BINAURAL SOUND LOCALIZATION BASED ON TIME-DELAY COMPENSATION AND SPATIAL GRID MATCHING

Baolong Zhao[1], Zhuo Fu[2], Jie Zhang[2], Yuezhao Chen[2], Runwei Ding[2]

[1] Shenzhen National Engineering Laboratory of Digital Television Co.,Ltd.
[2] Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Shenzhen Graduate School of Peking University, Shenzhen, 518055 CHINA.

blzhao@neldtv.org, {zhangjie827, chyuezhao}@sz.pku.edu.com,
gradyfu@hotmail.com, dingrunwei@pkusz.edu.cn

Abstract: Binaural source localization is a popular technique in various applications, such as hearing aids, mobile robot, video conference, etc. However, robust binaural cues estimate and suitable localization strategy are always limiting its performance. In this paper, a new algorithm for binaural sound source localization is presented, which is a two-layer model. In the first layer, a spectral weighting generalized cross correlation–phase transformation (GCC-PHAT) method is presented for robust time-delay estimation, by which the probabilistic azimuths of sound source are obtained. In the second layer, an improved algorithm is introduced, which is named Compensated Interaural Intensity Difference (CIID). Based on the probabilistic azimuth localization results and CIID features, spatial grid matching (SGM) is presented to provide a Bayesian model for localizing azimuth and elevation. Compared with three other algorithms, experimental results show that the proposed method has a robust result even in noisy environments.

Keywords: Sound source localization, Compensated interaural intensity difference, Spatial grid matching

1 Introduction

Binaural sound source localization (SSL) is an important technique in various fields such as speech capture, enhancement, hearing aids, hands free telephone devices, video conference, intelligent human-robot interactions (HRI), etc., for its easy-implementation with only two microphones. There are two significant binaural (interaural) cues based on differences in time and level of the sound arriving at two ears called interaural time difference (ITDs) and interaural intensity differences (IIDs). Last decades, Since “Duplex Theory” [1] and cochlear model [2] were proposed, a large amount of binaural localization algorithms have been developed in various experimental environments [3-5].

Actually, in many traditional methods, new obtained feature vector must match with each template to assure the direction of source. However, higher resolution needs more templates, and more time will be consumed for overall searching. In addition, most of them seldom consider the relationship between ITDs and IIDs. In fact, with the influence of ITD, the signals received by two ears have different starting points versus sound source, which will affect the extraction of IIDs. In general, the difference between starting points means time-delay.

As with these, this paper proposes a new sound source localization method based on time-delay compensation and spatial grid matching. In the first layer, the spectral weighting GCC-PHAT method is presented to estimate time-delay [8][9][11]. However, measuring error and surrounding interference induce the obtained time-delay incompletely reliable, that is, what can be made sure is the possible crude area the sound source located. By this method, candidate azimuths are prepared for the following operations and the matching times will drop dramatically.

In the second layer, a new algorithm named CIID is introduced, which is an improved algorithm of IID based on time-delay compensation and inspired by the algorithm used in [12][15]. With this algorithm, a Spatial Grid Matching (SGM) algorithm is utilized to refine azimuths and elevations. Since ITD offsets the influence of IIDs, it will be more effective than traditional IID.

Relation to prior works: This work has focused on the formulation of the CIID and SGM algorithm with two-layer model, which uses ITD's information to compensate IID. The work by Willert *et al.*[10] considers ITD's influence on IID in sub- frequency bands, but it needs larger storage space and has a worse result in noisy environment. The works by Li *et al.* [12] and Finger *et al.* [6] present multilevel models to reduce computational complexity, but they neglect the impact of ITD on IID. In [12], the binaural sound localization was based on a hierarchical framework, in which ITD, IID and spectral cues were utilized to make sure direction candidates, respectively. [6] had a similar consideration calibrating the localization space by binaural cues hierarchically. While CIID is related to recent approaches, using ITD to compensate IID was not involved in these earlier studies. SGM was proposed in [9], which divided the ambient environment into various grids so as to improve the localization resolution, yet minor grid means high computational expense. In addition, the interaural matching filter in [16] also provides a vital reference.

The rest of this paper is organized as follows: Sect.2 gives

978-1-4799-4719-5 /14/$31.00 ©2014 IEEE

the detailed binaural localization model in this paper including the spectral weighting GCC-PHAT for time-delay estimate, time-delay compensation and spatial grid matching algorithm. Experiments and analysis are shown in Sect.3. At last, some conclusions are drawn in Sect.4, respectively.

2 Binaural Localization Model

In this section, we will fully introduce our binaural localization model, including the spectral weighting GCC-PHAT for time-delay estimate, a new cue name CIID and the spatial grid matching algorithm to save time comsuption.

2.1 Spectral Weighting GCC-PHAT

In this subsection, the spectral weighting GCC-PHAT method is presented to find the average time-delay. The GCC-PHAT can be formulated as:

$$R\left(n\right)=\int_{-\pi}^{\pi}\frac{X_l\left(\omega\right)X_r^*\left(\omega\right)e^{j\omega n}}{\mid X_l\left(\omega\right)X_r^*\left(\omega\right)\mid}d\omega \tag{1}$$

where $R(n)$ is the generalized cross correlation function of signals between two ears. $X_l\left(\omega\right)$ and $X_r\left(\omega\right)$ are the Discrete Fourier Transforms (DFT) of the signals received by two ears.

The spectral weighting GCC-PHAT is to amplify peak of $R(n)$ by designing a weighting function. Firstly, the power spectrum of noise can be estimated from forepart of signal:

$$E_\eta(\omega)=\frac{1}{N}\sum_{n=1}^{N}\left|\eta_n(\omega)\right|^2 \tag{2}$$

where $\eta_n\left(\omega\right)$ denotes the spectrum of the n^{th} noise frame. N represents the number of noise frame. In addition, the power of current noisy signal frame is:

$$E_x\left(\omega\right)=\left|X\left(\omega\right)\right|^2 \tag{3}$$

Then posterior signal-to-noise ratio (SNR) versus frequency is given by:

$$\gamma(\omega)=max(\frac{E_X(\omega)}{E_\eta(\omega)},1) \tag{4}$$

Therefore, for two signals $x_l\left(n\right)$ and $x_r\left(n\right)$, the posterior SNR $\gamma_l\left(\omega\right)$ and $\gamma_r\left(\omega\right)$ can be obtained. When $E_X\left(\omega\right)>E_\eta\left(\omega\right)$ holds, Eq.(3) is a unbiased estimation of SNRs, and not vice versa, but is still valid for high noise levels. Hereby, silence/noise detection before evaluation is based on the voice activity detector (VAD) of the AMR-WB speech codec [14]. Then the proposed spectral weighting function is drawn as:

$$\Psi(\omega)=min((log_\kappa\gamma_r(\omega))^\beta,(log_\kappa\gamma_l(\omega))^\beta) \tag{5}$$

where κ can be adjusted in different environments to keep the weight reasonable. The greater overall SNR value is, the greater κ will be. Similarly, β is also adjustable, which controls the divergence of weighting coefficient. The greater β is, the greater divergence will be. Finally, the spectral weighting GCC-PHAT function can be written as:

$$R\left(n\right)=\int_{-\pi}^{\pi}\Psi(\omega)\frac{X_l\left(\omega\right)X_r^*\left(\omega\right)e^{j\omega n}}{\mid X_l\left(\omega\right)X_r^*\left(\omega\right)\mid}d\omega \tag{6}$$

Therefore, the time-delay can be achieved by detecting the peak of GCC-PHAT function formulated as:

$$\Delta\tau=argmax_nR\left(n\right) \tag{7}$$

Since $\Psi\left(\omega\right)$ is changed with SNR for acute peak of GCC-PHAT function, $\Delta\tau$ can be obtained accurately even in noisy environments. As to localization, all possible azimuths for each $\Delta\tau$ are trained offline, and the probability of each direction for the $\Delta\tau$ is stored as templates. When a new time-delay is obtained, all possible directions will be checked, and the most possible directions will be chosen and prepared for the second layer. For example, when $\Delta\tau$ is calculated as in -65^o the real direction may vary from -55^o, -65^o to -80^o as **Fig.1** shows. The error tolerance of lags for every azimuth is shown in **Fig.2**. Thus, the probability of direction G where the sound comes can be trained when $\Delta\tau$ is obtained as τ:

$$P(G\mid\tau)=\frac{sum(direction\in G,\Delta\tau=\tau)}{sum(\Delta\tau=\tau)} \tag{8}$$

which means the proportion between the number of sound source from direction G and from trained templates.

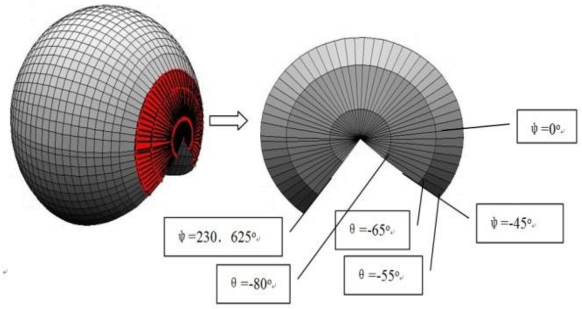

Fig. 1. All directions in the CIPIC database. The red area is the area of all possible directions

2.2 Time-Delay Compensation

In the second layer, the direction will be refined for both azimuth and elevation. Thereby, the performance of SSL relies heavily on the accuracy of the extraction of IIDs. CIID is to weaken the influence of ITD on IID, by which the time difference between binaural signals in the same

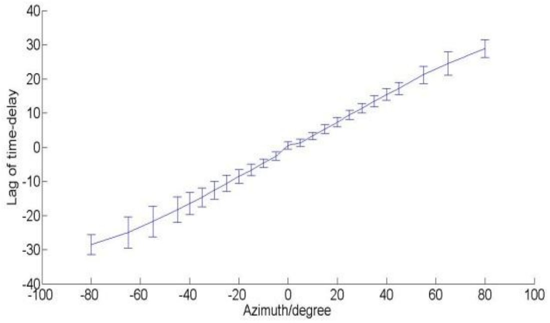

Fig. 2. The error tolerance of lags for every azimuth.

frame will be compensated.

After a *k* channels band-pass filtering, the sound signal is decomposed into *k* different signal components. In the m^{th} frequency channel, CIID can be achieved as:

$$W \odot x_l^m(n-\Delta\tau) = \lambda_m W \odot x_r^m(n) + \Delta\eta \qquad (9)$$

where W, λ_m, $\Delta\eta$ denote Square window function, real CIID in m^{th} channel and disparity of noises received by two ears. $\odot$ denotes an element-wise multiplication. From the standpoint of noise, Eq.(8) can be rewritten as:

$$\Delta\eta = W \odot x_l^m(n-\Delta\tau) - \lambda_m W \odot x_r^m(n) \qquad (10)$$

where $\Delta\eta$ can be thought as a zero-mean Gaussian noise. Since the task of CIID is to eliminate the difference between binaural signals as much as possible, $\Delta\eta$ will achieve the minimum value. Therefore, λ_m can be estimated by maximum likelihood estimate method using Minimum Mean Square Error criterion as:

$$\lambda_m = \min_{\lambda_m} \| W \odot x_l^m(n-\Delta\tau) - \lambda_m W \odot x_r^m(n) \|_2^2 \qquad (11)$$

λ_m can be obtained by partially differentiating versus λ_m. Assume this partial derivative be zero, then the result will be resolved as:

$$\lambda_m = \frac{\sum_N x_r^m(n) x_l^m(n-\Delta\tau)}{\sum_N x_r^m(n)^2} \qquad (12)$$

where *N* denotes the length of the window. In order to restrict λ_m to a proper range, the logarithmic operation is imposed to CIID. As a result, when the candidate azimuth θ_i is acquired in first layer, CIID can be described as:

$$CIID(\Delta\tau, m)|_{\theta_i} = 20log_{10}\lambda_m|_{\theta_i} \qquad (13)$$

2.3 Spatial Grid Matching

After Eq.(12), a k dimension vector CIID will be obtained. Moreover, the vectors of adjacent sound sources will yield a pair of similar feature vectors. As **Fig.1** shows,

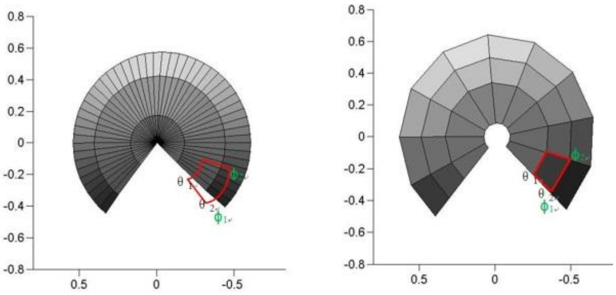

Fig. 3. The grid for candidate area when azimuth is calculated as -65° in the first layer.

sphere surface can be divided into many grids with a certain size. Those sound sources in a same grid are regarded as adjacent, whose feature vectors resemble to each other. On the contrary, when the geometrical distance between two different grids is farther, the corresponding feature vectors will be more different.

In this work, spatial grids are divided as follows: The resolution of azimuth is the same as the CIPIC database [7], which means different azimuth must be divided into different grids. Then five different adjacent elevations with same azimuth will be set in a same grid. As **Fig.3** shows, the red area identified in the left figure is thought as in a same grid as the red area did in the right figure.

Besides, a Gaussian Mixture Model (GMM) is constructed as template for each grid based on CIID vector during the training period. For an arbitrary grid, its azimuth distributes from θ_1 to θ_2, and its elevation distributes from φ_1 to φ_2 (see **Fig.3**). The GMM of this grid is trained offline.

As to localization, the problem will be simplified to find which grid the sound source is most matching. Let *G* denotes a grid and SSL can be mathematically formulated as:

$$\begin{aligned} G_g &= argmax_G P(G \mid CIID, \tau) \\ &= argmax_G \frac{P(CIID \mid G, \tau) P(G \mid \tau) P(\tau)}{P(CIID) P(\tau)} \\ &\propto argmax_G P(CIID \mid G) P(G \mid \tau) \end{aligned} \qquad (14)$$

where G_g denotes the grid located. The useful probabilities $P(G \mid \tau)$ and $P(CIID \mid G)$ are obtained in the previous layers, respectively.

Once the grid of sound source is assured, a further extraction operation should be done. The best direction should be refined from the five directions in the same grid, that is:

$$G_S = argmax_{g(i)} P(S \mid g(i)) \quad i \in [1,2,3,4,5 \qquad (15)$$

where G_S is the direction finally found for SSL and $g(i)$ denotes the i^{th} direction in *G*. The azimuth and elevation are described as:

$$\{\theta,\varphi\} = \{\frac{\theta_{g_1}+\theta_{g_2}}{2}, \varphi_{g_1} + (2\times i+1)\frac{\varphi_{g_2}-\varphi_{g_1}}{10} \qquad (16)$$

where $\theta_{g_1}, \theta_{g_2}, \varphi_{g_1}$ and φ_{g_2} compose the boundary of *G*.

3 Experiment and analysis

The CIPIC database of head-related impulse responses (HRIRs) [5] is applied in our experiments based on 20 groups of real sound signals for training grid templates. 100 groups of real sound signals (50 for human voice and 50 for music) are considered for each direction. The duration of each signal is 2 seconds. The sampling frequency is 44.1*kHz*. The result will be compared with the methods of Probabilistic Model [10], Hierarchical System [12] and Online Calibration [6]. The algorithm applied in this paper is CIID for short.

Here, two different experimental environments are considered, one for ideal condition without noise, the other for the condition with SNR 20dB (the additive white Gaussian noise environment). The result is obtained with different error tolerance according to the resolution of CIPIC.

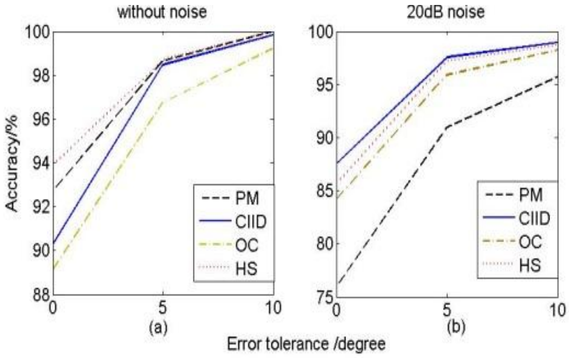

Fig. 4. The localization accuracy of azimuth in two different experimental environments

In **Fig.4** and **Fig.5**, the localization accuracy in two different experimental environments are shown. It can be found that under ideal conditions, there are little difference of performance among these four methods, and CIID has not achieved the best or the worst one. Strictly speaking, PM is the best. However, in noisy environment, our algorithm is the most robust one for both azimuth and elevation. The reasons lie in:

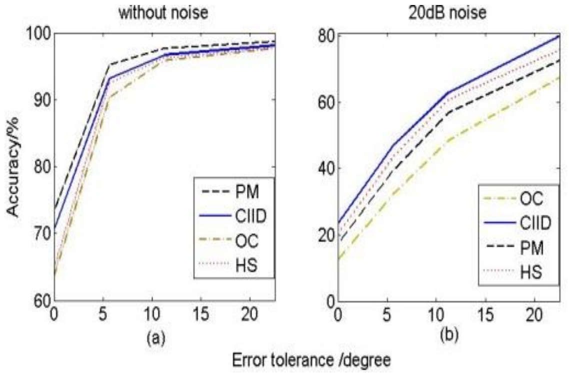

Fig. 5. The localization accuracy of elevation in two different experimental environments

- Apart from ITD and IID, spectral cues are also used in Hierarchical System, which are not effective enough in noisy environment.
- In Probabilistic Model, although ITD and IID are processed as a whole and the effects of ITD on IID are taken into account, it does not have robust scheme for time-delay estimation in quite noisy environment.
- The Online Calibration has not considered the influence of IID on ITD resulting in a worst performance than other methods.
- As to CIID, the improved GCC-PHAT algorithm can effectively deal with noise, therefore an accurate azimuth result can be achieved. Furthermore, CIID combines ITD with IID well. Also SGM has taken advantage of a searching way resembling decision tree for localization, which can save time complexity. As a result, CIID algorithm has the best performance.

4 Conclusions

In this paper, a new improved two-layer binaural sound localization model based on CIID and Spatial Grid Matching is presented. The whole circumstance is divided into spatial grids and different size means different resolution. In the first layer, spectral weighting GCC-PHAT is presented for acquiring all possible azimuths, which can extract more accurate time-delays in noisy environment. CIID is built upon the relation between ITD and IID. Under practical conditions, this method can achieve the accuracy over 90%. Accordingly, this two-layer model can effectively be well used for real human-computer interactions without increasing extra expense.

Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC, No.60875050, 60675025), National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247, 2012AA011705), Scientific and Technical Innovation

Commission of Shenzhen Municipality (No.JCYJ20120614152234873, CXC201104210010A, JCYJ20130331144631730,No.JCYJ20130331144716089).

References

[1] L. A. Jeffress, A place theory of sound localization, *J. comp. & phys. psyc.*, vol.61, pp.468-486, 1948.

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

[3] A. J. King, J. W. Schnupp, and T. P. Doubell, The shape of ears to come: dynamic coding of auditory space, *TRENDS in Cognitive Sciences*, vol. 5, no. 6, pp. 261-270, 2001.

[4] S. B. Andersson, A. A. Handzel, V. Shah, and P. Krishnaprasad, Robot Phonotaxis with Dynamic Sound-source Localization, *in Proc. IEEE ICRA*, vol. 5,pp. 4833-4838, 2004.

[5] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda, Sound localization for humanoid robots-building audio-motor maps based on the HRTF, *in Proc. IEEE IROS*, Beijing, China, pp. 1170-1176, Oct. 2006.

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

[7] V. Algazi, R. Duda, D. Thompson, and C. Avendano, The CIPIC HRTF database, *in Workshop IEEE Applic. of SPAA*, Mohonk, New York, pp. 99-102, Oct. 2001.

[8] H. Liu and X. F. Li, Time Delay Estimation for Speech Signal Based on FOC-Spectrum, *in Proc. INTERSPEECH*, Portland, USA, Sep. 9-13, 2012.

[9] X. F. Li, H. Liu and X. S. Yang, Sound Source Localization for Mobile Robot Based on Time Difference Feature and Space Grid Matching, *in Proc. IEEE IROS*, San Francisco, California, USA, pp. 2879-2886, 2011.

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

[11] C. H. Knapp, G. C. Carter, The generalized correlation method for estimation of time delay, *IEEE Trans. on ASSP*, pp. 320-327, 1976.

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

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

[14] TS 26.194, Adaptive Multi-Rate-Wideband speech codec, Voice Activity Detector, v6.0.0, *3GPP*, 2004.

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

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