A BINAURAL SOUND SOURCE LOCALIZATION MODEL BASED ON TIME-DELAY COMPENSATION AND INTERAURAL COHERENCE

Hong Liu, Jie Zhang

Engineering Lab on Intelligent Perception for Internet of Things (ELIP),
Shenzhen Graduate School, Peking University, China
hongliu@pku.edu.cn, zhangjie827@sz.pku.edu.cn

ABSTRACT

Binaural sound source localization is an important technique involving speech capture and enhancement. However, the simple array structure makes it hard to localize sources in complex noisy conditions. This paper presents a novel algorithm based on time-delay compensation (TDC) and interaural coherence for binaural sound localization. Firstly, the TDC of binaural signals is used to estimate interaural time-delay (ITD) and interaural intensity difference (IID) instead of generalized cross correlation and logarithmic energy ratio. Then the interaural coherence is utilized to select reliable frames and reduce the variance of ITDs. Finally, a hierarchical framework, which successfully reduces computation complexity, is applied to make a decision of location based on Bayesian rule. Our innovation lies in that both ITD and IID are foremost yielded by TDC. Compared with other popular algorithms, experiments show that the most extrusive superiority of this method is complexity for both time and storage.

Index Terms— Sound source localization, time-delay compensation, interaural coherence, hierarchical framework

1. INTRODUCTION

Binaural sound source localization (SSL) is an essential and popular technique in many applications such as video-conference, smart rooms, and human-computer interaction, just as the human auditory localization with the capability of pinpointing the sound source swiftly and accurately [1,2]. There are two significant binaural (interaural) cues based on differences in time and level of the sound arriving at two ears called interaural time differences (ITDs) and interaural intensity differences (IIDs) [3,4]. Last decades, a large amount of binaural localization algorithms have been developed in various experimental environments.

Most traditional methods are based on ITD or IID and seldom consider the influence on each other [5-9]. Intuitively, with the influence of ITD, the signals received by two ears have different starting points with respect to sound source, which affects the extraction of IID. Willert et al. presented two-dimensional frequency versus time-delay representation of binaural cues, so-called activity maps [10], and this idea has improved in [11]. Hierarchical system was proposed by Li et al. to cut down matching times [7]. However, ITDs are usually calculated by the classical generalized cross correlation (GCC) [12], and IIDs defined by logarithmic energy ratio, which namely means that two free-running progresses are required to reckon binaural cues.

Accordingly, this paper raises a novel time-delay compensation (TDC) algorithm, which can evaluate ITDs and IIDs by the same processor. Generally speaking, this mentality can effectively decrease the redundancy of realization. The interaural coherence is used to modified time-delay estimate by choosing reliable frames, and the newly GCC-TDC function is put forward to depress ITDs fluctuating. Then, a hierarchical framework based on Bayesian rule is adopted to reduce the computing complexity, in which ITD in the first layer is used to select candidate azimuths and IID to make a decision.

Relation to prior work: This work has focused on an improved version of TDC algorithm, and the localization progress has taken advantage of hierarchical framework. Although Willert et al. has proposed activity maps for binaural SSL, and we have developed TDC to consider the relationship between ITDs and IIDs as well as improve the performance in noisy environments [11], all those previous works count binaural cues in substeps rather than holistic computation, and serialization almost need more time than parallelization. Hierarchical system is put forward by Li et al., which can reduce time complexity but increase space complexity, because more layers need more priori templates. In addition, experiments verified that dividing frequency sub-bands has little help to localization but adding storages by an order of magnitude, because low-frequency signals such as speech can easily go around heads [4,11,13].

The rest of this paper is organized as follows: TDC and hierarchical algorithms are introduced in Sect.2 and Sect.3, respectively. Experiments and analysis are shown in Sect.4. At last, conclusions are drawn in Sect.5.

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

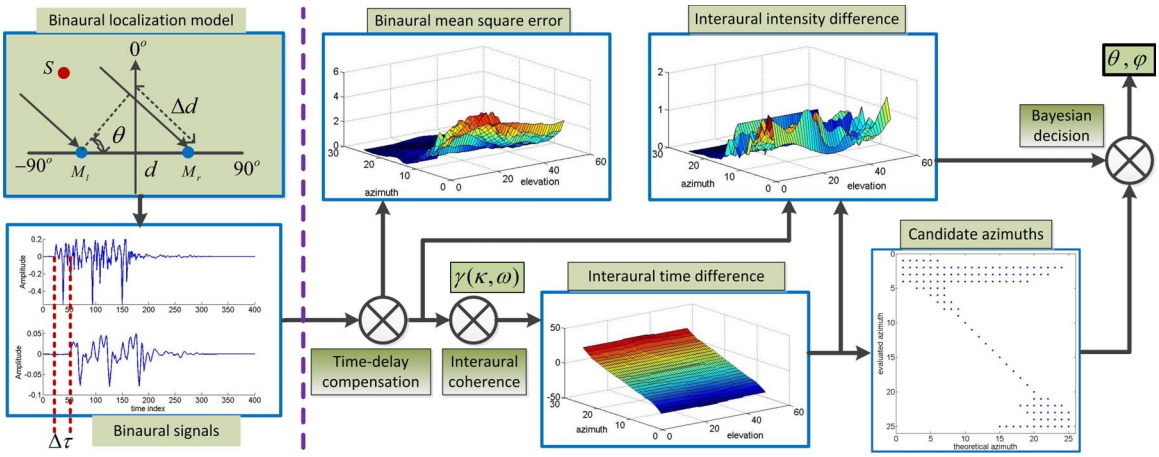

Fig. 1: A brief illustration of this binaural localization framework. The left part is modeling based on interaural-polar coordinate system. The core of right part is time-delay compensation, from which both ITD and IID can be solved.

2. TIME-DELAY COMPENSATION

2.1. Feature Extraction

Let $s(n)$ denote a sound source signal, and the received signals as $x_l(n)$ and $x_r(n)$ on the two microphones or ears, respectively (see Fig.1). Assume that binaural signals are counterparts of sound source with time-delay and attenuation so as to simplify analysis, it can be attained:

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

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

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

Therefore, take the idea of time-delay compensation into account, 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, Eq.(3) can be rewritten as:

$$\Delta v = W \odot x_l(n-\Delta\tau) - \lambda W \odot x_r(n) \tag{4}$$

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

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

Therefore, the parameters λ and $\Delta\tau$ can be estimated by maximum likelihood estimation as follows:

$$\frac{\partial y}{\partial \lambda} = \frac{\partial}{\partial \lambda}||W \odot x_l(n-\Delta\tau) - \lambda W \odot x_r(n)||^2 \tag{6}$$

Set this partial derivative to zero and λ, namely interaural intensity difference (IID), can be easily solved as:

$$\widetilde{\lambda} = \frac{\sum_N W^2(n) x_r(n) x_l(n-\Delta\tau)}{\sum_N W^2(n) x_r^2(n)} \tag{7}$$

where N denotes the length of window. As with time-delay $\Delta\tau$, it's difficult to compute from $\partial y/\partial\Delta\tau$ directly, but transformed into frequency domain instead, and Eq.(5) can be rewritten as:

$$Y(e^{j\omega}) = ||\boldsymbol{X}_l(e^{j\omega})e^{-j\omega\Delta\tau} - \lambda \boldsymbol{X}_r(e^{j\omega})||^2 \tag{8}$$

where $Y(e^{j\omega})$ and $\boldsymbol{X}(e^{j\omega})$ are the Fourier transform of variance and binaural signals processed by window function. Therefore, if let

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

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

$$\begin{aligned} \frac{\partial Y(e^{j\omega})}{\partial\Delta\tau} &= \frac{\partial}{\partial\Delta\tau}\Big(\boldsymbol{A}^*(e^{j\omega})\boldsymbol{A}(e^{j\omega})\Big) \\ &= \frac{\partial \boldsymbol{A}(e^{j\omega})}{\partial\Delta\tau} \cdot \frac{\partial Y(e^{j\omega})}{\partial \boldsymbol{A}(e^{j\omega})} \\ &= -j2\omega \boldsymbol{X}_l^*(e^{j\omega})\boldsymbol{A}(e^{j\omega})e^{-j\omega\Delta\tau} \end{aligned} \tag{10}$$

Let $\partial Y(e^{j\omega})/\partial\Delta\tau$ be zero, for $j\omega$ and $e^{-j\omega\Delta\tau}$ are not equal to zero, it will be obtained:

$$\boldsymbol{X}_l^*(e^{j\omega})\Big(\boldsymbol{X}_l(e^{j\omega})e^{-j\omega\Delta\tau} - \lambda \boldsymbol{X}_r(e^{j\omega})\Big) = 0 \tag{11}$$

where $*$ indicates complex conjugate. Then take Eq.(11) back to time domain using inverse discrete fourier transform, it can be shown as:

$$\begin{aligned} \delta(n-\Delta\tau) &= R(n) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\lambda \boldsymbol{X}_l^*(e^{j\omega})\boldsymbol{X}_r(e^{j\omega})}{\boldsymbol{X}_l^*(e^{j\omega})\boldsymbol{X}_l(e^{j\omega})} \cdot e^{j\omega n} d\omega \end{aligned} \tag{12}$$

where $R(n)$ is the proposed GCC-TDC function, which rather resembles the Roth weighting [14] based on an optimal filter with $x_l(n)$, $x_r(n)$ as the input and reference signals [15,16], respectively. Thereout, $\Delta\tau$ can be estimated as:

$$\widetilde{\Delta\tau} = arg\max_n R(n) \tag{13}$$

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

2.2. Interaural coherence

Based on the aforementioned analysis, ITDs and IIDs can be extracted from TDC. Combined with Eq.(7,12), we can draw that although there is a mutual relationship between ITD and IID, λ has an influence on the height of $R(n)$ in fact. On the contrary, λ is heavily relied on time-delay, thus halcyon ITDs should be calculated first. Hereby, interaural coherence (IC) is employed into GCC-TDC [17,18]. The energies of left and right ear are evaluated by the recursive averages as:

$$\begin{aligned} E_l(\kappa,\omega) &= \alpha\cdot|X_l(\omega)|^2 + (1-\alpha)\cdot E_l(\kappa-1,\omega) \\ E_r(\kappa,\omega) &= \alpha\cdot|X_r(\omega)|^2 + (1-\alpha)\cdot E_r(\kappa-1,\omega) \end{aligned} \tag{14}$$

where κ marks the frame index with each frame of $5.8ms$ duration. The smoothing factor α is determined from time constant T and sampling frequency f_s as $\alpha = 1/(T\cdot f_s)$ [19]. Here the IC function can be defined as:

$$\gamma(\kappa,\omega) = \frac{E_{lr}(\kappa,\omega)}{\sqrt{E_l(\kappa,\omega)\cdot E_r(\kappa,\omega)}} \tag{15}$$

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

$$E_{lr}(\kappa,\omega) = \alpha\cdot X_l(\omega)X_r^*(\omega) + (1-\alpha)\cdot E_{lr}(\kappa-1,\omega). \tag{16}$$

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. As a result, the proposed GCC-TDC can be modified with $\gamma(\kappa,\omega)$ as:

$$\widetilde{R}(n) = \frac{\lambda}{2\pi}\int_{-\pi}^{\pi}\boldsymbol{\gamma}(\kappa,\omega)\frac{\boldsymbol{X}_l^*(e^{j\omega})\boldsymbol{X}_r(e^{j\omega})}{\boldsymbol{X}_l^*(e^{j\omega})\boldsymbol{X}_l(e^{j\omega})}\cdot e^{j\omega n}d\omega \tag{17}$$

Fig.2 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 slighter for GCC-TDC is fundamentally in view of minimizing variance, which brings about more stable ITDs.

3. SOUND SOURCE LOCALIZATION

The task of sound source localization is to achieve azimuth θ and elevation φ, so to speak, ITD and IID are needed to changed into angels. Considering the geometrical relation in Fig.(1), it can be generated:

$$\theta = \sin^{-1}(\Delta d/d) = \sin^{-1}(\widetilde{\Delta\tau}c/df_s) \tag{18}$$

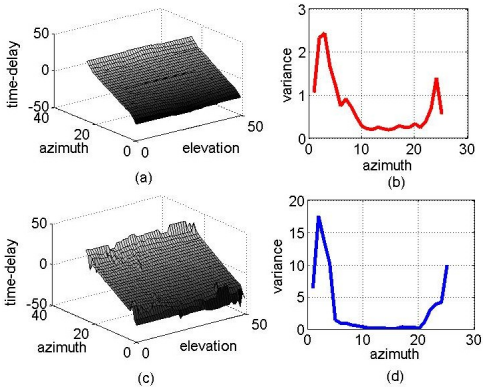

Fig. 2: The comparison of performance between GCC-TDC (upper) and GCC-PHAT (lower).

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

As to SSL, hierarchical localization framework is utilized. Firstly, the mean of time-delay $\overline{\tau_i}$ and the corresponding standard deviation σ_i can be trained for each azimuth θ_i. 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 comes a new sound source, the central azimuth is resolved and an available interval is achieved as follows:

$$\begin{aligned} P(\theta_i|\widetilde{\Delta\tau}) = P(\tau_i|\widetilde{\Delta\tau}) &= N(\widetilde{\Delta\tau}|\overline{\tau_i},\sigma_i^2) \\ \widetilde{\Delta\tau} &\subseteq (-3\sigma_i+\overline{\tau_i}, 3\sigma_i+\overline{\tau_i}) \end{aligned} \tag{19}$$

Then, consider intensity difference $\widetilde{\lambda}$ in the same train of thought, the average IID $\overline{\mu_j}$ and standard deviation δ_j can be trained for every direction. Based on the candidate azimuths in previous stage (see Fig.1), the probability of elevation φ_j and available interval of $\widetilde{\lambda}$ are obtained as:

$$\begin{aligned} P(\varphi_j|\theta_i,\widetilde{\lambda}) = P(\widetilde{\lambda}|\widetilde{\Delta\tau}) &= N(\widetilde{\lambda}|\overline{\mu_j},\delta_j^2) \\ \widetilde{\lambda} &\subseteq (-3\delta_j+\overline{\mu_j}, 3\delta_j+\overline{\mu_j}) \end{aligned} \tag{20}$$

Algorithm 1: *Sound Source Localization*

Input: ITD $\widetilde{\Delta\tau}$, IID $\widetilde{\lambda}$
Output: azimuth θ, elevation φ
Templates: ITDs, IIDs ;
if $\widetilde{\Delta\tau} \subseteq (-3\sigma_i+\overline{\tau_i}, 3\sigma_i+\overline{\tau_i})$ **then**
  $\theta_i \leftarrow acsin(\overline{\tau_i}c/df_s)$;
  $P(\theta_i|\widetilde{\Delta\tau}) \leftarrow N(\overline{\tau_i},\sigma_i^2)|_{\widetilde{\Delta\tau}}$;
end
while θ_i *exists* **do**
  if $\widetilde{\lambda} \subseteq (-3\delta_j+\overline{\mu_j}, 3\delta_j+\overline{\mu_j})$ **then**
    $P(\varphi_j|\theta_i,\widetilde{\lambda}) \leftarrow N(\overline{\mu_j},\delta_j^2)|_{(\widetilde{\Delta\tau},\widetilde{\lambda})}$;
  end
end
$(\theta,\varphi) \leftarrow arg\max_{(\theta_i,\varphi_j)} P(\theta_i|\widetilde{\Delta\tau})\cdot P(\varphi_j|\theta_i,\widetilde{\lambda})$;
return (θ,φ)

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

$$\begin{aligned}(\theta,\varphi) &= arg\max_{(\theta_i,\varphi_j)} P(\theta_i,\varphi_j|\widetilde{\Delta\tau},\widetilde{\lambda}) \\ &= arg\max_{(\theta_i,\varphi_j)} P(\theta_i|\widetilde{\Delta\tau})\cdot P(\varphi_j|\theta_i,\widetilde{\lambda})\end{aligned} \tag{21}$$

Then the detailed process is drawn in Algorithm 1.

4. EXPERIMENTS AND DISCUSSIONS

To evaluate our method, the CIPIC database is used in experiments which is measured by the U.C.Davis CIPIC Interface Laboratory including head-related impulse responses (HRIRs) for 45 different subjects [20]. The parameters used here are shown in Table 1.

Table 1: 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

The method in this paper is short by ICTDC, and the other three compared algorithms are TDC [11], Hierarchical System (HS) [7] and Probability Model (PM) [10], respectively. Experimental sound sources are captured in office environment with different signal-to-noise ratios (SNRs). The results of θ are illustrated in Table 2. We can see that in quite natural environment (40dB), all the four methods can achieve a very high accuracy of up to 90% and has little disparity, but when the SNR is 10dB, ICTDC has reached the best performance of increasing azimuthal accuracy by nearly 10%, which mainly owns to GCC-TDC obtaining more stable ITDs.

Table 2: The accuracy of θ in different SNRs.

SNR	40dB		10dB	
Error	$=0^o$	$\leq 10^o$	$=0^o$	$\leq 10^o$
ICTDC	**93.16**%	**99.56**%	**70.24**%	**92.72**%
TDC	90.28%	99.84%	62.64%	83.04%
HS	93.90%	99.87%	63.64%	84.13%
PM	92.72%	99.81%	58.92%	78.68%

With respect to elevation φ, a more obvious superiority has been reflected in Table 3. It can be obtained that HS lags behind the others seriously, because ICTDC, TDC and PM are the algorithms in same type based on considering the influence of ITD on IID. Besides, ICTDC has adopted interaural coherence function into time-delay estimation, which makes ITDs more robust from considerable reliable frames even in noisy environments.

The algorithm complexity is shown in Fig.3, from which it can been attained that this method requires the least complexity. Fig.3 a) counts the time consumption of these four algorithms by 800 times of random test. It's obvious that ICTDC successfully reduces the time consumption from **0.5**s of TDC down to **0.2**s approximately, which greatly relates to:

Table 3: The accuracy of φ in different SNRs.

SNR	40dB		10dB	
Error	$=0^o$	$\leq 11.25^o$	$=0^o$	$\leq 11.25^o$
ICTDC	**83.48**%	**94.80**%	**28.88**%	**55.44**%
TDC	70.48%	94.65%	25.56%	51.64%
HS	64.77%	95.23%	10.73%	32.10%
PM	65.49%	94.71%	25.76%	49.86%

- ICTDC calculates ITDs and IIDs all at once, which decreases the steps to evaluate binaural cues.
- The searching space is lessened to $O(n_a n_e)$ (see Fig.3 b)), where n_a, n_e and n_c denote the number of azimuth, elevation and frequency sub-bands, respectively, because ICTDC only needs to store the ITDs and IIDs in $n_a n_e$ directions referring in Algorithm 1, which derives from that dividing frequency contributes little benefits for TDC.
- The excellent matching strategy of hierarchical framework can also deflate candidates directions effectively.

Therefore, compared with others, ICTDC is more functional for SSL systems, especially for real-time sound source tracking, and so forth.

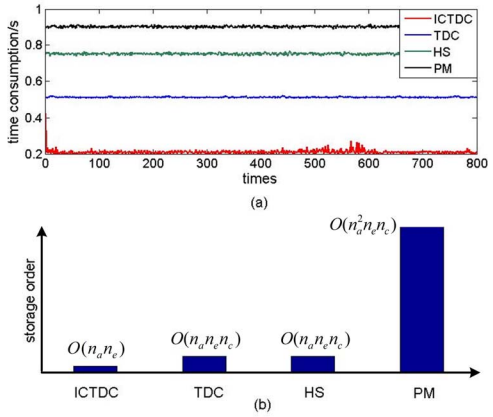

Fig. 3: a) Time consumption. b) The storage for templates.

5. CONCLUSIONS

In this paper, a novel binaural sound localization approach based on time-delay compensation (TDC) and interaural coherence is presented. This artifice not only increases the localization accuracy more or little, but the most importance of all is decreasing the complexity for both time consumption down to $0.2s$ and storage. The TDC relies on the influence of ITD on IID to extract binaural cues, which are foremost calculated by the same processor. Interaural coherence is applied into time-delay estimate, which can incline the variance of ITDs. The final localization is achieved by hierarchical system using Bayesian rule and searching by layers can effectively reduce matching times. Above all, our algorithm is more suitable for practical localization systems.

6. REFERENCES

[1] R. Flavio, C. Zhang, A. F. Dinei and E. B. Demba, "Using reverberation to improve range and elevation discrimination for small array sound source localization", IEEE Trans. on ASLP, vol.18, no.7, pp. 1781-1792, Sep.2010.

[2] B. G. Shinn-Cunningham, S. Santarelli and N. Kopco, "Tori of confusion: Binaural localization cues for sources within reach of a listener", J. Acoust. Soc. Amer., vol.107, no.3, pp. 1627-636, Mar. 2000.

[3] M. Raspaud, H. Viste and G. Evangelista, "Binaural Source Localization by Joint Estimation of ILD and ITD", IEEE Trans. on ASLP, vol.18, no.1, Jan. 2010.

[4] R. F. Lyon, and C. Mead, "An analog electronic cochlea", IEEE Trans. on ASSP, vol.36, pp. 1119-1134, 1988.

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

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

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

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

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

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

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

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

[13] N. Roman and D. Wang, "Binaural tracking of multiple moving sources", IEEE Trans. on ASLP, vol.16, no.4, pp. 728-739, May. 2008.

[14] P. R. Roth, "Effective measurements using digital signal analysis", IEEE Spectrum, vol. 8, pp. 62-70, Apr. 1971.

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

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

[17] M. Jeub, M. Dörbecker and P. Vary, "A semi-analytical model for the binaural coherence of noise fields", IEEE Signal Processing Letters, vol. 18, no. 3, Mar. 2011.

[18] M. Jeub, M. Schäfer, T. Esch and P. Vary, "Model-Based Dereverberation Preserving Binaural Cues", IEEE Trans. on ASLP, vol. 18, no. 7, Sep. 2010.

[19] 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.

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