2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Daejeon Convention Center
October 9-14, 2016, Daejeon, Korea

Probabilistic Binaural Multiple Sources Localization Based on Time-delay Compensation Estimator and Clustering Analysis

Hong Liu[1], Mengdi Yue[2] and Jie Zhang[3]

Abstract— **Sound source localization (SSL) is an essential technique in many applications, such as robot audition, human-robot interaction and speech capturing. However, SSL from a binaural input is still a challenging problem, particularly when multiple sources are active simultaneously. In this work, we propose a multi-sources localization framework based on the time-delay compensation (TDC) estimator and clustering analysis. The TDC estimator is a simultaneous operator to estimate binaural cues, which breaks the limitation of independent processors for binaural cues extraction. The multi-sources decision is realized by clustering analysis for the binaural cues of multiple signal frames. In experiments, we demonstrate that the localization performance is improved compared to the methods that assume the number of spatial stationary sources to be known. Results with both simulated and recorded impulse responses show that robust performance can be achieved with limited prior training, and our method is also adaptive to different sound activities.**

I. INTRODUCTION

Sound source localization (SSL) acts as an important role to recognize the direction of a sound source for humanoid robots, audio reproduction, scene analysis, etc. A significant number of source localization methods are proposed such as steered beamforming [1], high resolution spectral estimation [2] and time difference of arrival (TDOA) [3], [4]. The steered beamforming needs a priori knowledge of sound source and environmental noise which is difficult to obtained for realistic applications. High resolution spectral estimation requires the source to be stationary signal. TDOA algorithm has the characteristics of high computation efficiency and easy implementation, but it can not directly deal with multi-sources signals. Although SSL has been popular for the past several decades, it is still quite challenging to localize multiple sound sources, e.g., the cocktail scenarios [5].

For multiple sources localization, auditory scene analysis (ASA) and blind signal separation (BSS) are the mainly methods. BBS requires to be assumed statistical properties of signal and hybrid approaches, which is difficult to meet in actual acoustic environment. As human beings can accurately evaluate the sound sources in complex environments (including noise and reverberation), binaural SSL algorithm is present to do the same work. Binaural auditory processing based on human spatial hearing mechanism is a friendly and natural interaction technology. Since "Duplex Theory" [6] and cochlear model [7] were proposed, a large amount of binaural localization algorithms have been developed in various experimental environments [8]–[10].

There are two significant binaural cues for binaural SSL method [11] called interaural time difference (ITD) and interaural level difference (ILD). However, most traditional methods seldom consider the influence of binaural cues on each other. Intuitively with the impact of ITD, the signals perceived by two ears have different starting points with respect to the sound source, which affects the extraction of ILD. Furthermore, after binaural cues extraction, most previous algorithms use them to match the lookup computed from the head-related transfer functions (HRTFs). This would lead to a large amount of redundant solutions and more matching time. To solve the above problems, firstly, we present the time-delay compensation (TDC) estimator to evaluate the ITD and ILD simultaneously. The realization is simplified, as it breaks the limitation of two independent processors for binaural cues extraction. Secondly, a probabilistic strategy is introduced for ITD and ILD to calculate the probabilities of the direction of sound source.

In this paper, a multi-sources localization framework is proposed based on TDC estimator and clustering analysis. TDC estimator can estimate the binaural cues simultaneously. It is demonstrated that the joint of ITD and ILD can effectively overcome the overload selections of unwrapping ITDs and larger variance of ILDs in the low frequency bands. This single SSL, which consider both azimuth and elevation, is testified under different signal-to-noise ratio (SNR) conditions. Subsequently, this strategy is applied and extended to multi-sources scenarios. Since it is difficult to localize multi-sources only from one signal frame, we observe the spatial distributions of binaural cues in a long period of binaural audio. Through clustering the binaural cues, the centers are obtained which indicate the actual information on the sources, such that the multiple SSL achieved.

The rest of this paper is organized as follows: Sect.II introduces and analyzes the TDC estimator for binaural cues extraction. Sect.III shows the probabilistic strategy for single SSL. Sect.IV gives the *k*-means clustering for multiple SSL. Experiments and discussions are shown in Sect.V. At last, the conclusions and feature works are drawn in Sect.VI.

II. BINAURAL CUES EXTRACTION

A. Time-delay Compensation Estimator

In this section, a new binaural localization cues estimation method named time-delay compensation (TDC) [12] estima-

[1]Hong Liu and [2]Mengdi Yue are with the Key Laboratory of Machine perception (Ministry of Education), Shenzhen Graduate School, Peking University, Shenzhen, 518055 CHINA. {hongliu, yuemengdi}@pku.edu.cn

[3]Jie Zhang is with the Signal Information and Processing Lab, Delft University of Technology, 2628 CD Delft, The Netherlands. J.Zhang-7@tudelft.nl

978-1-5090-3762-9/16/$31.00 ©2016 IEEE

tor will be introduced. The propagation paths from the sound source to acoustic sensors are roughly parallel in the far-field scenario. Let $s(n)$ denote a sound source signal, we assume that binaural signals are counterparts of the sound source such that the differences among them lie in time-delay and attenuation. For briefly, we formulate the binaural audio as:

$$x_i(n) = a_i s(n-\tau_i) + v_i(n), \quad i \in \{l, r\}, \tag{1}$$

where n is the time index, a_i denotes the attenuation factor, τ_i is the time consumption propagating from the sound source to the two acoustic sensors, $v_i(n)$ represents the interference, l and r mean the left and right channels, respectively. Therefore, the interaural time-delay $\Delta\tau$ can be defined as:

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

Then, we expect to eliminate the differences of the binaural signals as much as possible. A monaural audio is compensated to match the other, i.e. passing through time alignment and magnitude stretching. This procedure is mathematically formulated as:

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

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 our goal is to minimize the error. From the standpoint of noises, (3) can be replaced by:

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

In an office environment, Δv is usually thought as zero-mean Gaussian noise. Hereby the variance of Δv is given by:

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

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

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

After setting this partial derivative to zero, namely, ILD λ can easily be solved as:

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

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

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

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

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

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

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

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

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

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

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

where $R(n)$ is the proposed GCC-TDC function, which rather resembles the Roth weighting [13], [14] based on an optimal filter with $x_l(n)$, $x_r(n)$ as the input and reference signals, respectively. Thereout, $\Delta\tau$ can be estimated by:

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

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

B. Time-frequency Analysis for TDC

The two cues considered in this paper, namely, the ILD and ITD, are based on the sliding short time fourier transform (STFT) spectra of the two observations. The ILD (in dB) at the κth frame is defined as:

$$\begin{aligned} ILD(\kappa,\omega) &= 20\log_{10}\mathscr{F}\{\lambda(\kappa,n)\} \\ &= 20\log_{10}\frac{|W(\omega)^2 X_r^\kappa(\omega) X_l^\kappa(\omega) e^{j\omega\Delta\tau}|}{|W^2(\omega) X_r^\kappa(\omega) X_r^\kappa(\omega)|}, \end{aligned} \tag{12}$$

where ω is angular frequency and X_r^κ and X_l^κ are the STFTs of the right and left channel of the binaural signal at frame κ, respectively. Therefore, ILD is simply the ratio in dB of the amplitudes of the right and left STFTs, i.e., the difference of the amplitudes in dB between the right and left STFTs. When one or both of the $|X^\kappa|$ is null, we consider the interaural differences as invalid, and discard the result. This holds for the ITD also. Nevertheless, a voice activity detector (VAD) is used to decide whether $|X^\kappa|$ is null, like the interaural coherence [15].

Based on the right and left spectra of the κth frame, we define the ITD (in seconds) as:

$$ITD(\kappa,\omega) = \frac{1}{\omega}\Big(\angle\frac{X_r^\kappa(\omega) X_l^\kappa(\omega)}{X_l^\kappa(\omega) X_l^\kappa(\omega)} + 2\pi p\Big), \tag{13}$$

where the integer p is the phase unwrapping factor, which is *a priori* unknown. The use of this factor is necessary by the fact that the angle of the ratio of the spectra is computed modulo 2π. The fact is that an unknown *priori* makes the phase become ambiguous above a certain frequency, which is mainly dependent on the size and shape of the head. This frequency, herein called the ITD ambiguity threshold, can be approximated more precisely with the following equation:

$$f_0 = \frac{c}{r\pi}, \tag{14}$$

where c is the speed of sound in air (344 m/s) and r is the average radius of the head. This threshold is however often averaged to 1500Hz and in the remainder of the paper this value will be used.

In order to retrieve the azimuth of a given frequency bin of the STFT pair, matching the ILD and ITD measurements of that bin to the measured ILD and ITD from the HRTF of the subject.

Since the HRTFs are assumed to be time-invariant, there is no dependency on the time index n. Instead, the HRTFs depend on the directional angle (θ, φ). By using the right and left HRTFs as functions of azimuth and frequency, $HRTF_r^s(\theta,\varphi,\omega)$ and $HRTF_l^s(\theta,\varphi,\omega)$, in place of the signal spectra in (12) and (13). We obtain the HRTF data lookup models of subject s for level difference $ILD^s(\kappa,\omega)$ and time difference $ITD^s(\kappa,\omega)$ as functions of directional angle (θ,φ) and frequency ω:

$$\begin{aligned} ILD^s(\kappa,\omega) &= 20\log_{10}\mathscr{F}\{\lambda^s(\kappa,n)\} \\ &= 20\log_{10}\frac{|HRTF_r^s(\theta,\varphi,\omega)HRTF_l^s(\theta,\varphi,\omega)e^{j\omega\Delta\tau}|}{|HRTF_r^s(\theta,\varphi,\omega)HRTF_r^s(\theta,\varphi,\omega)|} \end{aligned} \tag{15}$$

$$ITD^s(\kappa,\omega) = \frac{1}{\omega}\Big(\angle\frac{HRTF_r^s(\theta,\varphi,\omega)HRTF_l^s(\theta,\varphi,\omega)}{HRTF_l^s(\theta,\varphi,\omega)HRTF_l^s(\theta,\varphi,\omega)} + 2\pi p\Big) \tag{16}$$

Here again, the time difference depends on an arbitrary unwrapping factor p. This ambiguity is resolved by unwrapping the modulo 2π phase difference of the right and left HRTFs along the azimuth. The assumption is that the actual phase difference of the HRTFs does not show substantial hiatus across azimuth. Moreover, the phase unwrapping factor is assumed to be 0 at zero azimuth, where the phase difference should be as small as possible.

Considering the KEMAR dummy head in the CIPIC database [16], the HRTFs are measured for each azimuth with 0° elevation. The ILD and ITD as functions of azimuth and frequency for one particular head are illustrated in Fig. 1. The panels in the upper row show the smoothed ILD and ITD computed from the measured HRTFs. No processing across frequency is performed. Besides, using the averaged ITD and ILD estimates in (6) and (10), we can obtain the averaged ITD and ILD across (θ,φ), which are shown in the lower panels of Fig. 1.

The ITD is usually a relatively smooth function of azimuth, as shown in Fig. 1. This means that the standard deviation of the azimuth estimates based on ITD is relatively small. However, there may be several possible azimuth candidates due to the phase ambiguities in (13). The ILD is a more complex function of azimuth and must be smoothed across azimuth in order to become useful for azimuth lookup. Consequently, the azimuth estimates based on ILDs have a much larger standard deviation than those based on ITDs. In addition, the ILDs as functions of azimuth are not, in general, monotonic for all frequencies. On this occasion, the azimuth lookup is non-unique, yielding multiple possible azimuth estimates. In this condition, we take the azimuth closest to zero. This choice is made for analysis purposes only, since the ambiguity is lifted with the use of ITD in the estimation method shown in [17].

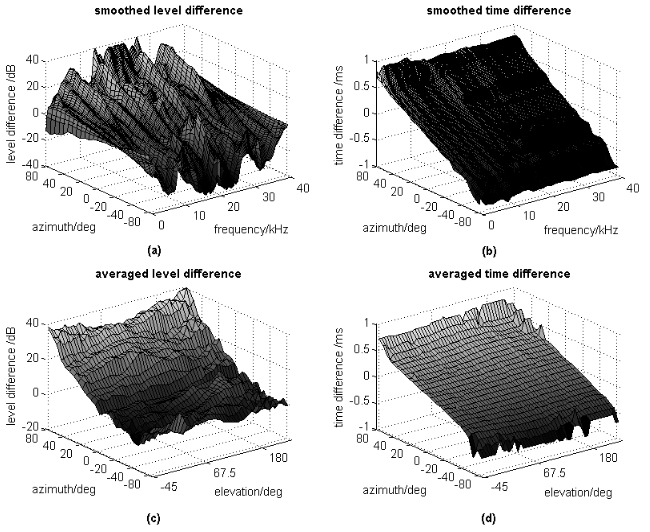

Fig. 1. Binaural cues estimation for KEMAR dummy head in CIPIC HRTF database. Upper: Smoothed HRTF lookup across azimuth. Lower: Averaged binaural cues across (θ,φ).

III. Probabilistic Single Source Localization

A. Azimuth Localization

As for the far field scenario of SSL, e.g., the works in [18], [19], the relationship between the time difference and azimuth is a sinusoidal function formulated as:

$$\sin\theta = \frac{\Delta\tau c}{2rf_s}. \tag{17}$$

Since the time delay $\Delta\tau$ is in sampling number, the time difference in millisecond is transformed by $\Delta\tau/f_s$. Fig. 2 shows an analysis for the time difference estimate using the HRTF lookup. The red dot-line represents the theoretical ITDs of horizontal azimuths, which are calculated by (17). The boxplots denote the corresponding evaluated mean value and variance of each azimuth that is obtained by Fig. 1(d). It is concluded that in the two broad sides of a head, the azimuths have larger variance, which result in more ambiguities for azimuth localization. Given an azimuth θ_i, we can view its ITD subjected to a normal probability density function (PDF) with mean $\mu_{\theta_i,ITD}$ and variance $\sigma_{\theta_i,ITD}$. Thus, when an ITD is evaluated, we can compute the probabilities of all azimuths that the sound source locates using a Gaussian distribution as:

$$p(\theta_i|ITD) = \frac{1}{\sqrt{2\pi}\sigma_{\theta_i,ITD}}\exp\Big(-\frac{(ITD-\mu_{\theta_i,ITD})^2}{2\sigma^2_{\theta_i,ITD}}\Big). \tag{18}$$

At this point, the azimuth of the sound source can be crudely evaluated by $arg\max p(\theta_i|ITD)$. Yet to overcome the

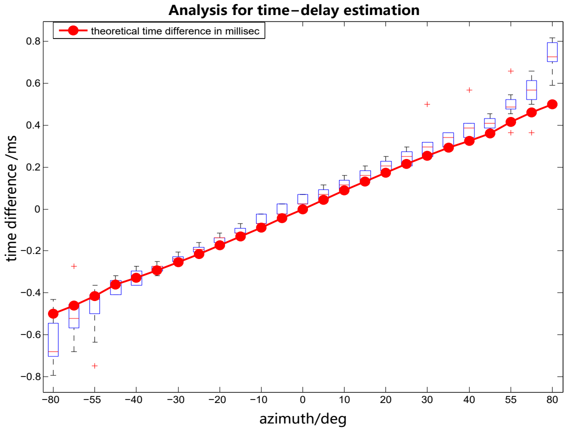

Fig. 2. Statistic analysis of time-delay estimate. The boxplot shows the estimated mean values and variances of time difference. The red line shows the theoretical time difference in millisecond.

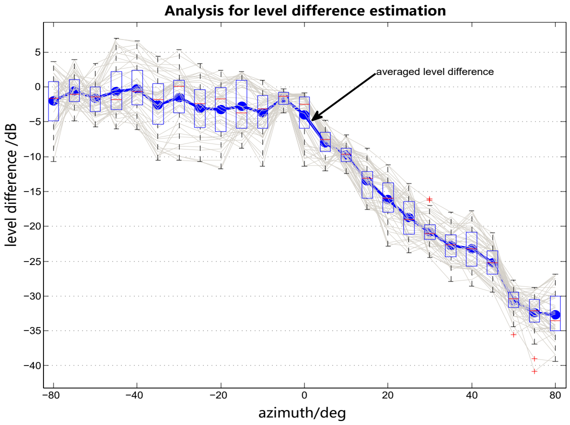

Fig. 3. Statistic analysis of level difference estimate. The boxplot shows the estimated mean values and variances of level difference. The bold blue line denotes the averaged ILD across elevations.

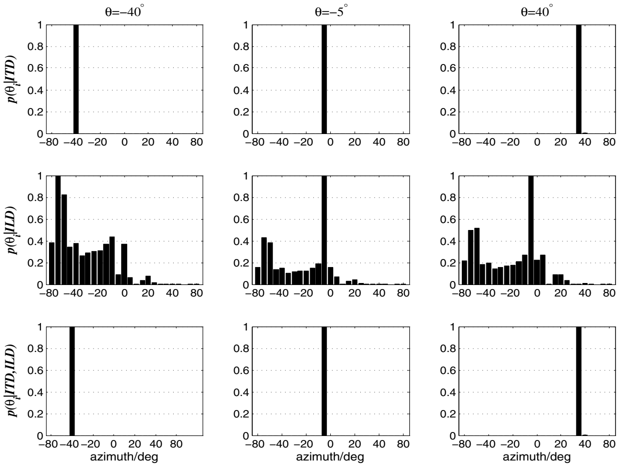

Fig. 4. The procedures of single sound source localization. The three columns express different azimuths, e.g., $-40°$, $-5°$ and $40°$, respectively. First row: The normalized probability of ITD-based localization. Second row: The normalized probability of ILD-based localization. Third row: The normalized probability of the joint of ITD and ILD based localization.

ambiguities of the two broad sides, it needs to consider the influences of ILD. Similarly, we analyze the level difference estimate for the HRTF dataset. The relationship between the ILDs (in dB) and azimuths are drawn in Fig. 3, which is counted from Fig. 1(c). The blue dot-line denotes the averaged ILD, and the boxplot represents the distribution of ILDs corresponding to each azimuth. Therefore, for the azimuth θ_i, a normal PDF of ILDs is obtained with mean $\mu_{\theta_i,ILD}$ and variance $\sigma_{\theta_i,ILD}$. It is observed that due to the effects of pinna, the ILDs do not have a linear distribution versus azimuth. On the left side of the head (i.e., $\theta < 0$), the ILDs attenuate slightly, while On the right side (i.e., $\theta > 0$), they decrease rapidly. In the same way, when an ILD is evaluated, the probabilities of azimuths of ILD-based localization can be computed using a Gaussian function as:

$$p(\theta_i|ILD) = \frac{1}{\sqrt{2\pi}\sigma_{\theta_i,ILD}} \exp\Big(-\frac{(ILD-\mu_{\theta_i,ILD})^2}{2\sigma^2_{\theta_i,ILD}}\Big). \quad (19)$$

In the context of the statistical analysis for conditional probabilities of ITD and ILD, we can get the joint probability of the sound source with given binaural cues, which is mathematically expressed as:

$$p(\theta_i|ITD,ILD) = p(\theta_i|ITD)\cdot p(\theta_i|ILD). \quad (20)$$

Like probabilistic models in [20], [21], we can resolve the joint azimuth localization by:

$$\theta = arg \max_{\theta_i} p(\theta_i|ITD,ILD). \quad (21)$$

Fig. 4 illustrates the conditional probabilistic distributions of azimuth localization in terms of ITD or ILD. The single sound source is located in $\theta \in \{-40°,-5°,40°\}$ of the horizontal plane, respectively. In the panel, the first row denotes the results of ITD-based localization, i.e., $p(\theta_i|ITD)$. The second row shows that of ILD-based, i.e., $p(\theta_i|ILD)$. And the last row depicts that of the joint of ITD and ILD, i.e., $p(\theta_i|ITD,ILD)$. It is concluded that the azimuth determined by the ITD is quite accurate, although we have not considered the different p for the phase unwrapping. That is, the averaged ITD has a strong ability to describe the actual azimuth. While observing the $p(\theta_i|ILD)$, we note that the ILD-based results contain more ambiguous, since ILD has larger variance as Fig. 3 shows. Generally, in the right ahead areas, the ILD can make acceptable decisions. By the joint of ILD and ITD, the results are improved. Although the visual effects of ILD are dispersive, it is essential to select correct p for the phase unwrapping, when we take the influences of frequency on ITD into account.

B. Elevation Localization

The probabilistic idea is referred here. First, the azimuth θ_i of the sound source is taken as a priori. Then, the exponential distance of the level difference is utilized to measure the probability of elevation, e.g.,

$$p(\varphi_j|\theta_i) = \exp(\widetilde{\lambda} - ILD|\theta_i), \quad (22)$$

where $\widetilde{\lambda}$ is obtain by (6), and ILD is looked up from the HRTFs in subsection II-B, such that the elevation is evaluated

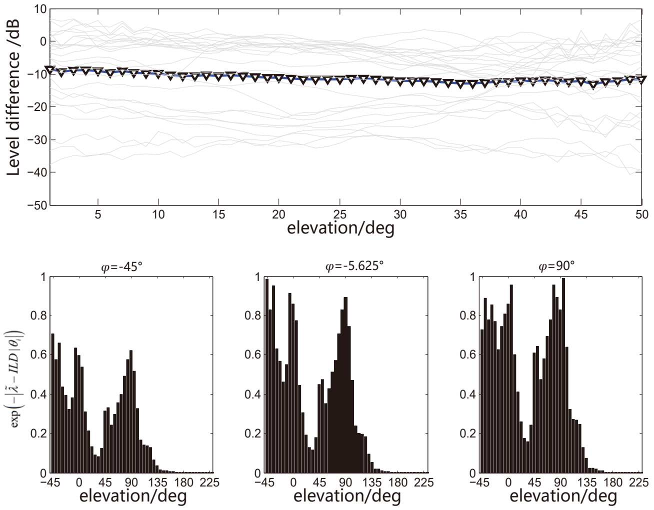

Fig. 5. The upper panel is the spatial distribution of ILD in dB versus elevation. The lower panel includes the probabilities $\exp(\widetilde{\lambda}-ILD|\theta_i)$ when the elevations of the sound source are $-45°$, $-5.625°$, $90°$, respectively.

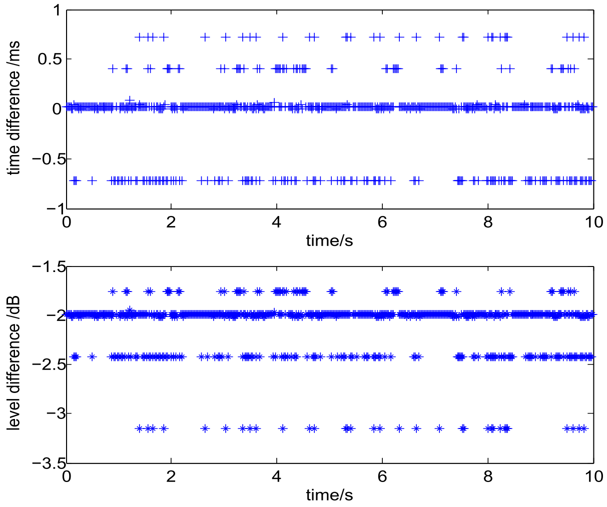

Fig. 6. The distributions of ITDs and ILDs for the binaural audio of 10s. The sound sources are positioned at $-40°$, $-20°$, $0°$ and $40°$, respectively, on the horizontal plane.

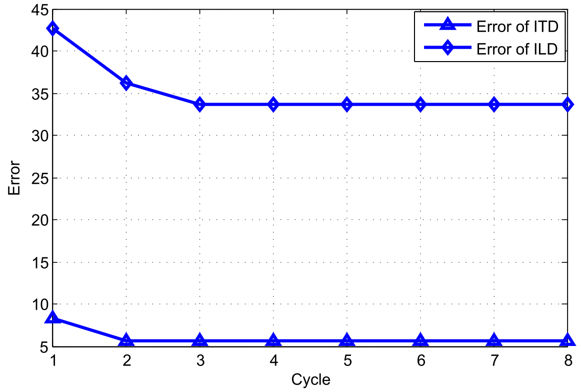

Fig. 7. The iterative errors of k-means for the ILDs and ITDs.

by the $arg\max p(\varphi_j|\theta_i)$, i.e.,

$$\varphi_j = arg\max p(\varphi_j|\theta_i). \tag{23}$$

In order to find out the performance of elevation localization, some realistic localization examples are displayed in Fig. 5. The upper panel is the spatial distribution of ILD in dB versus elevation, in which the shallow lines represent different azimuths, and the dark-triangle line is the averaged ILD. It can be observed that ILD is mainly affected by the azimuth and fluctuated slightly along with the elevation. In the lower panel, the elevations of the sound source are $-45°$, $-5.625°$, $90°$, respectively, and the subplots give the probability $\exp(\widetilde{\lambda}-ILD|\theta_i)$ of the three cases. We deduce that this method is generally useful for the elevations. Because the $\exp(\widetilde{\lambda}-ILD|\theta_i)$ has an obvious peak when the elevation are $-45°$, $-5.625°$, $90°$. Generally speaking, elevation localization is much more difficult than that of azimuth, because the ITD is only related by azimuth and ILD is partially related by the elevation.

IV. Multiple Source Localization

In practice, there are usually multiple sound sources needed to be localized, such as the "multi-person conversations" scenario. For the multi-sources localization, such as the HRI systems and video conference, azimuth is more important than the elevation in general. Therefore, the azimuth is the main focus in this section. It is severely difficult to realize merely by one frame of speech. Mostly, the azimuths of the sources are evaluated statistically from a long period of binaural audio. So we obtain a result (i.e., azimuth) from each frame by the binaural cues, and the final outcomes are regarded as the values that occur frequently.

We cluster the binaural cues obtained by the time-delay compensation estimator for all the speech frames first. For instance, Fig. 6 shows the binaural cues of four sources, which are positioned at $-40°$, $-20°$, $0°$ and $40°$, respectively, on the horizontal plane. The binaural audio is convolved by the KEMAR HRTFs with a Gaussian white noise. In Fig. 6, it can be obviously concluded that the ITDs are converged to four values (i.e., 0.75 *ms*, 0.4 *ms*, 0 *ms* and -0.75 *ms*), and the ILDs are also converged to four values (i.e., -1.75 dB, -2 dB, -2.45 dB and -3.2 dB). Note that the largest amounts of ITDs and ILDs are converged to 0 *ms* and -2 dB, respectively. That is to say, the source coming from the direction right ahead of a robot is easiest to be detected. Besides, when the sound sources last beyond 2 *s*, the binaural cues can show the four actual centers.

Consequently, we present an effective algorithm to manage converging the binaural cues. Fig. 7 illustrates the iterative errors of ITDs and ILDs using the popular k-means. For both the binaural cues, k-means can find out the centers within 4 steps. After that, we can exploit the aforehand proposed single SSL method to estimate the positions of the sources by the centers of binaural cues. Our multiple source localization strategy based on k-means is quite simple and high efficiency for the applications.

V. Experiments and Discussions

The CIPIC database [16] used in our experiments is measured by the U. C. Davis CIPIC Interface Laboratory, which

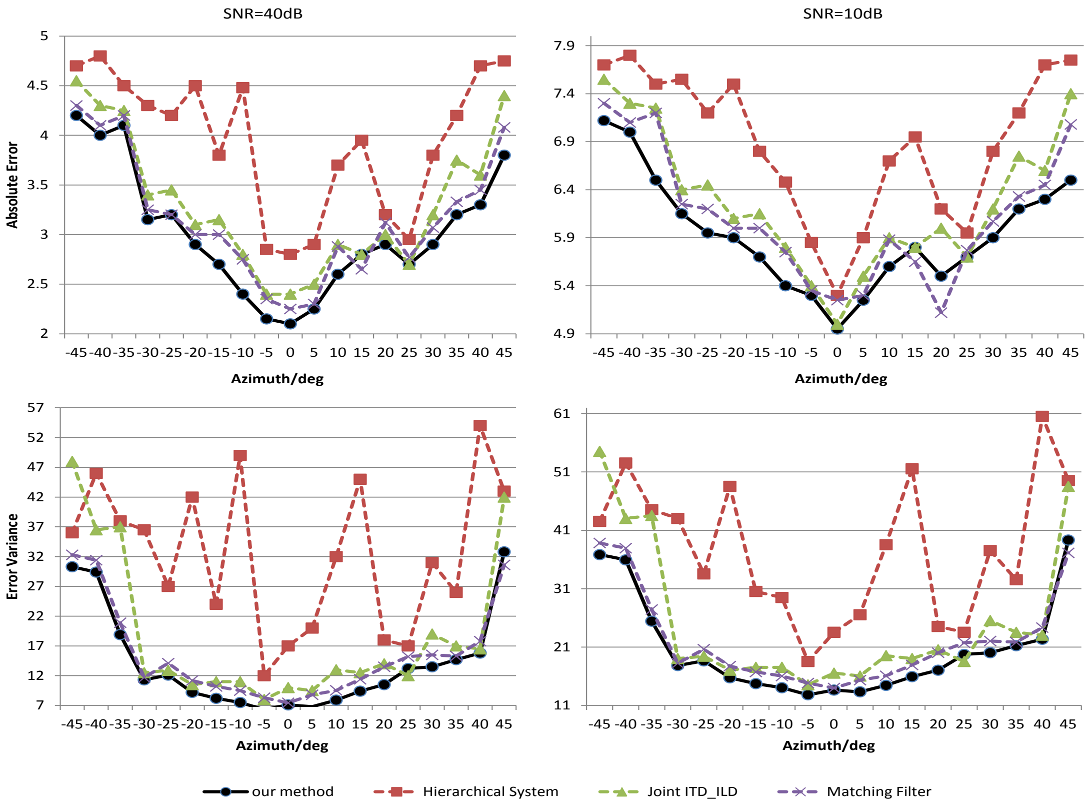

Fig. 8. The azimuthal results of two different SNRs (Left panel: 40dB, Right panel:10dB, Upper panel: Absolute Error, Lower panel: Error variance).

includes the HRTFs for 45 different subjects (i.e., 27 males, 16 females, and KENAR with large and small pinna). The HRTFs are tested at $1m$ distance with 25 different azimuths, 50 different elevations resulting in totally 1250 directions for each subject. The sound sources used in experiments are musical signals. The period of each sound for training and localization is 2 seconds and the sampling frequency is 44.1 kHz. We compare our method with Hierarchical System [20], Joint ITD_ILD [17] and Matching Filter [18]. The first algorithm is a three-layer framework, in which the ITD, ILD and spectral cues are used in the three layers, respectively. The second one puts forward a joint estimation of ILD and ITD for both single and multiple source localization issue. And the other one involves a new proposed binaural cue named interaural matching filter (IMF) for a hierarchical localization. We compare our method with them, since all of them refer the probabilistic idea.

In this paper, the results for test sets are based on different signal parts at $45 \times 1250 \times 100 \times 128 \times 5$, which means 45 subjects, 1250 directions, 100 sound signals and 5 sound activities processed over 256 sample points, which is also the length of the window. Our method is validated in different SNRs and with several different sound activities.

A. Azimuth Accuracy

We testify the performances of directional decision for a single sound source depending on the subject #003 in the CIPIC HRTFs. Here the central azimuths belong to $[-80°, -65°, -55°, -45° : 5° : 45°, 55°, 65°, 80°]$ in the frontal plane. The localization accuracy varies from the azimuth for the several compared methods are in Fig. 8. The left panel illustrates the azimuthal absolute errors and the respective error variances when the environmental SNR is 40dB, and the right panel shows the two norms as SNR is 10dB. These two different scenarios represent the typical office environments.

Our method (black-dot lines) gives much more consistent results, mimicking the behavior of the human perceptual behavior with more accuracy around frontal directions. The error variance stays very quiet and stable between -30° and 30°. As the SNR rising, our method still remains the best performance. On the other hand, the values for absolute error and error variance of others are generally higher than our method. It is obvious that the hierarchical system shows the widest fluctuations, both with respect to the absolute errors and error variance. And the other two algorithms also lag to ours. To seek the reasons, the proposed time-delay compensation estimator can offer robust binaural cues estimations, and the probabilistic localization strategy provides accurate azimuthal matching results. While the accurate binaural cues are hard to extract for the others, especially for the matching filter, which compute relative transfer functions as new localization cue and its design is influenced by the noise severely.

B. Elevation Accuracy

The elevation accurately of the source is observed and compared in this paper. In the database, the elevations vary

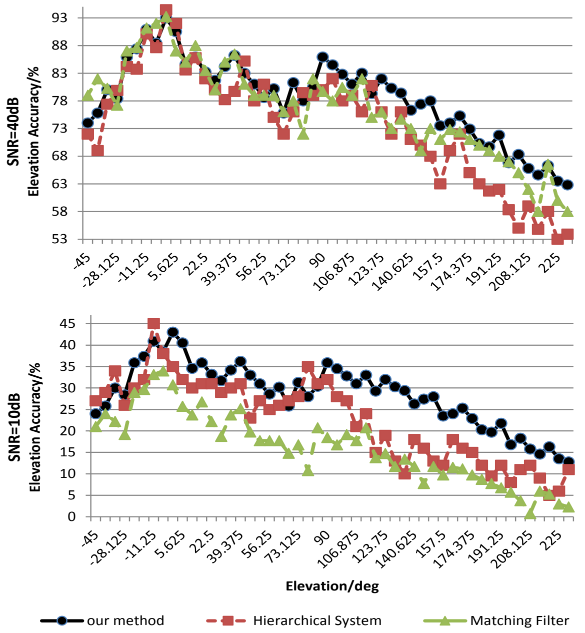

Fig. 9. The elevation localization accuracy of two different SNRs (Upper one: SNR=40dB, Lower one: SNR=10dB).

from $-45°$ to $230.625°$ in step of $5.625°$. The azimuth is fixed to 0°. Here we have omitted the Joint ITD_ILD for the elevation is fixed to 0° in their works [17], [22].

While putting a sound source at each direction and localizing repeatedly, the accuracy of elevation is shown as Fig. 9. In general, as the noise is slighter (SNR=40 dB), the results are quite acceptable. However, the accuracy fades rapidly as the noise rising. Most importantly, the overall results drop behind that of azimuth much. It is because that the elevation localization is much difficult than that of azimuth as proved in the subsection III-B. Generally speaking, our method has achieved the better result generally. Moreover, it is better to localize the same elevations (i.e., the frontal plane of the dummy head).

C. Performances of Sound Activities

We also test the dependence of the type of sound (instrument) on accuracy of the azimuth localization. In this experiment, individual instrumental sounds taken from the IOWA database [23] are panned to the azimuth angles using each of the 45 HRIR measurements of the 45 subjects of the CIPIC database. The mean absolute errors and the variances of the error are computed over the 25 angles. Thus the final results are averaged absolute errors and error variance over all the CIPIC subjects and sounds for a given IOWA instrumental sound class. Additionally, the environmental SNR is set to be 40 dB.

In total, more than 3000 instrumental sounds are used. Table I lists the results of this experiment. It can be seen that our method outperforms the results of the others, both for the absolute error and its variance generally. The errors are usually less (mostly under 5 degree) except for the double bass which could be explained by the fact that this instrument is low pitched and the tones have little energy in higher harmonics, which implies large errors in the ILD. Among the four algorithms, the matching filter obtains the suboptimal solution due to its superior ability to represent the relative transfer functions. The hierarchical system gets the worst performance because its binaural cues cannot recognize the influences by the frequency.

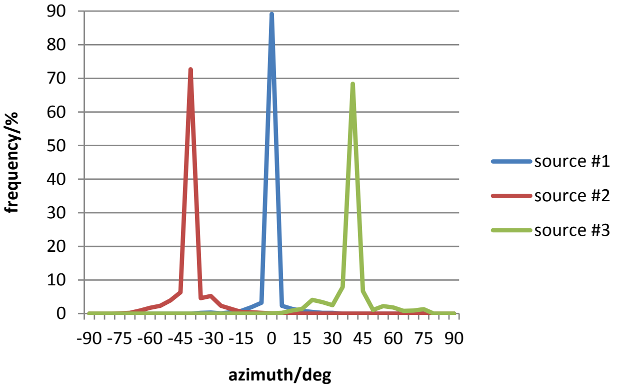

Fig. 10. The frequency of multiple sources localization for 500 times. The sources are located at $-40°$, 0°, 40°, respectively.

D. Accuracy of Multiple Sources Localization

To observe performances of multiple sources localization, the KEMAR HRTFs are take in the reminder of this section, the horizontal azimuth evenly divided into 32 directions with the step of 5°. We put three sound sources at $-40°$, 0°, 40°, respectively, and make the sources uttering continuously. By running multiple sources localization algorithm repeatedly for 500 times, the distribution of frequency of the sources are drawn as Fig. 10 shows. Based on *arg* maximizing the frequency functions, the locations of the sources are solved. As a consequence, we can conclude that the sources are well detected. Note that we achieve the best accuracy when the source at 0°, and the directions in the left or right side has a few mistakes as localized to the adjacent azimuths.

VI. Conclusions and Feature Works

In this work, a probabilistic binaural sound source localization method based on time-delay compensation (TDC) estimator and clustering analysis is proposed. The TDC estimator operates the binaural cues (including ITD and ILD) simultaneously, which is quite robust to work in noisy environments. For the single sound localization, we propose a probabilistic strategy for the azimuth as well as elevation based on the Gaussian distributions to match the lookup of the binaural cues. When multiple sources utter, the *k*-means clustering is involved to analyze in terms of the binaural audio containing multiple frames. Consequently, our method improves the performance of the directional decisions, and it also works for many sound activities localization task. In the future, we will try to apply it to the realistic robotic systems.

TABLE I

THE MEAN LOCALIZATION ERROR(VARIANCE) IN DEGREE OF THE FOUR COMPARED METHODS FOR DIFFERENT INSTRUMENTAL SOUND ACTIVITIES.

Sound	index	**Our method** (σ^2)	Joint ITD_ILD (σ^2)	Hierarchical system (σ^2)	Matching filter (σ^2)
D. Bass (arco)	289	**10.93(32.89)**	11.34(33.01)	11.48(42.77)	9.83(22.43)
D. Bass (pizz)	300	**15.20(48.30)**	15.24(52.58)	15.30(65.94)	15.23(49.54)
Bassoon	121	**3.10(1.93)**	3.24(2.01)	4.43(5.90)	3.28(2.59)
Cello (arco)	347	**4.20(5.32)**	4.23(6.00)	5.24(7.99)	4.34(5.90)
Cello (pizz)	330	**4.92(8.50)**	5.15(9.24)	6.17(14.87)	5.08(9.18)
Eb Clarinet	119	**3.78(5.34)**	4.74(6.84)	6.76(12.64)	4.10(5.92)
Bb Clarinet	139	**5.10(7.12)**	5.11(7.90)	6.95(12.67)	5.89(8.94)
Alto Flute	99	**3.78(3.84)**	4.29(4.48)	5.86(8.63)	4.19(4.21)
Flute	227	**4.11(6.35)**	4.57(6.67)	6.23(10.61)	4.75(6.82)
Bass Flute	102	**3.79(2.67)**	4.40(5.51)	5.81(9.76)	4.08(3.32)
Horn	96	**2.89(1.99)**	3.25(2.07)	4.43(6.44)	3.04(2.00)
Oboe	104	**4.88(5.92)**	5.13(6.68)	6.50(8.72)	4.89(6.05)
Piano	260	**4.17(7.19)**	4.58(7.76)	5.14(8.23)	4.23(7.65)
Soprano Sax	129	**3.25(4.16)**	4.33(4.36)	6.29(9.59)	3.97(4.29)
Alto Sax	192	**2.93(3.22)**	3.85(3.74)	5.86(11.70)	3.14(3.46)
Bass Trombone	131	**2.87(1.03)**	3.35(1.85)	4.04(3.89)	2.94(1.57)

VII. ACKNOWLEDGMENT

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), Specialized Research Fund for the Doctoral Program of Higher Education (No.20130001110011), Natural Science Foundation of Guangdong Province(No.2015A030311034), Science and Technology Innovation Commission of Shenzhen Municipality (No.JCYJ20130331144631730, No.JCYJ20130331144716089).

REFERENCES

[1] D. B. Ward and R. C. Williamson, “Particle filter beamforming for acoustic source localization in a reverberant environment,” *in IEEE International Conference onAcoustics, Speech, and Signal Processing.*, vol.2, pp. II–1777, 2002.

[2] C. T. Ishi, J. Even and N. Hagita, “Using multiple microphone arrays and reflections for 3d localization of sound sources,” *in IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS).*, pp. 3937C3942, 2013.

[3] A. Lombard, Y. H. Zheng, H. Buchner and W. Kellermann, “TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis,” *in IEEE Transactions on Audio, Speech, and Language Processing.*, vol. 19, pp. 1490–1503, 2011.

[4] 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 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS).*, pp. 2879-2886, 2012.

[5] N. Roman and D. L. Wang, “Binaural tracking of multiple moving sources,” *in IEEE Transactions on Audio, Speech, Language Process.*, vol. 16, no. 4, pp. 728–739, 2008.

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

[7] R. F. Lyon and C. Mead, “An analog electronic cochlea,” *in IEEE Transactions on Acoustics, Speech and Signal Processing.*, vol. 36, no. 7, pp. 1119–1134, 1988.

[8] C. Baumann and C. Rogers and F. Massen, “Dynamic binaural sound localization based on variations of interaural time delays and system rotations,” *in Journal of the Acoustical Society of America.*, vol. 138, no. 2, pp. 635-650, 2015.

[9] H. Jonas, L. Manuel, S. V. Jose and L, Francisco, “Sound Localization for Humanoid Robots - Building Audio-Motor Maps based on the HRTF,” *in IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS).*, pp. 1170-1176, 2006.

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

[11] S. T. Birchfield and R. Gangishetty, “Acoustic localization by interaural level difference,” *in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, vol. 4, pp. 1106–1109, 2005.

[12] H. Liu and J. Zhang, “A novel binaural sound source localization model based on time-delay compensation and interaural coherence,” *in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, pp. 1438–1442, 2014.

[13] P. R. Roth, “Effective measurements using digital signal analysis,” *in IEEE Spectrum.*, vol. 8, no. 4, pp. 62–70, 1971.

[14] C. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *in IEEE Transactions on Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[15] M. Jeub and P. Vary T. E. Scheferand, “Model-based dereverberation preserving binaural cues,” *in IEEE Transactions on Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1732–1745, 2010.

[16] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, “The cipic hrtf database,” *in IEEE Workshop Appl. Signal Process. Audio, Acoust (WASPAA).*, pp. 99–102, 2001.

[17] M. Raspaud, H. Viste, and G. Evangelista, “Binaural source localization by joint estimation of ILD and ITD,” *in IEEE Transactions on Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 68–77, 2010.

[18] H. Liu, J. Zhang, and Z. Fu, “A new hierarchical binaural sound source localization method based on interaural matching filter,” *in IEEE International Conference on Robotics ans Automation (ICRA).*, pp. 1598–1605, 2014.

[19] J. Zhang and H. Liu, “Robust Acoustic Localization via Time-Delay Compensation and Interaural Matching Filter,” *in IEEE Transactions on Signal Processing.*, vol. 63, no. 18, pp. 4771-4783, 2015.

[20] L. Danfeng and E. L. Stephen, “A bayes-rule based hierarchical system for binaural sound source localization,” *in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, pp. 521–524, 2003.

[21] J. Woodruff and D. L. Wang, “Binaural localization of multiple sources in reverberant and noisy environments,” *in IEEE Transactions on Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1503–1512, 2012.

[22] H. Viste and G. Evangelista, “Binaural source localization,” *in International Conference on Digital Audio Effects.*, no. 29, pp. 145–150, 2004.

[23] L. Fritts, “The iowamusic instrument samples,” [online] Available at: http://theremin.music.uiowa.edu, 1997.