Enhancing Loudspeaker-based 3D Audio with Room Modeling

Myung-Suk Song [#1], Cha Zhang [*2], Dinei Florencio [*3], and Hong-Goo Kang [#4]

[#] *Department of Electrical and Electronic, Yonsei University*
Yonsei University 134 shinchondong seodaemoon-gu, 120-749, Seoul, Korea
[1] earth112@dsp.yonsei.ac.kr [4] hgkang@yonsei.ac.kr

[*] *Microsoft Research*
One Microsoft Way, Redmond, WA 98052, USA
[2] chazhang@microsoft.com [3] dinei@microsoft.com

***Abstract*—For many years, spatial (3D) sound using headphones has been widely used in a number of applications. A rich spatial sensation is obtained by using head related transfer functions (HRTF) and playing the appropriate sound through headphones. In theory, loudspeaker audio systems would be capable of rendering 3D sound fields almost as rich as headphones, as long as the room impulse responses (RIRs) between the loudspeakers and the ears are known. In practice, however, obtaining these RIRs is hard, and the performance of loudspeaker based systems is far from perfect. New hope has been recently raised by a system that tracks the user's head position and orientation, and incorporates them into the RIRs estimates in real time. That system made two simplifying assumptions: it used generic HRTFs, and it ignored room reverberation. In this paper we tackle the second problem: we incorporate a room reverberation estimate into the RIRs. Note that this is a non-trivial task: RIRs vary significantly with the listener's positions, and even if one could measure them at a few points, they are notoriously hard to interpolate. Instead, we take an indirect approach: we model the room, and from that model we obtain an estimate of the main reflections. Position and characteristics of walls do not vary with the users' movement, yet they allow to quickly compute an estimate of the RIR for each new user position. Of course the key question is whether the estimates are good enough. We show an improvement in localization perception of up to 32% (i.e., reducing average error from 23.5° to 15.9°).**

I. Introduction[1]

Audio spatialization refers to techniques that synthesize a virtual sound image in order for the listener to feel as if the signals are originated by an actual source located at a certain position [1]. Spatialized audio can be rendered by headphones or loudspeakers. The latter relieves the user from wearing a headset, and is thus widely appreciated in home and desktop environments. However, spatialization performance of loudspeaker-based systems is significantly inferior.

One challenge in loudspeaker-based audio spatialization is the crosstalk caused by the contralateral paths from the loudspeakers to the listener's ears, which often damages the 3D cues of the spatialized audio. Crosstalk cancelation techniques have been studied to eliminate or minimize the crosstalk [2], [3], [4].

To perform crosstalk cancelation well, one must accurately model the acoustic path from the loudspeaker to the listener's position. Several methods to model the transfer functions have been proposed in the literature [5]. The simplest scheme is to use a free-field model: the sound field radiated from a monopole in a free-field is computed based on the distances from the sources to the observation points. Under the assumption that the human head can be modeled as a sphere, the expression for the sound field produced by a sound wave impinging on a rigid sphere has been formulated in [6]. An improvement over the spherical head model is to adopt the head related transfer function (HRTF) [7]. The HRTF is often measured in an anechoic chamber with dummy-heads to provide an acoustically realistic model of a human listener.

An additional hurdle is that crosstalk cancelation has to be done for the *current* position of the user's head. In other words, the HRTF has to be further composed with direct path delay and attenuation of the sound wave. Only then one can calculate more accurate transfer functions between the loudspeakers and the listener and use them for crosstalk cancelation [8]. For instance, early works [9], [10] used intrusive electromagnetic trackers to demonstrate this idea. Song et al. [11] recently built a binaural loudspeaker audio system based on a 3D face tracker with a single webcam, and showed experimentally that tracking and HRTF can indeed help improve the listener's 3D audio perception.

Finally, a major hurdle is reverberation. Real-world environments are often reverberant, which creates some additional challenges for crosstalk cancelation. Kyriakakis and Holman [12] noted that the performance of conventional crosstalk cancelation systems degrades in a realistic listening room in which reverberation exists in general. They proposed a solution that changes the layout of the system to ensure the direct path is dominant, so that it reduces the effect of reverberation. Unfortunately, even this small improvement is not always possible, as layout changes may not be practical. In theory,

MMSP'10, October 4-6, 2010, Saint-Malo, France.
978-1-4244-8112-5/10/$26.00

[1] Work performed during the first author's internship at Microsoft Research.

a better solution would be to take the room reverberation into consideration when computing the transfer functions between the loudspeakers and the listener. For instance, in [13], Lopez et al. used room impulse responses measured by dummy head to help crosstalk cancelation in reverberant rooms and it significantly improved results. However, that result is mostly a proof of concept: room impulse response is very difficult to measure, varies significantly with the head position and orientation relative to the room environment, and is notoriously hard to interpolate. For those reason, there has been no practical work that takes room reverberation into consideration during crosstalk cancelation.

In this paper, we explore a novel scheme for crosstalk cancelation, which explicitly considers room reverberation by modeling the room with a number of planar reflectors such as walls or ceilings. These models can be estimated with approaches such as [14]. In other words, instead of directly trying to measure the room impulse response (RIR), we model the room and obtain the RIR from that model. Indeed, while the RIR changes with every little movement of the user's head, an estimate of the new RIR can always be quickly obtained from the model. A similar approach has been applied to sound source localization, with dramatic improvements [15], [16]. By using the estimated RIR, and applying an equalization technique, we can improve the channel separation and thus the user's experience. The main question is, of course, whether such an estimated RIR is accurate enough (i.e., close to the true RIR) to yield reasonable results. Our experimental results clearly show that the estimated RIR is calculated close enough to the true RIR. By applying an equalization technique to the estimated acoustic transfer function that includes the reflections caused by the walls/ceilings of the listening room, we improve the listener's performance on estimating the virtual source position. At the center position, our subjective listening tests showed an improvement in localization perception of 32%, i.e., reducing the average error from 23.5° to 15.9°. Overall, the average improvement was 16%.

The remainder of this paper is organized as follows. Section II introduces conventional binaural audio systems. In Section III, a room-model based binaural audio system is investigated. Experimental results and conclusions are given in Section IV and V, respectively.

II. Conventional Binaural Audio Systems

The block diagram of a conventional binaural audio playback system with two loudspeakers is shown in Figure 1. Component $\mathbf{C}$ represents the transmission path or acoustic channel between the loudspeakers and the listener's ears. The binaural audio system consists of two major blocks: binaural synthesizer $\mathbf{B}$ and crosstalk canceller $\mathbf{H}$. The goal of the binaural synthesizer is to produce the sounds that should be heard by the listener's ear drums. In other words, we hope the signals at the listener's ears e_L and e_R shall be equal to the binaural synthesizer output x_L and x_R. Thus, the objective of the crosstalk canceller is to equalize the effect of the transmission path $\mathbf{C}$. If $\mathbf{C}$ is known or can be estimated, this can be achieved by taking an inversion of the $\mathbf{C}$ matrix [1][2].

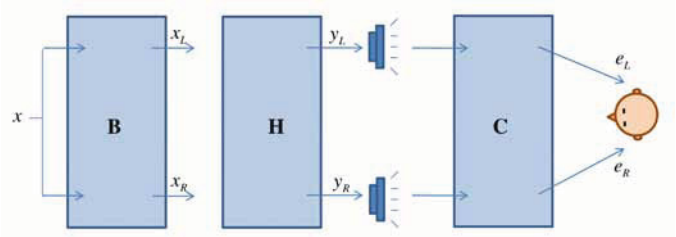

Fig. 1. Diagram of a typical binaural audio system with loudspeakers.

A. Binaural Synthesis

The binaural synthesizer synthesizes one or multiple virtual sound images at different locations around the listener using 3D audio cues. There are a number of well-known cues for the human auditory system to localize sounds in 3D, such as the interaural time difference (ITD), the interaural intensity difference (IID), and the directional filtering induced by the ear shape. In this paper, we follow the work in [8], [17] and use HRTF to synthesize binaural signals from a monaural source. When properly computed, the HRTF incorporates elements of all three 3D cues mentioned above. Specifically, one can filter the monaural input signal with the impulse response of the HRTF (for a given distance and angle of incidence) as:

$$\mathbf{x} = \begin{bmatrix} x_L \\ x_R \end{bmatrix} = \begin{bmatrix} B_L \\ B_R \end{bmatrix} x = \mathbf{B}x, \tag{1}$$

where x is the monaural input signal, B_L and B_R are the HRTFs between the listener's ears and the desired virtual source. The outputs of the binaural synthesizer x_L and x_R are the signals that should be reproduced at the listener's ear drums.

B. Crosstalk Cancelation

The acoustic path between the loudspeakers and the listener's ears (Figure 2) is defined as the acoustic transfer matrix $\mathbf{C}$:

$$\mathbf{C} = \begin{bmatrix} C_{LL} & C_{RL} \\ C_{LR} & C_{RR} \end{bmatrix}, \tag{2}$$

where C_{LL} is the transfer function from the left speaker to the left ear, and C_{RR} is the transfer function from the right speaker

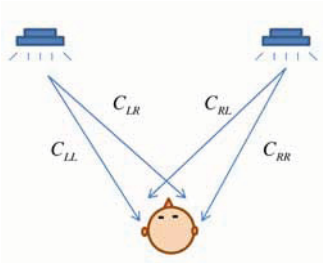

Fig. 2. Acoustic path between two loudspeakers and the listener's ears.

to the right ear. C_{RL} and C_{LR} are the transfer functions from contralateral speakers, which are called "crosstalks". For headphone applications, the two channels are completely separated, hence both C_{RL} and C_{LR} are zero. The binaural synthesis step alone is sufficient to generate great 3D auditory experiences for the user. However, for loudspeaker applications, the crosstalks will destroy the 3D cues of the binaural signal. We need to insert a crosstalk canceller to equalize the transmission path between the loudspeakers and the listener.

The crosstalk canceller matrix $\mathbf{H}$ can be calculated by taking the inverse of the acoustic transfer matrix $\mathbf{C}$.

$$\begin{aligned}\mathbf{H} = \mathbf{C}^{-1} &= \begin{bmatrix} C_{LL} & C_{RL} \\ C_{LR} & C_{RR} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} C_{RR} & -C_{RL} \\ -C_{LR} & C_{LL} \end{bmatrix} \frac{1}{D},\end{aligned} \tag{3}$$

where D denotes the determinant of the matrix $\mathbf{C}$. Note that we assume the listener's head position is known, e.g., given by a tracker [13], [17], [19]. In addition, since the acoustic transfer functions derived from the HRTFs have non-minimum phase characteristic, it is generally unstable to compute $\mathbf{H}$ by directly inverting of $\mathbf{C}$. Instead, $\mathbf{H}$ can be adaptively obtained by the least mean square (LMS) method [3], [18].

III. Enhanced Binaural Audio System with Room Modeling

As mentioned in the introduction, real-world environments are often reverberant, which complicates the computation of the acoustic transfer matrix $\mathbf{C}$. To include the indirect paths from the loudspeakers to the listener, the room impulse response must be carefully measured, as was done in [13]. However, the room impulse response may vary significantly as the listener moves around, which renders such measurement based schemes highly impractical.

As noted in [12], the main impact of reverberation on sound quality in immersive audio systems is due to discrete early reflections. Psychoacoustic experiments have confirmed that early reflections are the dominant source of frequency response anomalies when all other factors being equal [20]. In this paper, we propose to model such early reflections explicitly using a simplified room model. The key benefit of our approach over the measurement based approach is its capability to handle moving listeners: the early reflections can be computed through the image method [21] given the listener's position at any instance.

A. The Room Model

Rooms are diverse, and potentially complex environments. They may contain furniture, people, partial walls, doors, windows, nonstandard corners, etc. However diverse, they do have a few things in common. For instance, almost every room has four walls, a ceiling and a floor; the floor is leveled, and the ceiling almost always parallel to the floor; most walls are vertical, straight, and extend from floor to ceiling and from adjoining wall to adjoining wall. For personal binaural systems on the desktop, the two speakers are often placed on an office table, about 90 cm high. In addition, many objects that seem visually important are small enough that may actually be acoustically transparent for most frequencies of interest. Based on these observations, we model a typical room with 6 planar surfaces: 4 walls, the ceiling and the floor (or the table if the main reflection from below is due to the table).

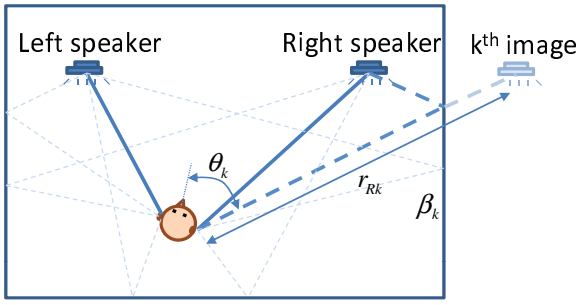

Fig. 3. Acoustic paths between two loudspeakers and the listener's ears with reflections.

Each planar surface is modeled with its 3D position and reflection coefficient. Their positions and the reflection coefficients can be measured or estimated by a number of methods. Several active 3D estimation methods from computer vision can be used [22], but do not provide estimates for reflection coefficients. To obtain estimates of reflection coefficients, acoustic measurements have to be performed. Again, several algorithms have been proposed for automatic acoustic room measurements. For instance, Ba et al. [14] actively probed the room by emitting a known signal, and estimated the room geometry and reflection coefficients by examining the received reflections of a compact microphone array. O'Donovan [23] used a 32-microphone spherical array to visualize the location of sound reflections. Antonacci and Aprea [24], [25] used a single microphone and either a moving source on a circular trajectory or multiple sources to estimate the coordinates of reflectors. Moebus [26] used MVDR beamforming with a single ultrasound transmitter/receiver pair mounted on a precision 2D positioning system to perform ultrasound imaging in air, with which the position and outline of obstacles can be determined.

While any of these methods could be used, in the following we assume a simplified planar room model similar to [14] is obtained for the test environment.

B. Enhanced binaural audio system with room modeling

When a sound source is placed inside a room, the sound wave at an arbitrary location can be represented by the superposition of a number of reflected sound waves. If the room contains only planar surfaces, the reflections can be modeled as direct sounds from various image sound sources, which are placed on the far side of the walls surrounding the real source [21].

As shown in Figure 3, the acoustic paths from the loudspeakers to the listener's ear drums can be represented by the summation of the impulse responses from the actual source and the imaged sources reflected by the walls surrounding the

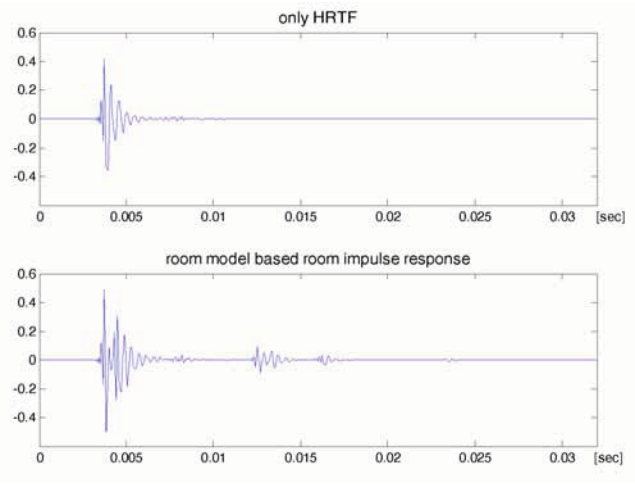

Fig. 4. Room impulse response computed with HRTF only and with the proposed scheme based on room modeling.

listener:

$$C_{mn} = \sum_{k=0}^{N} \frac{\beta_k}{r_{mk}} z^{-\Delta_{mk}} C_{mn}(\theta_k), \ m, n \in \{L, R\} \tag{4}$$

where N is the total number of planar surfaces, k denotes the index of the images of the loudspeakers, and the actual loudspeaker is represented as $k = 0$. m and n represent the indices for the left or right loudspeakers and left or right listener's ears, respectively. β_k, r_{mk}, and Δ_{mk} denote the reflection coefficient for the k^{th} wall, the distance between the k^{th} image of the m speaker and the listener, and the delay from the k^{th} image of the m speaker to the listener, respectively. $\Delta_{mk} = \frac{r_{mk}}{c}$, where c is the speed of sound. Note we assume the head size is much smaller than the distance between the image sources and the listener, hence both ears share the same r_{mk}. $C_{mn}(\theta_k)$ is the HRTF from the k^{th} image of m speaker to n ear. For instance, $C_{LL}(\theta_k)$ is the HRTF of the k^{th} image of the left speaker to the left ear. Finally, note that only first reflections of the walls are considered in this paper, although extending to multiple reflections is straightforward.

In short, the acoustic transfer function from m speaker to n ear is the summation of $C_{mn}(\theta_k)$ weighted by β_k, delayed by Δ_{mk}, and attenuated by distance r_{mk}. The overall acoustic transfer matrix $\mathbf{C}$ can be written as:

$$\mathbf{C} = \begin{bmatrix} \sum_{k=0}^{N} \frac{\beta_k}{r_{Lk}} z^{-\Delta_{Lk}} C_{LL}(\theta_k) & \sum_{k=0}^{N} \frac{\beta_k}{r_{Rk}} z^{-\Delta_{Rk}} C_{RL}(\theta_k) \\ \sum_{k=0}^{N} \frac{\beta_k}{r_{Lk}} z^{-\Delta_{Lk}} C_{LR}(\theta_k) & \sum_{k=0}^{N} \frac{\beta_k}{r_{Rk}} z^{-\Delta_{Rk}} C_{RR}(\theta_k) \end{bmatrix} \tag{5}$$

An example of room impulse response calculated based on the proposed room model method is shown in Figure 4. Based on this calculated room impulse response, we then compute the crosstalk canceller matrix $\mathbf{H}$ using the LMS method as in [3].

IV. Experimental Results

As we have mentioned before, there is no doubt that accounting for the room impulse response will improve results. However, since RIR varies significantly as the listener moves around, none of the existing approaches have successfully demonstrated the benefit of room modeling in practical environments. The key question we would like to answer is whether the proposed RIR estimation based on a room model is good enough to be useful in enhancing 3D sound perception. In order to evaluate the effectiveness of the proposed room model based scheme for crosstalk cancelation, we conducted subjective listening tests to compare it against the conventional methods where no room reverberation is considered [11]. Both systems adopted a webcam to perform 3D face tracking in order to obtain the listener's head position and orientation. The focus of the test is the listener's capability to accurately tell the directions of a set of virtual sound sources.

A. Test Setup

In our listening tests, the subjects were asked to identify the sound source directions between $-90°$ and $90°$ in azimuth, as shown in Figure 5. The two loudspeakers were located at $\pm 30°$, respectively. The virtual sound images were rendered at 10 pre-specified locations: $-90°$, $-75°$, $-60°$, $-45°$, $-30°$, $0°$, $15°$, $45°$, $60°$, $75°$, and $90°$. The distances from the center listening position to the loudspeakers and the virtual sound sources are about 0.6 m.

The subjects were asked to report their listening result on an answer sheet by indicating the sound source direction freely on the semi-circle. The presentation of the test signals and logging of the answers were controlled by the listener. Sound samples were played randomly and repetitions were allowed in all the tests. The monaural signal was a game-like stimulus, consisting of 5 sub-stimuli with 150 ms silent interval. The sub-stimulus was a pink noise with 16 kHz sampling rate. It was played 5 times in 25 ms duration with 50 ms silent interval.

Nine subjects participated the subjective study. Each subject was tested at 3 different positions: center, 20 cm to the left, and 20 cm to the right (Figure 5). These 3 positions are used to tabulate the results. They do not need to be precise: in both the conventional method and the proposed method, the subjects' head position and orientation were continuously obtained by a video-based motion tracker. The estimate of the acoustic transfer matrix $\mathbf{C}$ could thus be explicitly computed [13], [17], [19]. The results were evaluated by comparing the listener's results with the ground truth information, i.e, the

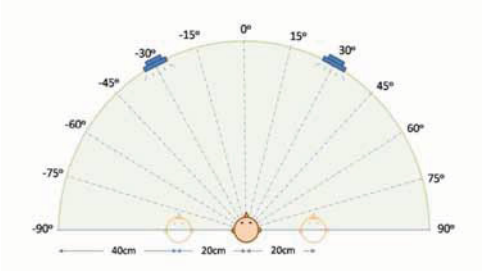

Fig. 5. The listening test configuration.

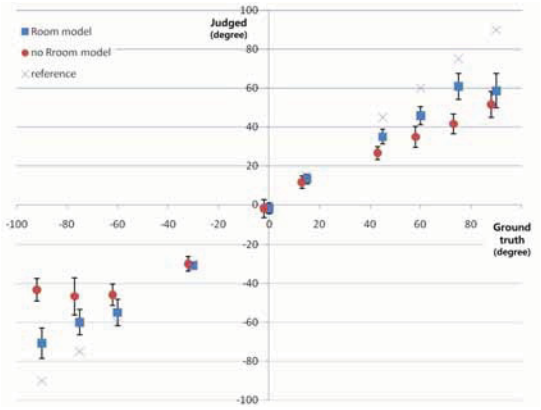

Fig. 6. Test results when the listener is at center.

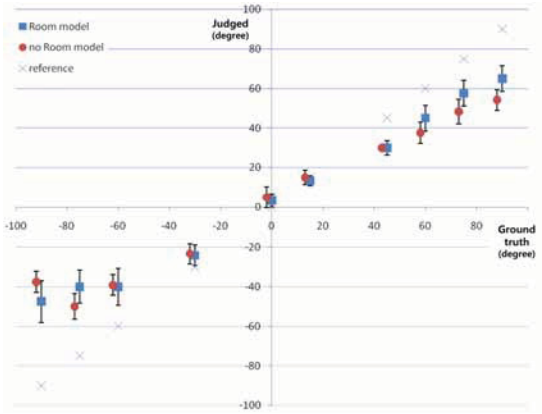

Fig. 7. Test results when the listener is at 20 cm left.

listener estimate of the target position with the "desired" target position. All tests were conducted in a normal laboratory room, with size about $5.6 \times 2.5 \times 3\,\mathrm{m}^3$. The listener's center position is located at 2.1 m away from the right wall and 1.2 m away from the front wall. The room model and the relative position between the loudspeakers and the room were measured by a tape, with accuracy up to 1 cm. The reflection coefficients of the walls are all set to be 0.5, which is a very crude approximation. Note that by using a method like the one proposed in [14] a more accurate estimate should be expected. The room reverberation time RT_{60} of the listening room is approximately 200 ms, calculated by utilizing the Sabine's equation [27].

B. Test Results

The average and standard deviation of azimuth angles identified by the 9 tested subjects are plotted in Figures 6-8, and summarized in Table I. In Figures 6-8 the squares represent the results of the proposed room model based binaural audio system, and the circles show those of the conventional system that does not consider room reverberation. The x-axis denotes the ground truth angles (ie., "target angles") and the y-axis represents the angles identified by the listener. The ground truth or reference angles are also marked in the figures with crosses. Identified angles closer to the reference (i.e., the "X") are better. For better visualization, the results of the traditional methods are plotted with a small offset to the left.

Figure 6 shows the results when the listener was at the center position. The virtual source between $-30°$ and $30°$ were almost always identified correctly, since they are inside the range of the two loudspeakers. When that is the case, crosstalk cancelation is not as critical, and localization performs reasonably well. When the virtual source moved out of the segment between the two loudspeakers, performance dropped. However, the proposed system showed much better accuracy in localizing the virtual sources compared with the conventional system. This demonstrates the effectiveness of the proposed approach of room modeling based crosstalk cancelation.

Even with room modeling, listeners still could not achieve perfect localization for virtual source outside the loudspeaker range. There are a number of reasons for this, besides the non-perfect reverberation modeling. Among contributing factors, we mention 1) there may be small offsets or errors between the estimated listener position and actual position, 2) the HRTFs used in both systems were not personalized, and 3) we did not incorporate a radiation pattern model for the loudspeaker. Of course, each of these can, with some effort, be accounted for.

Figure 7 shows the results when the listeners were placed at 20 cm to the left from the center position. Although both systems were designed under the assumption that the listeners' positions were known, the results are very different from the previous results obtained at the center position. This is understandable, since the the acoustic paths between the loudspeakers and the ears have been altered significantly. Generally speaking, the virtual sources located beyond $30°$ were identified more accurately compared with those beyond $-30°$. It was much easier to reproduce the virtual source on the right side than on the left, because the listener is much closer to the left speaker. The proposed room modeling based method still outperforms the conventional scheme, although the margin is much smaller compared with Figure 6. We believe the small margin can be attributed to the general difficulty in the listeners' localization capability when the loudspeakers are asymmetric [28].

Figure 8 shows the results when the listener was placed at 20 cm to the right from the center position. The overall trend is similar to the previous results. The proposed system based on the room model still shows better performance than the conventional system. The result is not exactly flipped over the previous results from the left position case, however, as the geometry of the room used in this test was not symmetric to

TABLE I
AVERAGE USERS'S ESTIMATION ERROR, IN DEGREES.

User's Position	Average error (no room model)	Average error (w/ room model)	Improvement (%)
center	23.5°	15.9°	32%
20cm left	22.2°	21.4°	4%
20cm right	29.2°	25.8°	11%

the center of the listener's location.

The average azimuth error for different listener's positions are reported in Table I. The azimuth error is calculated as a difference between the ground truth and the judged azimuth. The proposed room modeling based method shows the smaller error to localize the virtual sources than the conventional one for all three user positions. The speaker location is closer to the wall on the left, which explains the worse performance of the traditional algorithm when the user is 20 cm to the left (i.e., closer to the wall), as well the larger improvement when compared to moving 20 cm to the left.

We further conducted the student t-test to assess whether the results of the two systems are statistically different from each other. The absolute values of the difference between the ground-truth and the judged azimuth $\left|\text{Reference}_i - \text{Judged}_{i,n}\right|$ were compared, where i and n are the azimuth and subject index, respectively. The t-test score of the event that the results of the proosed system and the conventional system are drawn from two normal distribution with the same mean is merely 0.0023%, which shows that the difference is statistically very significant.

V. Conclusion

In this paper, we have explored the idea of enhancing conventional binaural audio systems with explicit room modeling. We showed that a simple room model with planar surfaces can be very useful to better calculating the acoustic transfer functions between the loudspeakers and the listener, which leads to better crosstalk cancelation and audio spatialization. Compared with measurement based schemes, our room model based approach has the advantage that the early reflections can be dynamically recomputed if the listener's head moves around. Note that previously published approaches (i.e., RIR estimation or measurement) were targeted mostly at proving the significance modeling the RIR, and were not practical. Thus, the proposed room modeling is a significant improvement over existing techniques: it actually makes room compensation practical. Indeed, our user studies with a real-time system implementing the proposed scheme showed an improvement of up to 32% in user's perception (for the center location), with a t-test score of 0.0023%. It can be expected that with better modeling, even better improvements can be obtained.

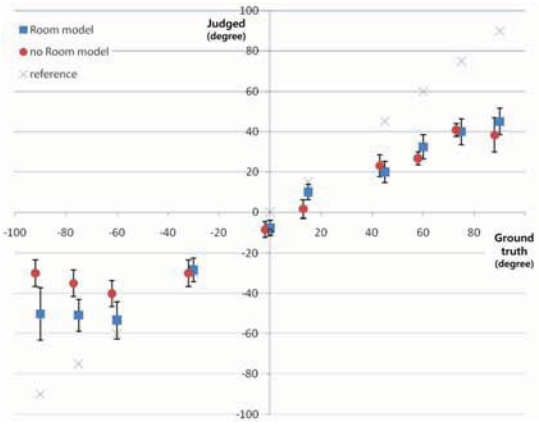

Fig. 8. Test results when the listener is at 20 cm right.

References

[1] D. Cooper and J. Bauck, "Prospects for transaural recording," J. Audio Eng. Soc., vol. 37, pp. 3-19, 1989.

[2] J. Bauck and D. Cooper, "Generalized transaural stereo and applications," J. Audio Eng. Soc., vol. 44, pp. 683-705, 1996.

[3] P. Nelson, H. Hamada, and S. Elliott, "Adaptive inverse filters for stereophonic sound reproduction," Signal Processing, IEEE Transactions on , vol.40, no.7, pp.1621-1632, 1992.

[4] A. Mouchtaris, P. Reveliotis, and C. Kyriakakis, "Inverse filter design for immersive audio rendering over loudspeakers," Multimedia, IEEE Transactions on , vol.2, no.2, pp.77-87, 2000.

[5] O. Kirkeby and P. Nelson, "Virtual source imaging using the stereo dipole", Proc. 103th Convention of the Audio Eng. Soc., 1997.

[6] O. Kirkeby, P. Nelson, and H. Hamada, "Acoustic fields generated by virtual source imaging systems", Proceedings of the Active 97, The international symposium on active control of sound and vibration, pp. 941-954, 1997.

[7] B. Gardner and K. Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone", MIT Media Lab, available on the World Wide Web at http://sound.media.mit.edu/resources/KEMAR.html

[8] A. Mouchtaris, J. Lim, T. Holman, C. Kyriakakis, "Head-related transfer function synthesis for immersive audio," IEEE Second Workshop on Multimedia Signal Processing, pp.155-160, 1998.

[9] P. Georgiou, A. Mouchtaris, I. Roumeliotis, and C. Kyriakakis, "Immersive Sound Rendering Using Laser-Based Tracking," Proc. 109th Convention of the Audio Eng. Soc., Paper 5227, 2000.

[10] T. Lentz, O. Schmitz, "Realisation of an adaptive cross-talk cancellation system for a moving listener," 21st AES Conference on Architectural Acoustics and Sound Reinforcement, 2002.

[11] M. Song, C. Zhang and D. Florencio, "Personal 3D audio system with loudspeakers,", IEEE International Workshop on Hot Topics in 3D, in conjunction with ICME 2010.

[12] C. Kyriakakis, T. Holman, "Immersive audio for the desktop," Proc. IEEE ICASSP, vol. 6, pp. 3753-3756, 1998.

[13] J. J. Lopez, A. Gonzalez and F. O. Bustamante, "Measurement of cross-talk cancellation and equalization zones in 3-D sound reproduction under real listening conditions", AES 16th International Conference, 1999.

[14] D. Ba, F. Ribeiro, C. Zhang and D. Florencio, "L1 Regularized Room Modeling with Compact Microphone Arrays," ICASSP 2010.

[15] F. Ribeiro, C. Zhang, D. Florencio and D. Ba "Using Reverberation to Improve Range and Elevation Discrimination for Small Array Sound Source localization," Audio, Speech and Language Processing, IEEE Transactions on, accepted for publication, 2010.

[16] F. Ribeiro, D. Ba, C. Zhang, and D. Florencio "Turning Enemies Into Friends: using Reflections To Improve Sound Source Localization," ICME2010.

[17] W. Gardner, "3-D audio using loudspeakers," Ph.D. thesis, Massachusets Institute of Technology, 1997.

[18] J. Lim and C. Kyriakakis, "Multirate adaptive filtering for immersive audio," Proc. IEEE ICASSP, vol. 5, pp. 3357 3360, 2001.

[19] S. Kim, D. Kong, and S. Jang, "Adaptive Virtual Surround Sound Rendering System for an Arbitrary Listening Position," J. Audio Eng. Soc., Vol. 56, No. 4, 2008.

[20] F. E. Toole, "Loudspeaker measurements and their relationship to listener preferences," J. Audio Eng. Soc., vol. 34, pp. 227-235, 1986.

[21] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," Acoustical Society of America, Vol. 65, No. 4 (1979), pp. 943-950.

[22] D. Kimber, C. Chen, E. Rieffel, J. Shingu, and J. Vaughan, "Marking up a world: visual markup for creating and manipulating virtual models," in *Proc. of IMMERSCOM*, 2009.

[23] A. O'Donovan, R. Duraiswami, and D. Zotkin, "Imaging concert hall acoustics using visual and audio cameras," in *Proc. of ICASSP*, 2008, pp. 5284–5287.

[24] F. Antonacci, A. Sarti, and S. Tubaro, "Geometric reconstruction of the environment from its response to multiple acoustic emissions," in *Proc. of WASPAA*, 2009.

[25] D. Aprea, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic reconstruction of the geometry of an environment through acquisition of a controlled emission," in *Proc. of EUSIPCO*, 2009.

[26] M. Moebus and A. Zoubir, "Three-Dimensional Ultrasound Imaging in Air using a 2D Array on a Fixed Platform," in *Proc. of ICASSP*, 2007.

[27] L. Beranek, *Concert and opera halls : How they sound*, Acoustical Society of America, 1996, p. 436.

[28] K. Foo and M. Hawksford and M. Hollier, "Three dimensional sound localisation with multiple loudspeakers using a pair-wise association paradigm with embedded HRTFs,", The 104th AES Convention, 1998.