n)$ denotes the corresponding interfering term, which is assumed to be an uncorrelated, zero-mean, stationary Gaussian random process, i denotes the microphone index, l and r refer to the left and right microphones, respectively. The impulse response

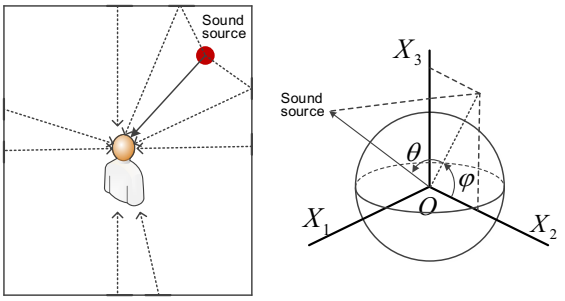


Fig. 1. Binaural localization model. (a) Signal model of binaural sound localization in reverberant environment. (b) The head-center interaural-polar coordinate system. The azimuth θ is the angle between a vector to the sound source and the midsagital plane (i.e., X2-X3 plane), and it varies from -90° to $+90^\circ$. The elevation φ is the angle from the horizontal plane to the projection of the source into the midsagital plane, and it varies from -90° to $+270^\circ$.

$h_i(n)$ involves two independent components, which include the acoustic property of room (i.e., reverberation) and head-related impulse responses (HRIRs). The modeling scheme in the reverberant environments is illustrated in Fig. 1(a). It can be seen that the propagation paths from the sound source to a receiver include a direct path and subsequent reflections. The HRIRs are derived from the training data along with the direct path, and the reflections contain the effect of reverberation. As shown in Fig. 1(b), azimuth θ and elevation φ measured in a head-center interaural-polar coordinate system (which is different with the vertical-polar coordinate system), are used to denote direction of the sound source, which follow the description of the CIPIC HRTF database [32].

In many practical applications, such as HRI, beamforming, the azimuth is more important than the elevation in general, azimuth estimation in complex acoustic environments is therefore the main focus of this work.

B. Reverberation weighting

Many works have been conducted for dereverberation in the past decades, such as inverse filtering [17], spectral subtraction [33], etc. Most of these methods are suitable for the systems with a single output, yet they may break the localization cues. In order to alleviate the contamination of reverberation to the binaural cues for sound localization, the reverberation weighting is proposed to suppress the reverberation.

Since RIR consists of the direct, early and late components [34], $h_i(n)$ can be defined as

$$h_i(n) = \begin{cases} h_i^E(n), & 0 \le n < T_L \cdot f_s \\ h_i^L(n), & T_L \cdot f_s \le n \le T_R \cdot f_s \\ 0, & n < 0 \end{cases} \tag{2}$$

where *E* and *L* stand for the early and late reverberation, respectively, and $h_i^E(n)$ contains the direct and early propagation paths of the sound source, $h_i^L(n)$ represents the late path, T_R refers to the reverberation time [35] and f_s is the sampling frequency. T_L denotes the onset time of late reverberation, which generally ranges from 50 ms to 100 ms [34]. Hence, the received binaural signals in Eq. (1) can be rewritten as

$$y_i(n) = \sum_{k=0}^{T_L f_s - 1} h_i^E(k)x(n-k) + \sum_{k=T_L f_s}^{T_R f_s} h_i^L(k)x(n-k) + v_i(n). \tag{3}$$

Since the two components contained in the RIR affect the sound signal in different ways, they are treated separately in what follows.

Suppressing the late and early reverberation can be done in the frequency domain using the late and early reverberation gains. According to the properties of late and early reverberation, the gains are calculated based on a spectral subtraction rule and the coherence of binaural signals, respectively. Since the spectral weighting has no influence on the coherence, the first step is to suppress the late reverberant components. To do this, the variances of late reverberation can be acquired by a simple statistical model for RIR [36]:

$$\tilde{h}^L(n) = m(n)e^{-\rho n f_s^{-1}}, n \ge 0, \tag{4}$$

where $m(n)$ is a standard Gaussian sequence with zero mean and unit standard deviation, ρ denotes the decay rate, which is determined by the reverberation time T_R:

$$\rho = \frac{3\ln(10)}{T_R}, \tag{5}$$

where T_R can be estimated by Schroeder's method [35]. In Eq. (4), the late reverberant components can be considered as an uncorrelated noise process if the energy of direct path is smaller than all reflections [37].

Here, the received binaural signals are enframed by a Hamming window, and then transformed to the frequency domain through STFT. The variance of the late reverberant signal in the frequency domain can be estimated by an estimator proposed in [22]:

$$\sigma^2_{y_i^L}(\kappa,\omega) = e^{-2\rho T_L} \cdot \sigma^2_{y_i}(\kappa - N_L, \omega), \tag{6}$$

where $\sigma^2_{y_i^L}(\kappa,\omega)$ denotes the variance of the reverberant signal, N_L is the number of frames corresponding to T_L, κ is the frame index, ω denotes the frequency index. Here, the spectral variance of the reverberant speech signal is calculated through recursive averaging:

$$\sigma^2_{y_i}(\kappa,\omega) = \alpha_1 \cdot \sigma^2_{y_i}(\kappa-1,\omega) + (1-\alpha_1)|Y_i(\kappa,\omega)|^2, \tag{7}$$

where α_1 is a smoothing factor, which is set to 0.95, and $Y_i(\kappa,\omega)$ denotes the STFT coefficients of y_i, which obeys the zero-mean complex Gaussian distribution.

Hereby, a posteriori signal-to-interference ratio can be computed by

$$\eta_i^L(\kappa,\omega) = \frac{|Y_i(\kappa,\omega)|^2}{\sigma^2_{y_i^L}(\kappa,\omega)}. \tag{8}$$

Then, the weighting gains used to suppress the late reverberant components contained in the binaural signals are calculated based on the spectral magnitude subtraction rule. They are formulated as

$$G_i^L(\kappa,\omega) = 1 - \frac{1}{\sqrt{\eta_i^L(\kappa,\omega)}}. \tag{9}$$

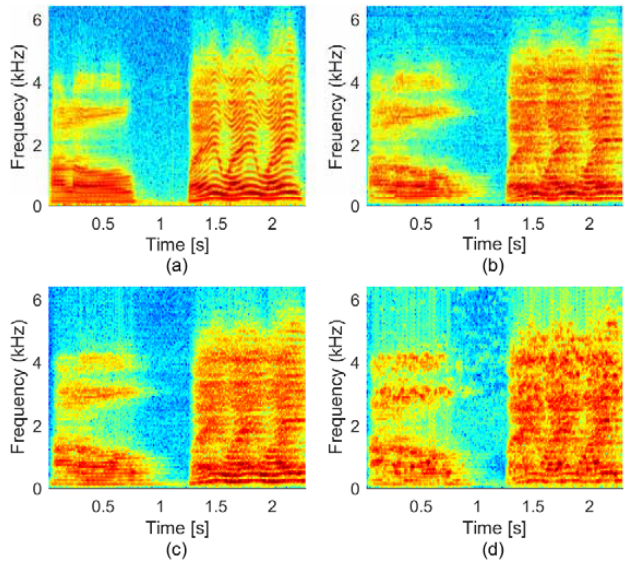


Fig. 2. The effect of reverberation weighting on spectrograms. (a) Original signal. (b) Reverberant signal with reverberation time $T_R = 0.5$ s. (c) Late reverberation-suppressed signal. (d) Reverberation suppressed signal.

The late reverberation-suppressed binaural signals can be obtained by

$$\tilde{Y}_i(\kappa,\omega) = Y_i(\kappa,\omega)\cdot G_i^L(\kappa,\omega). \tag{10}$$

Subsequently, the goal is to suppress the early reverberant components. The coherence-based method is adopted to keep the coherent parts unaffected and to remove all non-coherent signal parts, since the direct-path signal shows a high coherence among binaural signals. The coherence is calculated by

$$\Gamma_{y_l y_r}(\kappa,\omega) = \frac{|\Phi_{y_l y_r}(\kappa,\omega)|}{\sqrt{\Phi_{y_l y_l}(\kappa,\omega)\Phi_{y_r y_r}(\kappa,\omega)}}, \tag{11}$$

where $\Phi_{y_i y_i}(\kappa,\omega)$ and $\Phi_{y_l y_r}(\kappa,\omega)$ refer to the weighted short-term auto- and cross-power spectral densities, respectively, which can be recursively evaluated by

$$\Phi_{y_i y_i}(\kappa,\omega) = \alpha_2 \Phi_{y_i y_i}(\kappa-1,\omega) + |\tilde{Y}_i(\kappa,\omega)|^2, \tag{12}$$

$$\Phi_{y_l y_r}(\kappa,\omega) = \alpha_2 \Phi_{y_l y_r}(\kappa-1,\omega) + \tilde{Y}_l(\kappa,\omega)\tilde{Y}_r^*(\kappa,\omega), \tag{13}$$

where $(\cdot)^*$ denotes the complex conjugate operation and α_2 represents a recursion factor [38], which is determined by

$$\alpha_2 = e^{-\frac{\mathcal{L}}{4tf_s}}, \tag{14}$$

where $\mathcal{L}$ denotes the frame length, and t refers to the time duration for coherence estimation. Similar to [38], t is set to 0.1 s.

Inspired by [39], the obtained coherence is applied to compute the weighting gains for suppressing the early reverberation, which is formulated as

$$G_i^E(\kappa,\omega) = \frac{\mathrm{Re}\{\Phi_{y_l y_r}(\kappa,\omega)\} - \Gamma_{y_l y_r}(\kappa,\omega)\Phi_{y_i y_i}(\kappa,\omega)}{\Phi_{y_i y_i}(\kappa,\omega)\big(1-\Gamma_{y_l y_r}(\kappa,\omega)\big)}, \tag{15}$$

where $\mathrm{Re}\{\cdot\}$ returns the real part of its argument.

Applying the early reverberant gains to $\tilde{Y}_i(\kappa,\omega)$, the dereverberated output signals can be obtained:

$$\hat{Y}_i(\kappa,\omega) = \tilde{Y}_i(\kappa,\omega)\cdot G_i^E(\kappa,\omega). \tag{16}$$

An illustration of dereverberation results in a spectrogram sense is shown in Fig. 2. It can be seen that the speech spectrum of reverberant speech ($T_R = 0.5$ s) becomes fuzzy due to the influence of reverberation compared to the original speech. This phenomenon leads to inconsistent extraction of binaural cues for localization. After reverberation weighting, the dereverberated speech spectrogram becomes clearer and the reverberation tail in the spectrogram is obviously shortened. Meanwhile, it has enhanced peaks that cause a little distortion, yet the acoustic quality of the dereverberated signal is acceptable by listening test.

III. Estimating Localization Cues

A. Online binaural cues extraction

In this part, two frequency-dependent acoustic cues are computed from binaural recordings after reverberation weighting, namely ITD and IID [40]. The two physical localization cues used in this paper are extracted based on the sliding STFT spectra of the dereverberated binaural signals. Hereinafter, the online binaural cues extraction and sound localization are implemented based on each individual frame. For each frame of the binaural signals, the IID (in dB) can be calculated by

$$\Delta I(\omega) = 20\log_{10}\Big|\frac{Y_r(\omega)}{Y_l(\omega)}\Big|, \tag{17}$$

where $Y_l(\omega)$ and $Y_r(\omega)$ are the STFTs of left and right channel of the binaural signals, respectively. When one or both of the $|Y_i(\omega)|$ is null, the interaural differences are regarded as invalid, so that the information of this frame is disregarded.

With the frequency spectra of binaural signals, the ITD is extracted by

$$\Delta T_p(\omega) = \frac{1}{\omega}\Big(\angle\frac{Y_r(\omega)}{Y_l(\omega)} + 2\pi p\Big), \tag{18}$$

where p denotes the phase unwrapping factor, which is a *priori* integer. Since the angle corresponding to the spectral ratio is calculated modulo 2π, the factor p is necessary for the correct ITD estimation. However, p makes IPD become ambiguous above a certain frequency, which mainly depends on the size and shape of listener's head. As the average radius of human head is about 0.07 m, its corresponding ITD range is about [-0.5, 0.5] ms. The binaural signals are analyzed over the frequency range of 0-8 kHz, so the range of p is [-7, 7]. For each possible p in this range, the corresponding $\Delta T_p(\omega)$ is calculated to estimate azimuth for determining the following correct phase unwrapping factor in the online localization stage. In addition, the parameter p indexes sound positions. A negative p corresponds to a position on the left side ($\theta < 0$) and positive p indicates the position on the right side. In this case, possible values of p depend on the physical layout of microphone sensors and sources. Note that Eq. (17) and Eq. (18) are used to extract the binaural cues from the binaural recording for subsequent online localization.

B. Offline templates establishment

In order to retrieve the azimuth from an STFT pair for a given frequency bin, the IID and ITD estimates of that

bin are required to be matched with the estimated IID and ITD from the HRTFs of all subjects, respectively. It is assumed that the HRTFs are time-invariant, and they are only dependent on the azimuthal angle θ. In this work, the templates consisting of the intensity difference $\Delta I_s^{\mathcal{T}}(\theta,\omega)$ and the time difference $\Delta T_s^{\mathcal{T}}(\theta,\omega)$ for each individual subject *s*, need to be established before the online localization. The templates establishment can be accomplished offline. Similar to the online IID extraction in Eq. (17), IID templates can be established by

$$\Delta I_s^{\mathcal{T}}(\theta,\omega) = 20\log_{10}\Big|\frac{\mathrm{HRTF}_{\mathrm{r}}^{\mathrm{s}}(\theta,\omega)}{\mathrm{HRTF}_{\mathrm{l}}^{\mathrm{s}}(\theta,\omega)}\Big|, \tag{19}$$

where $\mathcal{T}$ stands for the template, $\mathrm{HRTF}_l^s(\theta,\omega)$ and $\mathrm{HRTF}_r^s(\theta,\omega)$, respectively, represent the frequency-domain HRTFs on the left and right ears for azimuth θ and subject *s*. Similarly, ITD templates based on HRTF can be built as

$$\Delta T_{s,p}^{\mathcal{T}}(\theta,\omega) = \frac{1}{\omega}(\angle\frac{\mathrm{HRTF}_r^s(\theta,\omega)}{\mathrm{HRTF}_l^s(\theta,\omega)} + 2\pi p). \tag{20}$$

In Eq. (20), the ITD also depends on the phase unwrapping factor. In the offline training stage, the exact position of a calibration source related to the head is known in advance, such that the theoretical ITD can be calculated, with which we can choose the correct *p* of the ITD templates for Eq. (20). The ambiguity is eliminated through unwrapping modulo 2π for the phase differences of the HRTFs across frequencies. The actual phase differences of the HRTFs are assumed to be a continuous function versus frequency. Besides, *p* is supposed to be 0 when $\theta = 0°$, where the phase difference ought to be pretty small. Since the correct phase unwrapping factor $\hat{p}$ can be determined in this case, the unwrapped ITD templates can be obtained as $T_{s,\hat{p}}^{\mathcal{T}}(\theta,\omega)$, which is simplified to $\Delta T_s^{\mathcal{T}}(\theta,\omega)$ in the context. Note that Eq. (17) and Eq. (18) are used for online localization based on the binaural audio, while Eq. (19) and Eq. (20) are applied for offline training templates using HRTFs. Taking the CIPIC database [32] as an example, the ITDs and IIDs for one specific head are shown in Fig. 3(a) and (b) in terms of the azimuth and angular frequency, respectively.

Let N_s denote the total number of subjects, the averaged ITD and IID templates over all the subjects can be obtained by

$$\Delta\bar{I}^{\mathcal{T}}(\theta,\omega) = \frac{1}{N_s}\sum_{s=1}^{N_s}\Delta I_s^{\mathcal{T}}(\theta,\omega), \tag{21}$$

$$\Delta\bar{T}^{\mathcal{T}}(\theta,\omega) = \frac{1}{N_s}\sum_{s=1}^{N_s}\Delta T_s^{\mathcal{T}}(\theta,\omega). \tag{22}$$

The obtained $\Delta\bar{I}^{\mathcal{T}}(\theta,\omega)$ and $\Delta\bar{T}^{\mathcal{T}}(\theta,\omega)$ are shown in Fig. 3(c) and (d). It can be seen that the smoothed $\Delta\bar{T}^{\mathcal{T}}(\theta,\omega)$ (or $\Delta\bar{I}^{\mathcal{T}}(\theta,\omega)$) and $\Delta T_s^{\mathcal{T}}(\theta,\omega)$ (or $\Delta I_s^{\mathcal{T}}(\theta,\omega)$) are in the same range and follow similar changing trend. Hence, the smoothed versions of binaural cues can be representative of an individual subject, and they are thus used for the online localization process. From Fig. 3(a) and (c), it can also be concluded that, for a specific angular frequency, ITDs are significantly influenced by the azimuth, and for a specific azimuth, the ITDs almost stay invariant in terms of the frequency. However, the IIDs in Fig. 3(b) or (d) vary with both the azimuth and frequency. This phenomenon explicitly reveals that the IID measurement can be employed to resolve the phase unwrapping factor, which will be detailed in Section IV.

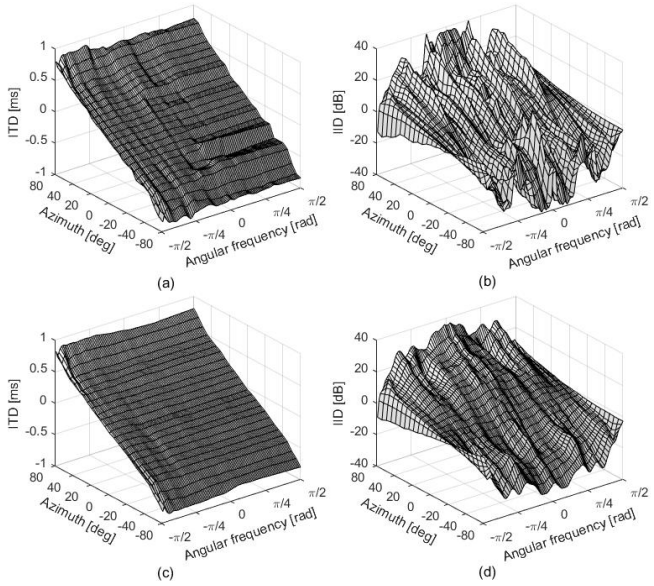


Fig. 3. Binaural cues distribution versus azimuth and angular frequency in the CIPIC HRTF database. (a) ITD and (b) IID for subject #21. (c) Averaged ITD and (d) averaged IID over all subjects.

IV. Generalized parametric mapping and correct phase unwrapping factor estimation

A. Generalized parametric mapping

In Section III, we present how to extract the binaural cues and how to establish the ITD and IID templates. In this section, the generalized parametric mapping is proposed to describe the mapping relationship between binaural cues and azimuths by finding more generalized scaling factors, which builds a generalized parametric model to guide templates lookup for azimuth estimation. This generalized parametric mapping improves the robustness and adaptability of the sound localization system when it is applied to different artificial heads.

As a typical human head has a nearly spherical and uniform surface, the ITD produced by a sound source at azimuth θ can be approximately defined based on the diffraction theory [41] (has also been validated in Fig. 3). However, the formulation between the IID and azimuth does not obey a monotonic function but a more complex function as a typical example depicted in Fig. 3. Since the scale of IID is proportional to the sine of the azimuth [40], an approximation is made to simplify analysis but without loss of generality. Here, for azimuth θ, the corresponding ITD and IID are related to $\Delta\tau(\theta)$ and $\Delta\varepsilon(\theta)$, respectively, which can be approximatively defined as

$$\begin{aligned}\Delta\tau(\theta) &= \gamma\frac{\theta+\sin\theta}{c},\\ \Delta\varepsilon(\theta) &= \sin\theta,\end{aligned} \tag{23}$$

where c denotes the propagation speed of the sound signal in the air (set to 344 m/s), γ represents the "head radius" and its value is set to be the mean head radius in the CIPIC database,

namely 0.07 m. Based on Eq. (23), the parametric ITD and IID model for subject *s* can be formulated as

$$\begin{aligned}\Delta T_s^{\mathcal{P}}(\theta,\omega) &= \alpha_s(\omega)\Delta\tau(\theta),\\ \Delta I_s^{\mathcal{P}}(\theta,\omega) &= \beta_s(\omega)\Delta\varepsilon(\theta),\end{aligned} \tag{24}$$

where $\mathcal{P}$ stands for "parametric", $\alpha_s(\omega)$ and $\beta_s(\omega)$ are the scaling factors for the ITD and IID model, respectively. The two scaling factors depend on the frequency and the subject, which are applicable for all the values of θ. These scaling factors are introduced to give the closest match to the measured ITD and IID templates in the training stage (see Section III-B). Since human head cannot perfectly conform to the head used for modeling Eq. (23) in realistic scenarios, $\Delta\tau(\theta)$ and $\Delta\varepsilon(\theta)$ may only reflect the global changing trend for different θ.

Let N_a denote the number of azimuth, such that $\theta_j, j = 1, 2, \cdots, N_a$. With the $\Delta\tau(\theta)$ and $\Delta\varepsilon(\theta)$ in Eq. (23), we can separately define their vectors as

$$\begin{aligned}\Delta\boldsymbol{\tau} &= [\Delta\tau(\theta_1), \Delta\tau(\theta_2), \cdots, \Delta\tau(\theta_{N_a})]^T,\\ \Delta\boldsymbol{\varepsilon} &= [\Delta\varepsilon(\theta_1), \Delta\varepsilon(\theta_2), \cdots, \Delta\varepsilon(\theta_{N_a})]^T.\end{aligned}$$

With regard to the ITD and IID templates established in the training stage, we can define the vectors of ITD and IID templates for each subject *s* as

$$\begin{aligned}\Delta\boldsymbol{T}_s^{\mathcal{T}}(\omega) &= \left[\Delta T_s^{\mathcal{T}}(\theta_1,\omega), \Delta T_s^{\mathcal{T}}(\theta_2,\omega), \cdots, \Delta T_s^{\mathcal{T}}(\theta_{N_a},\omega)\right]^T,\\ \Delta\boldsymbol{I}_s^{\mathcal{T}}(\omega) &= \left[\Delta I_s^{\mathcal{T}}(\theta_1,\omega), \Delta I_s^{\mathcal{T}}(\theta_2,\omega), \cdots, \Delta I_s^{\mathcal{T}}(\theta_{N_a},\omega)\right]^T.\end{aligned}$$

In order to make the parametric ITD and IID give the closest match to the measured ITD and IID templates, the optimal scaling factors for the parametric ITD and IID model can be found by solving the following optimization problems:

$$\underset{\alpha_s(\omega)}{\text{minimize}} \quad \left\|\alpha_s(\omega)\Delta\boldsymbol{\tau} - \Delta\boldsymbol{T}_s^{\mathcal{T}}(\omega)\right\|_2^2, \tag{25}$$

$$\underset{\beta_s(\omega)}{\text{minimize}} \quad \left\|\beta_s(\omega)\Delta\boldsymbol{\varepsilon} - \Delta\boldsymbol{I}_s^{\mathcal{T}}(\omega)\right\|_2^2. \tag{26}$$

Through minimizing the objective functions in Eq. (25) and Eq. (26) (which respectively refer to the square of ℓ_2-norm for the difference between $\alpha_s(\omega)\Delta\boldsymbol{\tau}$ and $\Delta\boldsymbol{T}_s^{\mathcal{T}}(\omega)$ and the difference between $\beta_s(\omega)\Delta\boldsymbol{\varepsilon}$ and $\Delta\boldsymbol{I}_s^{\mathcal{T}}(\omega)$), the optimal scaling factors $\alpha_s(\omega)$, $\beta_s(\omega)$ for the ITD and IID model can be obtained for each subject.

With regard to Eq. (25) and Eq. (26), least squares method is adopted to calculate the optimal scaling factors. The optimal scaling factors $\alpha_s(\omega)$ and $\beta_s(\omega)$ for each subject are given by

$$\begin{aligned}\alpha_s(\omega) &= \frac{c\sum_{j=1}^{N_a}(\theta_j + \sin\theta_j)\cdot\Delta T_s^{\mathcal{T}}(\theta_j,\omega)}{\gamma\sum_{j=1}^{N_a}(\theta_j+\sin\theta_j)^2},\\ \beta_s(\omega) &= \frac{\sum_{j=1}^{N_a}\Delta I_s^{\mathcal{T}}(\theta_j,\omega)\cdot\sin\theta_j}{\sum_{j=1}^{N_a}\sin^2\theta_j}.\end{aligned} \tag{27}$$

However, using the $\alpha_s(\omega)$ and $\beta_s(\omega)$ from each specific subject for sound source localization is complex in practical situations. Fortunately, averaged parameters over the on-the-shelf subjects in the HRTF datasets are enough for an expected azimuth localization accuracy. Then, the generalized scaling factors $\alpha_g(\omega)$ and $\beta_g(\omega)$ are obtained by calculating the expected (averaged) values of $\alpha_s(\omega)$ and $\beta_s(\omega)$ over all the subjects, which are utilized to build a generalized parametric model.

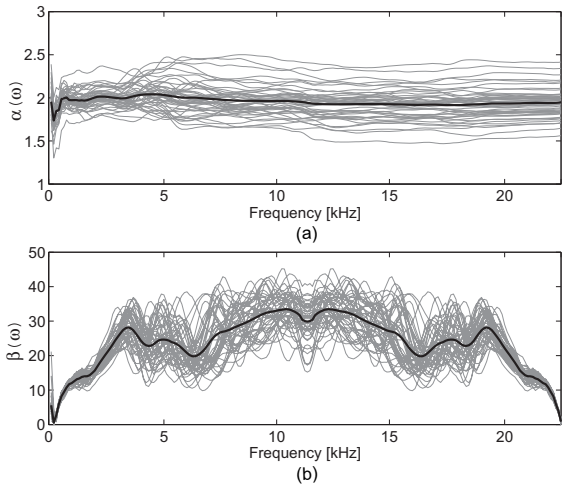


Fig. 4. The changing trends of the scaling factors for ITD (a) and IID (b) model in terms of frequency and subject.

Fig. 4 shows the changing trends of the scaling factors for the ITD and IID model in terms of frequency and subject, respectively. In Fig. 4, the grey curves represent the scaling factors of different subjects, and the two black solid curves denote the averaged results. It can be seen that the $\alpha_s(\omega)$ and $\beta_s(\omega)$ for each subject follow the similar changing trend in terms of frequency. Therefore, the generalized scaling factors calculated by least squares method can represent their dependence on subjects well. Here, the obtained generalized scaling factors are applied to the parametric ITD and IID model in Eq. (24), which can improve the adaptability of our method for different artificial heads. The generalized parametric model is then formulated as

$$\begin{aligned}\Delta T^{\mathcal{P}}(\theta,\omega) &= \alpha_g(\omega)\gamma\frac{\theta+\sin\theta}{c},\\ \Delta I^{\mathcal{P}}(\theta,\omega) &= \beta_g(\omega)\sin\theta.\end{aligned} \tag{28}$$

B. Correct phase unwrapping factor estimation

With the online measured ITD and IID in Section III-A, their corresponding azimuth can be estimated based on the generalized parametric model formulated in Eq. (28). In this section, the ITD-based and IID-based azimuth estimations in [24] are used to find the correct phase unwrapping factor. Firstly, the above generalized parametric model is used to estimate azimuths based on the online measured ITD and IID. Then, the obtained ITD-based and IID-based azimuth estimates are combined to find the correct phase unwrapping factor, which is used to unwrap the measured ITD. In order to retrieve the azimuth from the binaural cues, it is necessary to inverse the generalized parametric model. In detail, the

parametric azimuth estimates from the online measured ITD $\Delta T_p(\omega)$ and IID $\Delta I(\omega)$ can be respectively computed by

$$\hat{\theta}^{\mathcal{P}}_{T,p}(\omega) = f^{-1}\Big(\frac{c}{\gamma\alpha_g(\omega)}\Delta T_p(\omega)\Big), \tag{29}$$

$$\hat{\theta}^{\mathcal{P}}_{I}(\omega) = \arcsin\frac{\Delta I(\omega)}{\beta_g(\omega)}, \tag{30}$$

where $f^{-1}(\cdot)$ is the inverse function of $f(\theta) = \theta + \sin\theta$. Since $f^{-1}(\cdot)$ is difficult to be calculated directly, a polynomial approximation of $f^{-1}(\cdot)$ over the interval of interest is achieved by a Chebyshev series, which is formulated as

$$\hat{f}^{-1}(z) = \frac{z}{2} + \frac{z^3}{96} + \frac{z^5}{1280}. \tag{31}$$

Here, $\hat{\theta}^{\mathcal{P}}_{T,p}(\omega)$ and $\hat{\theta}^{\mathcal{P}}_{I}(\omega)$ are rough azimuth estimates, because the $\Delta\tau(\theta)$ and $\Delta\varepsilon(\theta)$ in Eq. (23) are defined with an approximation model of human head. However, the azimuth estimations based on Eq. (29) and Eq. (30) are computationally efficient for the direct mapping between binaural cues and azimuth, so they are only used to determine the correct phase unwrapping factor.

To this end, we can obtain some rough azimuth estimates using the ITD and IID computed from the dereverberated signals (in Section II). The candidate estimate computed by Eq. (30) is unique, while multiple solutions are resolved from Eq. (29) due to different phase unwrapping factor p. Hence, in order to make use of the accurate ITD information for sound localization, the correct phase unwrapping factor $\hat{p}$ need to be determined first. It is solved by matching the azimuth estimates from ITD and IID:

$$\hat{p} = \underset{p}{\operatorname{argmin}}\, |\hat{\theta}^{\mathcal{P}}_{T,p}(\omega) - \hat{\theta}^{\mathcal{P}}_{I}(\omega)|. \tag{32}$$

Based on the correct phase unwrapping factor $\hat{p}$ selected by Eq. (32) and the measured ITD $\Delta T_p(\omega)$, the unwrapped ITD $\Delta T_{\hat{p}}(\omega)$ are prepared for the following localization process.

V. Sound Source Localization

To help illustrate the proposed localization algorithm, some azimuth estimates based on ITD, IID and their combination are presented in Fig. 5. In this figure, two-dimensional histograms as functions of azimuth and frequency are used to describe the results, where the color represents the normalized distance of a candidate azimuth (i.e., the darker, the more likely). We illustrate four different azimuthal cases, i.e., $-30°$, $0°$, $30°$ and $65°$. The first row of panels is only based on ITDs through Eq. (29). It can be seen that the ITD-based azimuth estimation becomes more and more ambiguous with increasing frequency. The second row of panels shows similar histograms only based on IIDs through Eq. (30). It can be seen that although there is no ambiguity for this case, it has a larger standard deviation than the azimuth estimates based on ITD, especially at low frequencies. It can be concluded that the ITD-based estimates are more ambiguous at high frequencies and IID-based localization has a larger standard deviation at low frequencies.

Although both ITD and IID are functions of azimuth, they can also be related to each other. A joint evaluation of these quantities is proposed in [24] to improve the azimuth estimation. The IID-based estimation is just used to correct ITD extraction, and the final estimation is decided by the corrected ITD. The third row of panels in Fig. 5 shows the results based on the joint estimation of ITD and IID through Eqs. (32, 29), and it is much better than the ITD-based or IID-based case. As the definitions of $\Delta\tau(\theta)$ and $\Delta\varepsilon(\theta)$ in Eq. (23) are not absolutely accurate but computationally efficient, the above rough azimuth estimates from ITD and IID are used to determine the correct phase unwrapping factor. Since the azimuth estimates from ITD and IID are complementary over frequency, both of their estimates should be considered for the final azimuth determination. These motivate us to adopt the template matching to combine both ITD and IID for a precise azimuth estimation.

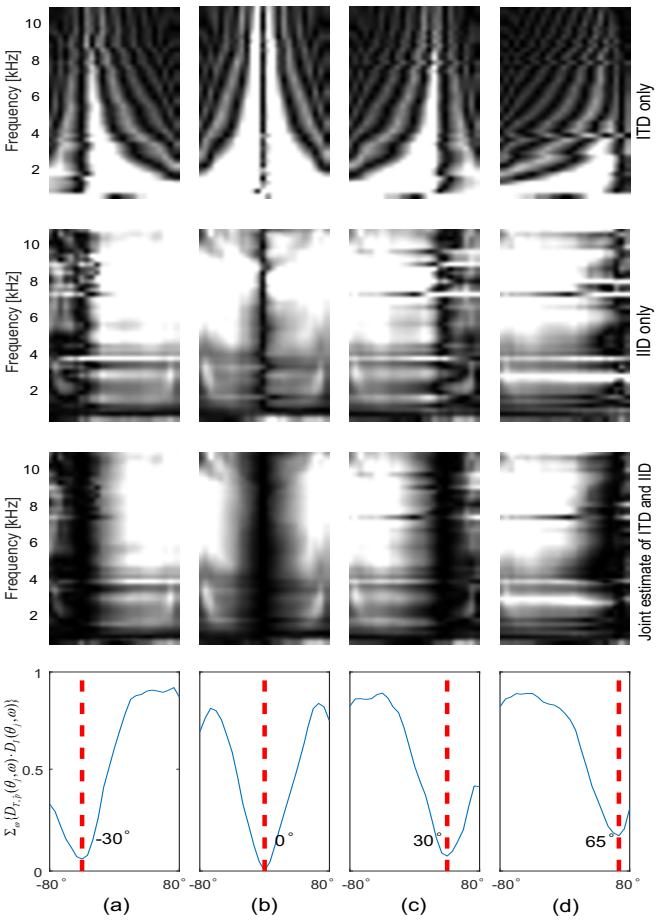


Fig. 5. Histogram of azimuth estimates for four different azimuth angles, i.e., $-30°$, $0°$, $30°$ and $65°$, (a)-(d), respectively. First row: based on ITD only. Second row: based on IID only. Third row: based on the joint estimate of ITD and IID. Bottom row: normalized marginal distance distributions of the localized azimuths.

Template matching is introduced to take raw estimates from both ITD and IID into account for a precise azimuth estimation, which is used to improve the localization accuracy of the SSL system for noisy and reverberant environments. More specifically, the normalized distances between the unwrapped ITD and the averaged ITD templates are computed. Similarly, the normalized distances between the measured IID and the averaged IID templates are also computed. Then, the obtained normalized distances from ITD and IID are combined over

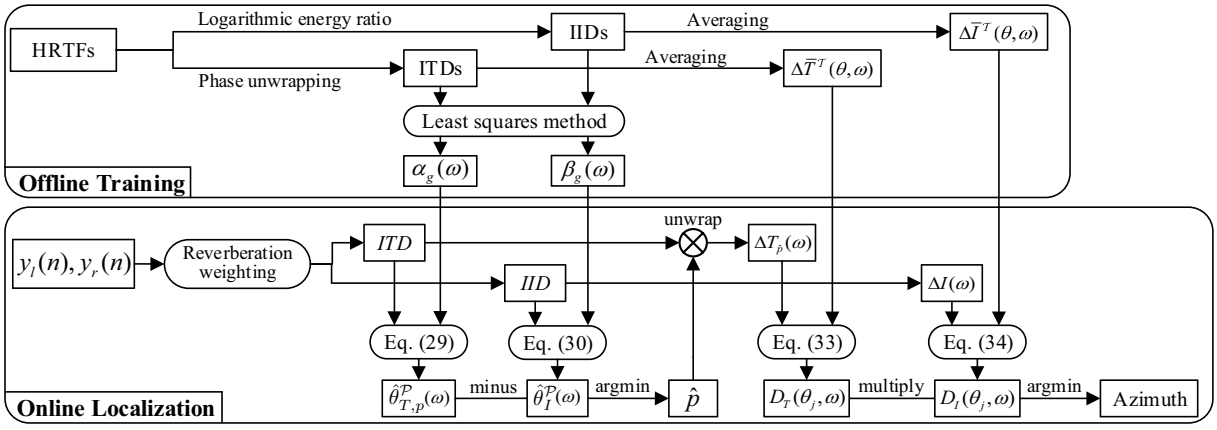


Fig. 6. Flowchart of the proposed method. The upper part is an offline process generating templates to train generalized parametric model. The lower online localization estimates azimuth through reverberation weighting, generalized parametric mapping and template matching. Here, rectangle box denotes a value and rounded rectangle box represents a function.

frequency for a precise azimuth estimation (similar thoughts can also be found in [18], [25], etc).

In detail, the normalized distance between the unwrapped ITD $\Delta T_{\hat{p}}(\omega)$ and the averaged ITD templates $\Delta\bar{T}^{\mathcal{T}}(\theta_j,\omega)$ for each azimuth θ_j can be calculated by

$$D_T(\theta_j,\omega)=\frac{|\Delta T_{\hat{p}}(\omega)-\Delta\bar{T}^{\mathcal{T}}(\theta_j,\omega)|}{\max\limits_{\theta_j}(|\Delta T_{\hat{p}}(\omega)-\Delta\bar{T}^{\mathcal{T}}(\theta_j,\omega)|)}. \tag{33}$$

Similarly, the normalized distance between the measured IID $\Delta I(\omega)$ and the averaged IID templates $\Delta\bar{I}^{\mathcal{T}}(\theta_j,\omega)$ for each azimuth θ_j is obtained by

$$D_I(\theta_j,\omega)=\frac{|\Delta I(\omega)-\Delta\bar{I}^{\mathcal{T}}(\theta_j,\omega)|}{\max\limits_{\theta_j}\left(|\Delta I(\omega)-\Delta\bar{I}^{\mathcal{T}}(\theta_j,\omega)|\right)}. \tag{34}$$

Since ITD is robust at low frequencies and IID is reliable at high frequencies, they are complementary to each other for azimuth estimation. Therefore, the normalized distances of ITD and IID are combined at each frequency to overcome the unreliable estimates. Specifically, $D_T(\theta_j,\omega)$ and $D_I(\theta_j,\omega)$ are multiplied and summed over frequency to obtain hybrid normalized distances. In this way, the hybrid distances become smaller at the ground-truth direction. This kind of combination can sharpen the curve of hybrid distances at the ground-truth azimuth (as shown in the fourth row of panels in Fig. 5), such that the source direction is more recognizable. Finally, the precise azimuth estimate $\hat{\theta}$ is found by minimizing the hybrid distances:

$$\hat{\theta}=\underset{\theta_j}{\operatorname{argmin}}\sum\nolimits_{\omega}\{D_T(\theta_j,\omega)\cdot D_I(\theta_j,\omega)\}. \tag{35}$$

In Fig. 5, the fourth row of panels shows the corresponding distance distributions for the four different azimuths, it can be seen that the correct estimation is achieved for all the cases.

The flowchart of our method is illustrated in Fig. V, which includes two modules, i.e., offline training and online localization. In the offline training process, the generalized scaling factors and the averaged ITD and IID templates are extracted from training data. With regard to the online localization process, binaural signals are firstly processed by reverberation weighting. The generalized parametric model and template matching are then used to achieve and refine azimuth estimation. The online localization process is briefly summarized in Algorithm 1.

VI. Experiments and Analyses

This section evaluates the effectiveness of the proposed method under complex acoustic conditions. Firstly, simulated experimental scenario and setup are shown in Section VI-A. Then, the evaluations of our method in simulated environments are presented in Section VI-B. Finally, Section VI-C evaluates the localization performance of our method in a realistic indoor environment.

A. Experimental setup

The CIPIC HRTF database [32] used in this paper is collected by the U. C. Davis CIPIC Interface Laboratory. It contains HRTFs for 45 different subjects, which include 27 males, 16 females, and KEMAR with large and small pinnae. The HRTFs are measured at source-to-sensors distance of 1 m with 25 different azimuths and 50 different elevations, such that 1250 directions for each subject are considered in total.

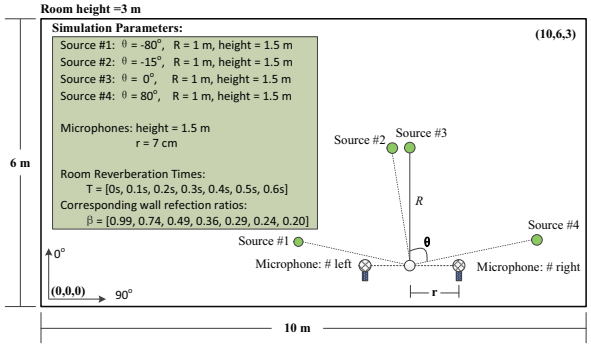


Fig. 7. Simulation scene and parameters of experimental environments. The average radius of heads in the CIPIC datasets is 7 cm approximately.

An enclosure of (10 m × 6 m × 3 m) is simulated using the Roomsim toolbox [42] based on image method [43]. The head is placed at the position (6, 2, 1.5) m. A Chinese pop musical signal sampled at 44.1 kHz, which consists of human

Algorithm 1: *Binaural sound source localization*

Input: $y_i(n), i = l, r$

Output: Estimated azimuth $\hat{\theta}$

Templates: ITDs, IIDs, scaling factors

Apply STFT to $y_i(n)$ using Hamming window;

Compute the late reverberation weighting gains $G_i^L(\kappa,\omega)$ based on spectral subtraction rule;

Suppress the late reverberation $\tilde{Y}_i(\kappa,\omega) = Y_i(\kappa,\omega)\cdot G_i^L(\kappa,\omega)$;

Compute the early reverberation weighting gains $G_i^E(\kappa,\omega)$ based on the coherence of binaural signals ;

Suppress the early reverberation $\hat{Y}_i(\kappa,\omega) = \tilde{Y}_i(\kappa,\omega)\cdot G_i^E(\kappa,\omega)$;

Extract binaural cues $\Delta T_p(\omega)$ and $\Delta I(\omega)$ from the dereverberated binaural signals using Eq. (17), Eq. (18);

Estimate crude ITD-based azimuth $\hat{\theta}_{T,p}^{\mathcal{P}}$: $\hat{\theta}_{T,p}^{\mathcal{P}}(\omega) = f^{-1}\Big(\frac{c}{\gamma\alpha_g(\omega)}\Delta T_p(\omega)\Big)$;

Compute crude IID-based azimuth $\hat{\theta}_I^{\mathcal{P}}$: $\hat{\theta}_I^{\mathcal{P}}(\omega) = \arcsin\frac{\Delta I(\omega)}{\beta_g(\omega)}$;

Determine the correct phase unwrapping factor $\hat{p}$: $\hat{p} = \underset{p}{\operatorname{argmin}}\,|\hat{\theta}_{T,p}^{\mathcal{P}}(\omega) - \hat{\theta}_I^{\mathcal{P}}(\omega)|$;

Determine $\Delta T_{\hat{p}}(\omega)$ by unwrapping the online measured ITD $\Delta T_p(\omega)$ with $\hat{p}$;

Compute the normalized distance based on the unwrapped ITD: $D_T(\theta_j,\omega) = \frac{|\Delta T_{\hat{p}}(\omega) - \Delta\bar{T}^{\mathcal{T}}(\theta_j,\omega)|}{\max\limits_{\theta_j}(|\Delta T_{\hat{p}}(\omega) - \Delta\bar{T}^{\mathcal{T}}(\theta_j,\omega)|)}$;

Compute the normalized distance based on the measured IID: $D_I(\theta_j,\omega) = \frac{|\Delta I(\omega) - \Delta\bar{I}^{\mathcal{T}}(\theta_j,\omega)|}{\max\limits_{\theta_j}\big(|\Delta I(\omega) - \Delta\bar{I}^{\mathcal{T}}(\theta_j,\omega)|\big)}$;

Precise azimuth estimation: $\hat{\theta} = \underset{\theta_j}{\operatorname{argmin}}\sum_\omega\{D_T(\theta_j,\omega)\cdot D_I(\theta_j,\omega)\}$;

return $\hat{\theta}$

voice and instrumental activities that always exist, is utilized as the sound source signal, and it has no silence. Since the major focus of this paper is azimuth estimation, the elevation angle is set to 0 degree and the sound source is positioned at variable horizontal angles with respect to the head. The subject #21 (i.e., Kemar head) in the CIPIC HRTF database is used for evaluation. The simulation scene and detailed parameters are illustrated in Fig. 7. The source can be seen in the far field with source-to-sensors distance of 1 m. The additive diffuse noise is white Gaussian noise. The binaural signals are enframed by a Hamming window of 256 samples with a frame shift of 128 samples. Note that the values of the online extracted ITD and IID are limited in the ranges [-1, 1] ms and [-40, 40] dB, respectively. If the extracted ITD or IID is beyond these ranges, the current binaural cues are regarded as invalid and disregarded.

B. Experiments in simulated environments

1) Evaluation of reverberation weighting: First of all, we investigate the effect of the reverberation weighting on sound localization performance. In this part, different reverberation times varying from 0 s to 0.6 s at an interval of 0.1 s are considered for the simulated environment.

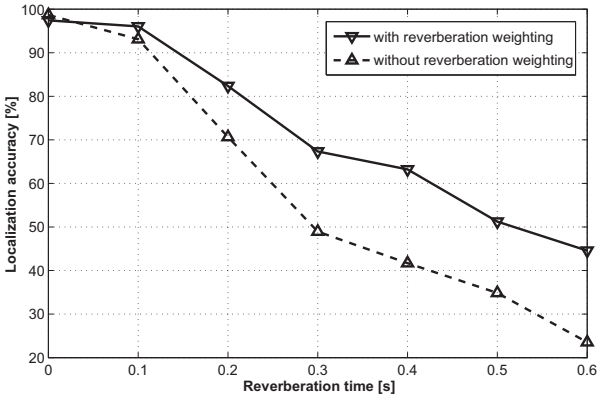


Fig. 8. Comparison of localization accuracies between the method with and without reverberation weighting where tolerance = 0°.

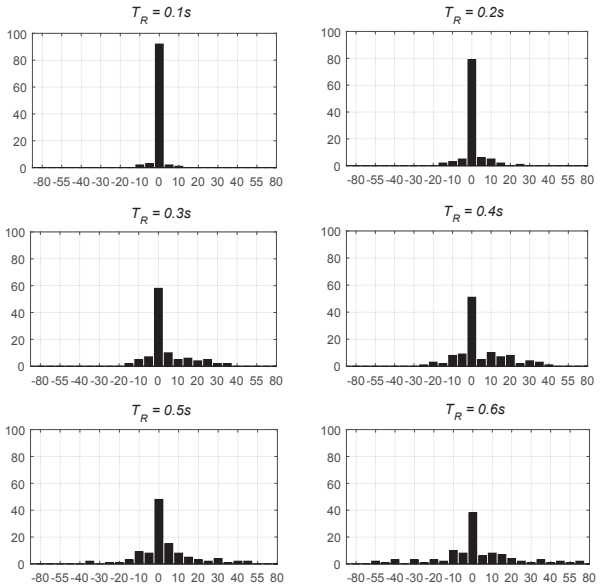


Fig. 9. Histograms of the localization results where azimuth $\theta = 0°$ and reverberation time ranges from $T_R = 0.1$ s to $T_R = 0.6$ s.

The azimuth estimation results with and without reverberation weighting at different reverberation times are shown in Fig. 8. The results are averaged over 100 trails with a tolerance of 0°. It can be seen that the reverberation weighting improves localization accuracy, particularly in strong reverberant environments, e.g., the cases when $T_R \geq 0.3$ s. Nevertheless, the result without reverberation weighting is a little better when $T_R = 0$ s, because reverberation weighting would bring a little distortion to the original binaural signals. More detailed localization results can be found in Fig. 9, where the sound source is fixed at $\theta = 0°$. The six subplots illustrate the results at different reverberation times. It can be observed that though the reverberation time increases to 0.6 s, the proposed method achieves an accuracy of nearly 40%, and the false results are judged as the directions around $0°$. Therefore, the reverberation weighting is helpful to sound localization in reverberant environments.

TABLE I
THE LOCALIZATION ACCURACIES OF AZIMUTH θ AT DIFFERENT REVERBERATION TIMES

Reverberation time	0.1 s			0.3 s			0.5 s		
Tolerance	0°	5°	10°	0°	5°	10°	0°	5°	10°
our method	$\mathbf{91.03\%}$	$\mathbf{93.41\%}$	$\mathbf{97.26\%}$	$\mathbf{46.84\%}$	$\mathbf{54.16\%}$	$\mathbf{71.04\%}$	$\mathbf{29.47\%}$	$\mathbf{37.26\%}$	$\mathbf{46.86\%}$
Parisi et al. [30]	85.38%	90.36%	95.39%	38.93%	46.24%	59.34%	22.26%	27.85%	38.59%
Raspaud et al. [24]	83.27%	86.51%	90.12%	29.54%	35.65%	42.19%	15.73%	19.31%	25.87%

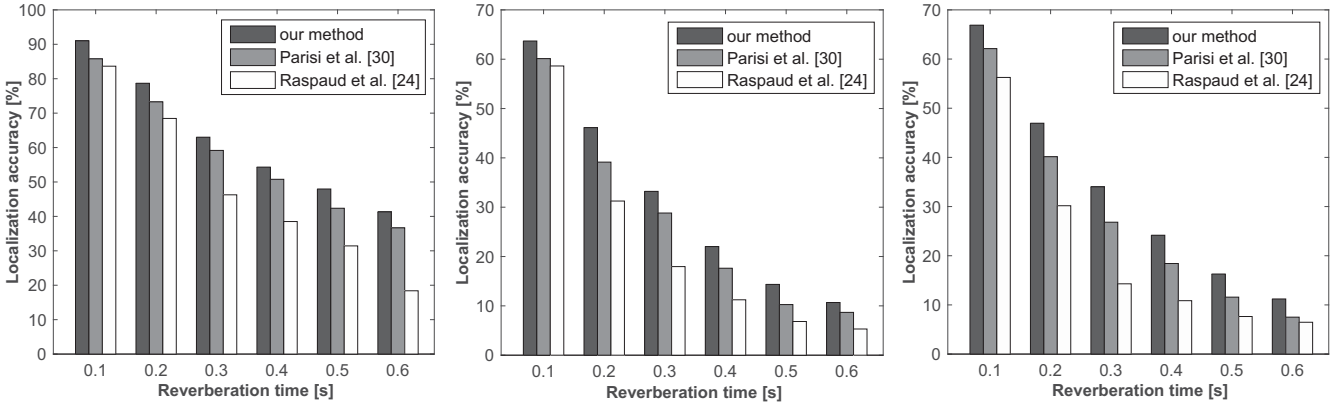


Fig. 10. Localization accuracies of different azimuths $\theta =$ (a) -15°; (b) -80°; (c) 80° at different reverberation times where tolerance = 0°.

2) Azimuth estimation in reverberant environments: In this part, several related methods are compared with the proposed method under different reverberant conditions. The comparison methods include the works proposed by Raspaud *et al.* [24] and Parisi *et al.* [30], because they have similar localization framework or reverberation-preprocessing stage with the proposed method. Table I compares their azimuth estimation results in terms of localization tolerance and reverberation time. It can be seen that the proposed method achieves higher localization accuracies than the other two methods. For example, the accuracy of our method is improved by about 10% and reaches 71.04% for the case when $T_R = 0.3$ s with 10° tolerance. With regard to the localization resolution, it should be clarified that the tolerance 0° does not mean without error, but error $< 5^\circ$ instead, and the tolerance 5° indicates error $< 10^\circ$. The resolution is determined by the offline HRTFs measurement, i.e., how to divide the localization space of interest.

Generally, Raspaud's method is more effective in anechoic environments, but its performance degrades rapidly when reverberation time increases. In fact, Raspaud's method obtains the worst results in the environments with strong reverberation, because the extraction of ITD and IID by Raspaud's method is deteriorated by the influence of reverberation. Parisi's method performs somewhat better than Raspaud's via using cepstral prefiltering. In Parisi's method, the cepstral prefiltering is used to dereverberate the binaural audio for valid time difference. Compared to Parisi's method, the proposed method achieves better results at various reverberation times. The reason is that the reverberation weighting can achieve better dereverberation than cepstral prefiltering, so that the binaural cues are better preserved. In addition, the template matching can integrate the robust azimuth estimations of ITD and IID across frequency bands. Moreover, the generalized parametric mapping improves the reliability of the azimuth estimates based on ITD and IID through finding the generalized scaling factors for non-specified subjects.

Some more detailed comparisons are shown in Fig. 10, where three azimuths (i.e., $-15^\circ, -80^\circ, 80^\circ$) are separately estimated by the above three methods at different reverberation times. It can be seen that our method outperforms the others, especially under strong reverberant conditions. Since the extraction of binaural cues is influenced by the reverberation, Raspaud's method becomes ineffective in the reverberant environments. With the generalized parametric mapping and template matching, our method can obtain more precise joint estimates of ITD and IID, thus our method is more robust and generalized than Parisi's approach.

TABLE II
THE LOCALIZATION ACCURACY VERSUS NORMALIZED TIME CONSUMPTION OF LOCALIZATION PROCESS

	our method	*Parisi et al. [30]*	*Raspaud et al. [24]*
Accuracy	68.53%	64.02%	63.86%
Norm.-time	1.0	0.93	0.91

In order to compare the computational complexities of the localization processes in different methods, we measured the time consumption of Matlab implementations of the three algorithms that are compared in this section. Table II compares the localization accuracy with 0° tolerance versus the time consumption. The execution times are normalized by that of the proposed method, such that the proposed method is regarded as the benchmark. The results are obtained under the condition with a reverberation time of 0.2 s, and same binaural cues extracted from the binaural signals preprocessed by reverberation weighting, are given to localization processes of the three comparison methods. From Table II, the time consumption of the proposed method is slightly higher than methods in [24] and [30] because of the azimuth refinement through template matching, but the localization performance

TABLE III
THE LOCALIZATION ACCURACIES OF AZIMUTH θ AT DIFFERENT SNRS

SNR	Environment without noise			20 dB			10 dB		
Tolerance	0°	5°	10°	0°	5°	10°	0°	5°	10°
our method	$\mathbf{93.16\%}$	$\mathbf{98.03\%}$	$\mathbf{99.85\%}$	$\mathbf{89.03\%}$	$\mathbf{97.45\%}$	$\mathbf{99.31\%}$	$\mathbf{69.63\%}$	$\mathbf{85.12\%}$	$\mathbf{92.13\%}$
IMF [44]	92.16%	96.52%	98.56%	86.49%	96.54%	97.26%	68.58%	81.94%	90.65%
Online Calibration [19]	89.12%	96.76%	99.24%	84.26%	95.92%	98.24%	58.94%	67.52%	75.23%
Hierarchical System [18]	93.90%	98.70%	99.87%	85.64%	97.21%	98.72%	63.64%	79.50%	84.13%

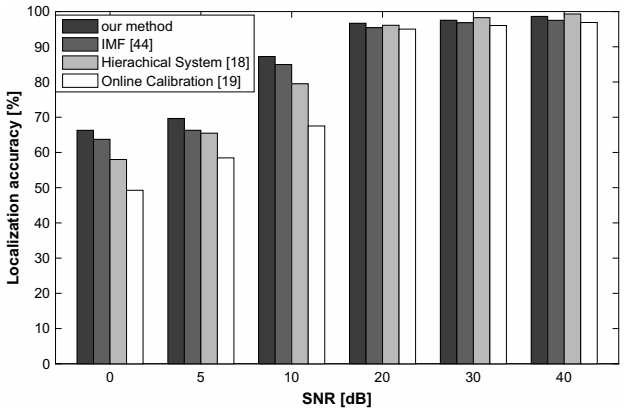


Fig. 11. Localization accuracies of our method compared with several popular methods at different SNRs with 5° tolerance.

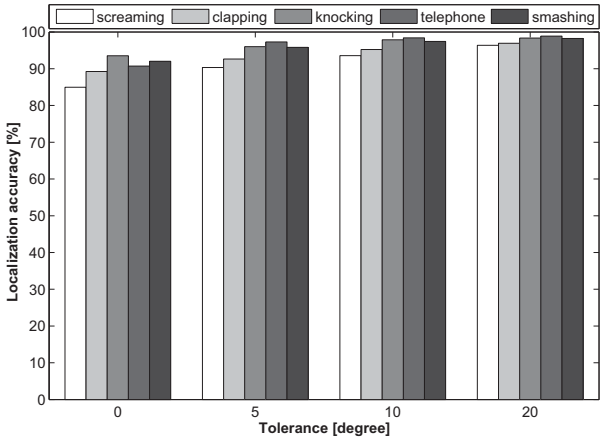


Fig. 12. Localization accuracies of different sound activities with different tolerances at SNR = 20 dB.

of our method is better through combing the robust estimations of ITD and IID. Due to the fact that the localization processes in [24] and [30] are similar, there is small difference among their time consumption.

3) Azimuth estimation in noisy environments: Some comparisons with several state-of-the-art methods, including a classical Hierarchical System [18], Online Calibration [19] and Interaural Matching Filter (IMF) [44], are carried out in the noisy environments without reverberation, i.e., $T_R = 0$ s. In fact, [18] and [19] belong to the hierarchical methods using different binaural cues. In [44], the IMF was proposed as a new localization cue, which contains somewhat relative transfer function information, to achieve a real-time sound localization in noisy environments. All these methods use both ITD and IID for sound localization, and the two-step localization process in our method can also be viewed as a two-stage hierarchical localization strategy. That is why they are involved for comparison in this part.

Table III shows the localization results at different SNRs. It can be seen that our method obtains the best performance when SNR decreases. In detail, the performance among these four algorithms has small gaps in the environment without noise. In this case, the performances of all the methods exceed 89% with a tolerance of 0°, and over 99% with a tolerance of 10°, that means, they have somehow satisfied the accuracy requirement in the quiet environments. The Online Calibration gets the worst performance, because it only takes the raw ITDs and IIDs as localization cues, while the other approaches still use other additional cues, e.g., IMF, spectral differential cues. IMF achieves a comparable performance with our method in most cases, but its performance would be worse in reverberant environments. For the Hierarchical System, ITD, IID and spectral cues are involved in three layers, respectively, it thus works better than the Online Calibration. However, in the mild noisy environments (e.g., SNR = 10/20 dB), the proposed method achieves favourable results compared with others. This superiority mainly owes to two aspects: 1) the generalized parametric mapping with generalized scaling factors efficiently determines the correct phase unwrapping factor to provide an unwrapped ITD for the subsequent template matching; 2) the usage of template matching with the averaged ITD and IID templates can effectively combine the robust estimates from ITD and IID at different frequency bands. Since ITD and IID are separately more robust for low and high frequency bands according to the Duplex theory [10], template matching helps to provide a more robust localization performance under the adverse conditions. Therefore, the generalized parametric mapping and the template matching are effective and reliable for a specific subject in the training set, which enables the proposed approach to have better consistency and robustness in a variable environment.

More detailed localization accuracies of the four algorithms at different SNRs are illustrated in Fig. 11 where tolerance = 5°. The proposed method performs with obvious superiorities under these noisy conditions. For high SNRs, there are small gaps among the four methods. For low SNRs, our method and IMF achieve comparable performances, which are much better than the other two methods. In the strong noisy environments (e.g., SNR $<$ 5dB), the performances of the Hierarchical System and Online Calibration degrade rapidly. In these cases, they cannot calculate the correct ITD, because it is difficult for the classical GCC-PHAT [7] to extract the notable spectral peak that marks the correct time delays.

4) Sound Activity Localization: In order to evaluate the robustness of our method to different types of sound sources,

Fig. 13. Sound localization in realistic environment. (a) Experimental scene of realistic environment. (b) The dummy head used for binaural recording.

we test five different sound activities under the noisy condition where SNR = 20 dB. These activities include *clapping hands*, *knocking on a door*, *telephone ringing*, *screaming* and *glass smashing*, which are common in people's daily life. These sound activities are recorded in an office environment, which are convolved with the simulated BRIRs to generate binaural signals. The localization results of these activities are shown in Fig. 12, we can see that they are well localized in the horizontal direction. For instance, when the tolerance is $0°$, the azimuth estimation accuracy is higher than 85%. Note that the localization performance of *screaming* is slightly worse than the other four. This phenomenon is caused by the sounding property that the intensity of *screaming* mainly converges to the high frequency bands, which makes harder for ITD estimation. As stated before, ITD is more ambiguous in the high frequency bands because of the phase unwrapping. Fortunately, these localization results are good enough and acceptable for the practical applications, which also verifies the robustness and adaptability of the proposed method to different types of sound sources.

C. Localization in realistic environment

In order to testify the proposed method in a more realistic setting, the KU100[1] dummy head is used to collect the binaural signals. The dummy head is positioned in an office as a typical indoor environment, and its setup is shown in Fig. 13(a). The Fig. 13(b) shows the structure of the dummy head, which is a replication of human head, equipped with microphones within the "ears". The sound signal is collected by the two "ears" and transferred to a computer through an ICON[2] mobile sound card with a sampling frequency of 44.1 kHz. The experimental office room is of dimensions (6 m $\times$ 5 m $\times$ 3 m). Since the walls and roof of the room are made of painted concrete and the floor is resilient, the reverberation time of this room is around 0.3 s. The SNR of this environment is about 20 dB. The KU100 dummy head is placed at the center of the room. The distance between sound sources and the head is set to 1 m. A sound source moves from -40° to 40° at an interval of 10°. For each direction, 20 groups of audio data are recorded, which are from the speech data of 10 males and 10 females in the TIMIT database [45]. The average length of these audios is 2 s. As the focus of this work is azimuth estimation, both the heights of microphones and sound sources are set to 1.5 m, namely they are on the same horizontal plane. Since the distance between two ears of this dummy head is 18 cm, which is similar to the average distance between the two ears in the CIPIC HRTF database, the previous models and templates are directly used here.

[1]https://www.neumann.com/?lang=en&id=current_microphones&cid=ku100_description

[2]http://www.iconproaudio.com

TABLE IV
THE AVERAGE LOCALIZATION ACCURACIES OF AZIMUTH θ COMPARED WITH SEVERAL POPULAR METHODS IN REALISTIC ENVIRONMENT

Tolerance	$0°$	$5°$	$10°$
our method	$\mathbf{84.60\%}$	$\mathbf{87.41\%}$	$\mathbf{89.36\%}$
Parisi et al. [30]	81.43%	83.58%	87.26%
Raspaud et al. [24]	76.28%	77.57%	81.93%
IMF [44]	80.13%	82.26%	85.65%
Online Calibration [19]	71.29%	74.52%	75.34%
Hierarchical System [18]	77.21%	79.67%	83.49%

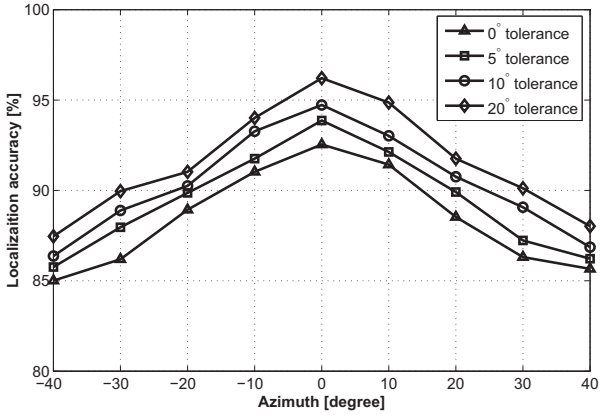


Fig. 14. The average localization accuracies of our method for different tolerances and directions in realistic environments.

Fig. 14 shows the average accuracies of azimuth estimation using our method with different localization tolerances. It can be seen that the localization accuracy reaches over 85% with $5°$ tolerance, which benefits from the proposed generalized parametric mapping and the joint estimation of ITD and IID improved by the template matching. Besides, our method obtains higher accuracies for the directions in front of the head, while its performance declines with increasing azimuth (absolute value). This phenomenon is due to that the shield of human head makes the ITD estimation with larger error at lateral directions.

The performances of different methods in the realistic indoor environment are compared in Table IV. Based on their models and templates under the above simulated condition, the localization accuracies with different tolerances are obtained. It can be seen that when both noise and reverberation are present, our method performs much better with different tolerances as compared to other approaches. Using the reverberation-

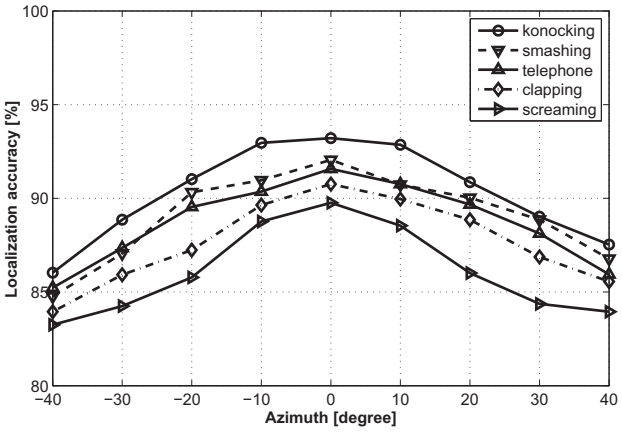


Fig. 15. The average localization accuracies of different activities in different directions with 5° tolerance.

preprocessing part, the method in [30] obtains second-best performance, which is much better than [24]. Both [24] and [30] only adopt ITD for final sound localization, so they work worse than the proposed method. The performance of IMF is slightly lower than [30], because reverberation disturbs the extraction of IMF. Since spectral cues are used in Hierarchical System, it works better than Online Calibration. It can be concluded that the methods with reverberation-preprocessing part work better than those without it, which also verifies the effectiveness of reverberation weighting. Besides, the effective combination of robust estimates from ITD and IID makes our method more reliable against noise and reverberation.

In order to testify the adaptability of our method in the natural indoor environments, the aforementioned five different sound activities emitted by a loudspeaker, are recorded through the KU100 artificial head with the identical setup. Their recorded binaural audio are used to evaluate the proposed method. The average localization results for these sound activities are shown in Fig. 15. It shows that the accuracy of *knocking* is best, and that of *screaming* is worst, which is consistent to the results presented in Section VI-B4. Therefore, these real experiments reveal that the generalized parametric mapping and the template matching are effective and reliable for sound localization in a realistic environment.

VII. Conclusions and future works

In order to achieve a reliable azimuth localization in the realistic environments including noise and reverberation, we proposed a novel sound source localization method based on reverberation weighting and generalized parametric mapping in a binaural context. The reverberation weighting effectively suppresses the influence from the indoor reverberation, and precisely preserves interaural time/intensity difference for the following sound localization. Binaural cues, namely frequency-domain ITD and IID, are extracted from the dereverberated signals. The azimuth estimation is refined by combining them together through a two-step localization process with the generalized parametric model and template matching.

The proposed generalized parametric mapping optimizes the nonlinear mapping relationships between ITD/IID and azimuth, which builds a generalized parametric model through finding the generalized scaling factors by solving a least squares problem. The generalized parametric model is computationally efficient and improves the adaptability of the proposed method to different artificial heads. The two-step localization process effectively refines azimuth estimation based on the generalized parametric model and template matching. Through achieving rough azimuth estimation based on the generalized parametric model, the correct phase unwrapping factor is quickly determined to unwrap ITD. The template matching effectively combines the robust raw estimates from the unwrapped ITD and the online measured IID across frequency to achieve a precise azimuth estimation. Besides, it also indicates the dependence between ITD and IID. Experiments in both simulated and realistic environments demonstrate the effectiveness and adaptability of our method for various types of sound sources, environments as well as artificial heads. Since the proposed method only considers the azimuth information of a single sound source, the future work may focus on elevation estimation and multi-sound source localization.

Acknowledgment

The authors would like to thank the three anonymous reviewers for their constructive comments and suggestions.

References

[1] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[2] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. Multimed.*, vol. 10, no. 3, pp. 538–548, 2008.

[3] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. ICASSP*, pp. 2814–2818, 2015.

[4] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2171–2186, 2016.

[5] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34, no. 3, pp. 276–280, 1986.

[6] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Proc. ICRA*, vol. 1, pp. 1033–1038, 2004.

[7] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[8] D. S. Talagala, W. Zhang, T. D. Abhayapala, and A. Kamineni, "Binaural sound source localization using the frequency diversity of the head-related transfer function," *J. Acoust. Soc. Amer.*, vol. 135, no. 3, pp. 1207–1217, 2014.

[9] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 728–739, 2008.

[10] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psychol.*, vol. 41, no. 1, pp. 35–39, 1948.

[11] R. F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, no. 7, pp. 1119–1134, 1988.

[12] M. Heckmann, T. Rodemann, F. Joublin, C. Goerick, and B. Scholling, "Auditory inspired binaural robust sound source localization in echoic and noisy environments," in *Proc. IROS*, pp. 368–373, 2006.

[13] K. Youssef, S. Argentieri, and J.-L. Zarader, "A binaural sound source localization method using auditive cues and vision," in *Proc. ICASSP*, pp. 217–220, 2012.

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

[15] X. Li and H. Liu, "Sound source localization for hri using foc-based time difference feature and spatial grid matching," *IEEE Trans. Cybern.*, vol. 43, no. 4, pp. 1199–1212, 2013.

[16] J. Zhang and H. Liu, "Robust acoustic localization via time-delay compensation and interaural matching filter," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4771–4783, 2015.

[17] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 774–784, 2006.

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

[19] H. Finger, S.-C. Liu, P. Ruvolo, and J. R. Movellan, "Approaches and databases for online calibration of binaural sound localization for robotic heads," in *Proc. IROS*, pp. 4340–4345, 2010.

[20] J. Benesty and J. Chen, "A multichannel widely linear approach to binaural noise reduction using an array of microphones," in *Proc. ICASSP*, pp. 313–316, 2012.

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

[22] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, 2010.

[23] K. Youssef, S. Argentieri, and J.-L. Zarader, "A learning-based approach to robust binaural sound localization," in *Proc. IROS*, pp. 2927–2932, 2013.

[24] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 68–77, 2010.

[25] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, 2010.

[26] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 1–13, 2011.

[27] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1503–1512, 2012.

[28] C. Pang, J. Zhang, and H. Liu, "Direction of arrival estimation based on reverberation weighting and noise error estimator," in *Proc. Interspeech*, pp. 3436–3440, 2015.

[29] H. Liu, C. Pang, and J. Zhang, "Binaural sound source localization based on generalized parametric model and two-layer matching strategy in complex environments," in *Proc. ICRA*, pp. 4496–4503, 2015.

[30] R. Parisi, F. Camoes, M. Scarpiniti, and A. Uncini, "Cepstrum prefiltering for binaural source localization in reverberant environments," *IEEE Signal Process. Lett.*, vol. 19, no. 2, pp. 99–102, 2012.

[31] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Process.*, vol. 59, no. 3, pp. 253–266, 1997.

[32] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Work. Appl. Signal Process. to Audio Acoust.*, pp. 99–102, 2001.

[33] K. I. Furuya and A. Kataoka, "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1579–1591, 2007.

[34] A. Tsilfidis, A. Westermann, J. M. Buchholz, E. Georganti, and J. Mourjopoulos, "Binaural dereverberation," in *The technology of binaural listening*, pp. 359–396, 2013.

[35] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Amer.*, vol. 37, no. 3, pp. 409–412, 1965.

[36] K. Lebart, J. M. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[37] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Univ.Eindhoven, Eindhoven, The Netherlands, Jun. 2007.

[38] A. Westermann, J. M. Buchholz, and T. Dau, "Binaural dereverberation based on interaural coherence histogramsa," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 2767–2777, 2013.

[39] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 6, pp. 709–716, 2003.

[40] R. O. Duda, "Elevation dependence of the interaural transfer function," in *Binaural and spatial hearing in real and virtual environments*, pp. 49–75, 1997.

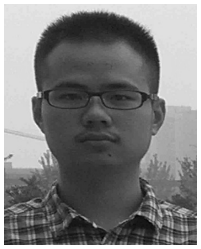
[41] R. M. Stern, G. J. Brown, and D. Wang, "Binaural sound localization," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, pp. 147–185, 2006.

[42] D. Campbell, K. Palomaki, and G. Brown, "A matlab simulation of shoebox room acoustics for use in research and teaching," *Comput. Inf. Syst.*, vol. 9, no. 3, pp. 48–51, 2005.

[43] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[44] H. Liu, J. Zhang, and Z. Fu, "A new hierarchical binaural sound source localization method based on interaural matching filter," in *Proc. ICRA*, pp. 1598–1605, 2014.

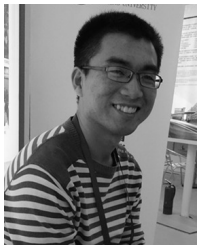
[45] J. Garfolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," *National Institute of Standards and Technology*, 1993.



Cheng Pang received the B.E. degree in mechatronic engineering in 2013. He is currently a Ph.D. student in the School of Electronics Engineering and Computer Science (EECS), Peking University (PKU), China. His current research interests are speech and audio signal processing, with a focus on sound source localization, speech enhancement and speech separation.

Hong Liu received the Ph.D. degree in mechanical electronics and automation in 1996, and serves as a full professor in the School of EECS, 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 Trans. on Signal Processing, and IEEE Trans. on PAMI.



Jie Zhang was born in Anhui Province, China, in 1990. He received the M.S. degree at the School of Electronics and Computer Engineering, Shenzhen Graduate School, Peking University, China. Currently, he is working on Ph.D. in the Circuits and Systems group of Delft University of Technology (TU Delft).

His current research interests are audio signal processing, speech enhancement, localization, wireless sensor networks and distributed optimization.

Xiaofei Li received the Ph.D. degree in electronics from Peking University, Beijing, China, in 2013. He is currently a Postdoctoral Researcher at INRIA (French Computer Science Research Institute), Montbonnot Saint-Martin, France. His research interests include multimicrophone speech processing for sound source localization, separation and dereverberation, single microphone signal processing for noise estimation, voice activity detection, and speech enhancement.