An Interactive 3D Audio System with Loudspeakers

Myung-Suk Song, Cha Zhang, *Senior Member, IEEE,* Dinei Florencio, *Senior Member, IEEE,* and Hong-Goo Kang, *Member, IEEE*

***Abstract*—Traditional 3D audio systems using two loudspeakers often have a limited sweet spot and may suffer from poor performance in reverberant environments. This paper presents a novel binaural 3D audio system that actively combines head tracking and room modeling into 3D audio synthesis. The user's head position and orientation are first tracked by a webcam-based 3D head tracker. The system then improves its robustness to head movement and strong early reflections by incorporating the tracking information and an explicit room model into the binaural synthesis and crosstalk cancellation process. Sensitivity analysis on the room model shows that the method is reasonably robust to modeling errors. Subjective listening tests confirm that the proposed 3D audio system significantly improves the users' perception and ability for localization.**

***Index Terms*—3D audio, head tracking, room modeling, loudspeaker, spatial audio.**

I. Introduction

Recent advances in computation, displays, and networking technology have brought about many new interactive multimedia applications in areas as diverse as telepresence, gaming, remote surgery, etc. Most of these applications strive to provide an immersive experience to the user, e.g., improve image quality by providing high resolution displays or even 3D displays, improve responsiveness by adopting powerful CPU/GPUs, enlarging network bandwidth and shortening network delay, improve system robustness by having quality monitoring and management, security solutions, etc. One aspect of multimedia, however, seems to be lagging behind: realistic audio. Consider, for example, 3D immersive environments which is one key area of interactive multimedia, where multiview imaging is often used to capture real-world scenes. The multiview videos are then transmitted via multiview video compression and streaming schemes, and viewed with free viewpoint rendering on 2D or 3D displays. Such topics have attracted a lot of research interest recently [1]. On the other hand, immersive audio, the counterpart that is indispensable in such systems, seems to have received little attention. Consider an immersive teleconferencing application. Thanks to high quality multiview video capture and 3D rendering, the users can enjoy a faithful sense of the remote attendees' locations in the 3D space. When a remote attendee speaks, it is natural to request the perceived sound source be originated from the same 3D location. Traditional audio spatialization systems can render a virtual sound image in order for the listener to feel as if the signals were emitted by a source located at a certain position in 3D space [2][3]. However, to match the increased user's expectations, we need further improvements in these three-dimensional audio spatialization systems.

The above 3D audio effect may be achieved through *wave field synthesis* [4], which renders a whole sound field to the room through a large number of loudspeakers [5]. Nevertheless, such a solution is expensive and non-scalable. A better solution is to rely on a high quality three-dimensional audio spatialization system. These systems render a *virtual* sound image in order for the listener to feel as if the signals were emitted by a source located at a certain position in 3D space [2][3]. Either headphones or a small number of loudspeakers (two in our system) can synthesize such spatialized audio effects, though the latter is often more appealing in immersive applications since it does not require the user to wear headphones.

There are three popular techniques for loudspeaker-based 3D audio: *Ambisonics* [6], *amplitude panning* [7], and *binaural synthesis* that utilizes the head related transfer functions (HRTF) [8]. Ambisonics and amplitude panning are widely used panning techniques. In both methods, the virtual sound source is rendered at various locations by controlling the output amplitude of the loudspeakers. When two loudspeakers are available, however, they can only reproduce virtual sources in the line segment between loudspeakers. In addition, results degrade significantly if the user gets closer to one of the two loudspeakers. Binaural synthesis is capable of placing the virtual sound beyond the loudspeakers' boundaries due to the use of the HRTF that faithfully represents the transfer function between the sound sources and human ears. Since the loudspeaker-based system has the so-called crosstalk issue (caused by the contralateral paths from the loudspeakers to the listener's ears), a filter bank module known as crosstalk cancellation [9][10][11] needs to be placed in front of the audio reproduction module.

In practice, crosstalk cancellation still presents many challenges. One obstacle is that a crosstalk canceller based 3D audio system works only if the user is in a small zone, called *sweet spot*, because the cancellation filter bank is very sensitive to the user's head location and orientation. A head tracking module has been proposed to overcome the problem [12][13][14][15], where the listener's head movement is continuously tracked to adaptively control the crosstalk canceller. Electromagnetic trackers were used in [16][17], though such devices are expensive and uncomfortable to wear, defeating the purpose of using loudspeakers for 3D audio. Non-intrusive methods using webcams and face tracking techniques have been also proposed [13][18][19]. Nevertheless, these early works did not fully evaluate the effectiveness of their systems due to the limited computational resources and the inaccuracy of the face tracking techniques at that time.

Another major hurdle of operating a 3D audio system in real-world environments is reverberation. Reverberation will change the transfer functions between the loudspeakers and the ears, and significantly reduce the effectiveness of the

Fig. 1. Our personal 3D audio system with one webcam on the top of the monitor, and two loudspeakers.

crosstalk cancellation module (typically designed with a free field assumption). The degradation effect has been noticed by Kyriakakis and Holman [18]. They presented a solution to reduce the negative effect of reverberation by changing the layout of the environment to ensure that the direct path is dominant. However, such a solution is not always feasible. Lopez et al. [20] proposed to model the room reverberation explicitly during the process of computing the transfer functions between the loudspeakers and the listener. They used room impulse responses (RIR) measured by a dummy head to help crosstalk cancellation in reverberant rooms and showed significant improvement. Nevertheless, their results were at best a proof of concept. In real-world environments, the RIR is very hard to measure. It varies significantly with the head position and orientation related to the room environment, and cannot be easily interpolated.

In this paper, we make significant progress towards building an interactive loudspeaker-based 3D audio system. We propose, implement, and test the first *real-time* webcam face tracking based binaural 3D audio system (as shown in Fig. 1). The system accurately estimates the head position and orientation using a webcam-based 3D face tracker [21], and uses that to dynamically move the sweet spot to the user's position. Furthermore, we introduce a powerful room modeling technique to dynamically compensate for a few early reflections of the room transfer function. Instead of measuring the accurate RIR at each listening point, e.g. using dummy head, we model the room with a number of planar reflectors such as walls or ceiling. These room models can be estimated with various approaches such as [27][29][30]. In other words, instead of directly measuring the RIR, we utilize a geometric model of the room to obtain the current RIR, which can be quickly and continuously updated given the user's head position and orientation. By applying an estimated acoustic transfer function that includes the reflections caused by the walls/ceilings of the room to the crosstalk canceller, we improved the channel separation and achieve better spatialized audio effect.

Subjective listening tests were conducted to evaluate the effectiveness of 3D audio synthesis using head tracking alone, and in combination with room modeling. Subjects were asked to identify the virtual sound source locations at different head positions. The results were compared with the ground truth information to independently measure the impact of the head tracking module and the room model based system on human localization accuracy. Our experimental results showed the clear advantage of the proposed system over traditional 3D audio systems without head tracking. They also showed that the estimated RIR using room modeling was able to improve the listener's performance on identifying the virtual source position.

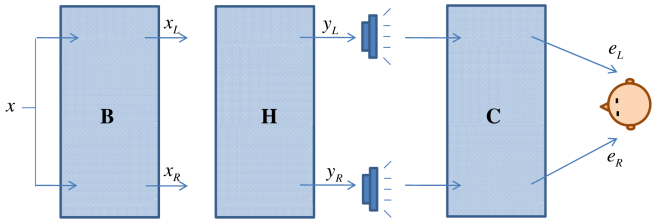

Fig. 2. Schematic of binaural audio system with loudspeakers.

The rest of the paper is organized as follows. Section II illustrates conventional binaural audio systems. The dynamic 3D audio system with head tracking is described in Section III. Section IV introduces the proposed room model based binaural audio system. Performance evaluation and subjective test result are presented in Section V and Section VI, respectively. Finally, Section VII concludes the paper.

II. Conventional Binaural Audio System

The block diagram of a typical binaural audio playback system with two loudspeakers is shown in Fig. 2. Block **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 **B** and crosstalk canceller **H**. The goal of the binaural synthesizer is to compute the sound that should be heard by the listener's ear drum. In other words, we hope that 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. The role of crosstalk canceller is to compensate for the transmission path [2][9].

A. Binaural Synthesis

The binaural synthesizer synthesizes virtual sound images at specified locations around the listener using a monaural audio signal and the HRTF [8][12]. Since HRTFs incorporate most of the physical cues that human relies on for source localization, one can filter the monaural input signal with the head related impulse response (HRIR) 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, B_L, and B_R are the monaural input signal and HRTFs between the listener's ears and the desired virtual source, respectively. The output of binaural synthesis x_L and x_R are the signals that need to be reproduced at the listener's ear drums.

B. Crosstalk Cancellation

An acoustic transfer matrix $\mathbf{C}$ representing the acoustic paths between the two loudspeakers and the listener's ear drums is defined as :

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

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 to the right ear. C_{RL} and C_{LR} are the transfer functions from contralateral speakers, called "crosstalk". Canceling the crosstalk is the main challenge for loudspeaker-based 3D audio systems. The common practice is to insert a crosstalk canceller in order to equalize the transmission path between the loudspeakers and the listener.

In principle, the crosstalk canceller matrix $\mathbf{H}$ could be obtained by taking the inverse of the acoustic transfer matrix $\mathbf{C}$, i.e.,

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

where $D = C_{RR}C_{LL} - C_{LR}C_{RL}$ denotes the determinant of the matrix $\mathbf{C}$. Since the acoustic transfer functions derived from the HRTFs are non-minimum phase, it is generally unstable to compute $\mathbf{H}$ by the direct inversion of $\mathbf{C}$. Instead, we obtain the crosstalk canceller matrix $\mathbf{H}$ by an adaptive least mean square (LMS) method [10][31].

III. Dynamic Binaural Audio System with Head Tracking

The conventional loudspeaker-based binaural audio system described in the previous section works well when the listener stays at a fixed position corresponding to the presumed binaural synthesizer $\mathbf{B}$ and acoustic transfer matrix $\mathbf{C}$. However, the performance rapidly degrades when the listener moves away from that sweet spot. To keep the virtual sound source at the same location even when the head moves, the binaural synthesizer needs to dynamically update its matrix $\mathbf{B}$ to reflect the movement. In addition, the acoustic transfer matrix $\mathbf{C}$ also needs to be updated, which leads to changes for the crosstalk canceller matrix $\mathbf{H}$. The updates of $\mathbf{B}$ and $\mathbf{H}$ were referred as "dynamic binaural synthesis" and "dynamic crosstalk canceller", respectively [15].

This section presents a binaural audio system that is capable of applying a 3D model based face tracker to steer the sweet spot to the listener's head position in real-time [24]. The working flow of the interactive 3D audio system is as follows. First, the position and orientation of the listener's head are detected and tracked. The HRTF filters are then updated using the tracking information. Delays and level attenuation from the speakers to the ears are also calculated to model the new acoustic transmission channel. Finally, the filters for both binaural synthesis and crosstalk cancellation are updated. Each processing step of the system is described in detail below.

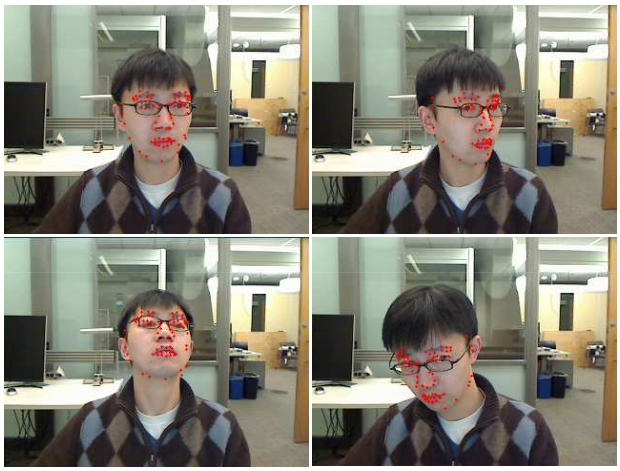

Fig. 3. The tracker adopted in our system tracks the head position and orientation with high accuracy.

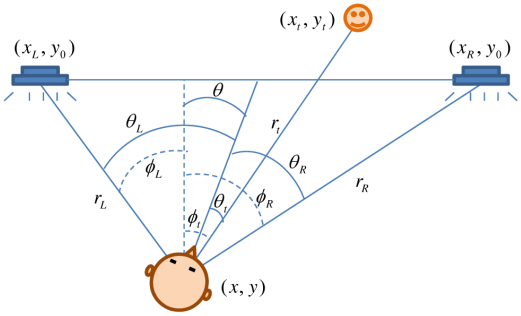

Fig. 4. Configuration of dynamic binaural synthesis and dynamic crosstalk canceller. The listener is located at position (x, y) with head orientation θ, and the virtual source is at position (x_t, y_t).

A. Head Tracking

We adopt the 3D face model based head tracker developed in [21]. Given the input video frames from the monocular webcam, a face detector [22] is first applied to find faces in the scene. A face alignment algorithm [23] is then used to fit a 3D face model on top of the detected face. The face model is then tracked based on tracking feature points on the face. We refer the reader to [21] for more technical details. A few examples of the tracked faces are shown in Fig. 3.

The 3D head tracker outputs the head's position and orientation in the 3D world coordinate of the webcam, assuming the calibration parameters of the webcam are known. The position and orientation information is then transformed into the world coordinate of the loudspeakers, which requires the mutual calibration between the webcam and the loudspeakers. In the current implementation, we assume the webcam is placed in the middle of the two loudspeakers, and its height is roughly measured and given to the system as a known parameter.

B. Dynamic Binaural Synthesis

The dynamic binaural synthesizer renders the virtual sources at the specified locations with the given head tracking information. The synthesizer matrix $\mathbf{B}$ needs to be adaptive to accommodate the relative changes of the virtual source

position caused by head movement. Fig. 4 shows a simplified 2D configuration of the binaural synthesizer and the dynamic crosstalk canceller. The virtual source location (x_t, y_t) is unchanged, but the head position (x, y) and orientation θ are changing. The head tracker provides the latest update on (x, y) and θ, which is then used to recompute the appropriate HRTF based on azimuth θ_t and distance r_t between the virtual sound source and the listener. The filters for the dynamic binaural synthesizer $\mathbf{B}$ are then updated, so that the virtual source remains at the fixed location rather than moving with the listener.

C. Dynamic Crosstalk Canceller

When the listener moves around, the acoustic transfer function between the loudspeakers and the ears will be changed. These changes shall be accounted for in the system. To determine the transfer function between the listener and the left speaker, the HRTF of azimuth θ_L is retrieved from the HRTF database obtained from [25]. Similarly, for the transfer function between the listener and the right speaker, the HRTF of azimuth θ_R is chosen.

The listener's movement also changes the distance between the listener and each loudspeaker, which in turn changes the attenuation and time delay of the sounds from the loudspeakers to the listener's head position. The new time delays d_L and d_R can be calculated based on r_L, r_R (Fig. 4) and the sound speed c. The amplitude level can be estimated by considering the spherical wave attenuation for the specific distances r_L and r_R. In summary, the new acoustic transfer matrix $\mathbf{C}_d$ is defined as:

$$\mathbf{C}_d = \begin{bmatrix} \frac{r_0}{r_L} z^{-d_L} C_{LL} & \frac{r_0}{r_R} z^{-d_R} C_{RL} \\ \frac{r_0}{r_L} z^{-d_L} C_{LR} & \frac{r_0}{r_R} z^{-d_R} C_{RR} \end{bmatrix}, \tag{4}$$

where r_0 is the distance between the loudspeakers and the listener in the conventional binaural audio system and C_{LL}, C_{LR}, C_{RL} and C_{RR} are the transfer functions when the listener is at the perpendicular bisector of the loudspeakers, respectively. If we set $d_R = 0$ the delays d_L can be computed as:

$$d_L = \frac{(r_R - r_L) f_s}{c} \tag{5}$$

where f_s, and c are the sampling frequency and the velocity of sound wave, respectively. To account for cases where $r_L > r_R$, a constant delay can be added to the computation. Fractional delay is approximated by the nearest integer since there was no noticeable degradation.

The dynamic crosstalk canceller $\mathbf{H}_d$ for the moving listener is the inverse of the new acoustic channel model $\mathbf{C}_d$:

$$\begin{aligned} \mathbf{H}_d = \mathbf{C}_d{}^{-1} &= \begin{bmatrix} \frac{r_0}{r_L} z^{-d_L} C_{LL} & \frac{r_0}{r_R} z^{-d_R} C_{RL} \\ \frac{r_0}{r_L} z^{-d_L} C_{LR} & \frac{r_0}{r_R} z^{-d_R} C_{RR} \end{bmatrix}^{-1} \\ &= \frac{1}{r_0} \begin{bmatrix} r_L z^{d_L} & 0 \\ 0 & r_R z^{d_R} \end{bmatrix} \begin{bmatrix} C_{LL} & C_{RL} \\ C_{LR} & C_{RR} \end{bmatrix}^{-1}. \end{aligned} \tag{6}$$

As can be seen in Eq. (6), $\mathbf{H}_d$ can be separated in two matrices or modules. The second matrix represents the conventional crosstalk canceller, while the first matrix is the term to adjust the time difference and intensity difference due to the variations in distance from each loudspeaker to the listener's position.

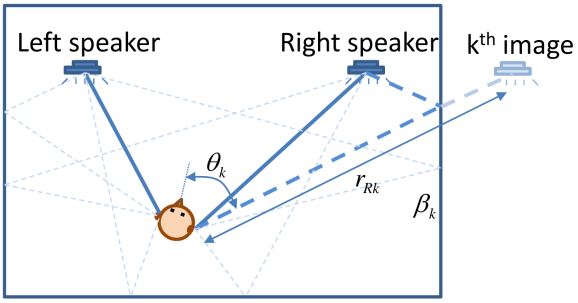

Fig. 5. Acoustic path between two loudspeakers and listener's ears with wall reflections. β_k, r_{Rk}, and θ_k denote the reflection coefficient for the k-th wall, the distance between the k-th image of R speaker and listener, and degree between the k-th image of speaker and listener, respectively.

IV. Room Model Based Binaural Audio System

The computation of the acoustic transfer matrix $\mathbf{C}$ becomes more complicated in real-world environments due to reverberation. Theoretically, the problem could be solved by measuring the RIR between indirect paths from the loudspeakers to the listener as was done in [20]. However, such a scheme is highly impractical since the RIR varies significantly as the listener moves around and cannot be easily interpolated. In this section, we describe an alternative approach: a binaural audio system that accounts for reverberation by explicitly modeling early reflections using a simplified room model. The underlying principle is that early reflections are the dominant source of frequency response anomalies and sound quality degradation in an immersive audio system [18][32]. The key benefit of our approach over the measurement based approach is its capability to handle moving listeners by computing the early reflections through the image method [33] with the listener's position tracked by a face tracking module at any instance. Note, late reverberation that has important effects in subjective distance perception is not the main concern of this paper.

A. Room Model

We assume the user is located in 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). The position and orientation of these planar surfaces can be estimated in various ways [28][27]. In the following discussion we assume the planar room model has been obtained for the test environment.

B. Room Model based Binaural Audio System

In a reverberant room, the sound field at an arbitrary location can be represented by a superposition of numerous reflected sound sources. When a rectangular enclosure room is assumed, the reflection parts 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. This is known as

the image method [33]. As shown in Fig. 5, the acoustic path from each speaker to each ear drum can be represented by the summation of the impulse responses from the actual source and the imaged sources reflected by 6 walls surrounding the listener:

$$C_{mn} = \sum_{k=0}^{N} \frac{\beta_k}{r_{mk}} z^{-\Delta_{mk}} C_{mn}(\theta_k), \quad \text{where } m, n = L \text{ (left) or } R \text{ (right).} \tag{7}$$

Here N denotes the total number of planar surfaces; k denotes the index for images of the speaker ($k = 0$ is the index for the actual loudspeaker). Note that this paper considers only the first reflections of the walls, although extending to multiple reflections is straightforward. m and n represent the indices for 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 m speaker and listener, and the delay from k-th image of m speaker to the listener, respectively. $\Delta_{mk} = \frac{r_{mk}}{c}$, where c is the speed of sound. Note that the head size is assumed to be 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 example, $C_{LL}(\theta_k)$ is the HRTF of the k-th image of the left speaker to the left ear.

When all the N first order reflections are taken into account, the acoustic transfer matrix $\mathbf{C}$ is:

$$\mathbf{C} = \begin{bmatrix} \sum_{k=0}^{N} \frac{\beta_k z^{-\Delta_{Lk}}}{r_{Lk}} C_{LL}(\theta_k) & \sum_{k=0}^{N} \frac{\beta_k z^{-\Delta_{Rk}}}{r_{Rk}} C_{RL}(\theta_k) \\ \sum_{k=0}^{N} \frac{\beta_k z^{-\Delta_{Lk}}}{r_{Lk}} C_{LR}(\theta_k) & \sum_{k=0}^{N} \frac{\beta_k z^{-\Delta_{Rk}}}{r_{Rk}} C_{RR}(\theta_k) \end{bmatrix}. \tag{8}$$

The acoustic path $\mathbf{C}$ is determined by the following procedure. First the reflection coefficients and configuration of walls are measured or estimated. After determining each imaged source using the model, the HRTFs for the left and right ears are calculated, corresponding to the direction of each image. Taking into account the reflection coefficient of the walls and the distance to the location of the listener, the final RIRs are determined by summing the responses of the individual imaged sources caused by the 6 walls as well as the direct path from the actual loudspeakers. Note these RIRs depend on the listener's head position and shall be updated every time the user moves his/her head. Fig. 6 depicts an example of RIR calculated based on the proposed room model method. Based on this calculated RIR, the crosstalk canceller matrix $\mathbf{H}$ is computed using the LMS method as in [10] and [31].

As we mentioned earlier, only the early reflections part of the room reverberation is considered in this paper. Note that the interpretation of reverberation is somewhat different when it is considered as a perceptual element or a transfer function. The perception of reverberation gives us information about target distance and room size, and is clearly dominated by the late reverberation. In contrast, the energy-modifying characteristics of the room transfer function are typically dominated by a few early (strong) reflections. Since we cannot cancel the late reverberation by compensating only for the early reflections, the target depth cannot be made shorter than the distance to the speakers (making the target further away is easy by simply adding more reverberation). However, compensating the early reflections can correct for a large portion of the acoustic energy at the ears, as the early arriving energy is mostly responsible for the localization in azimuth and elevation.

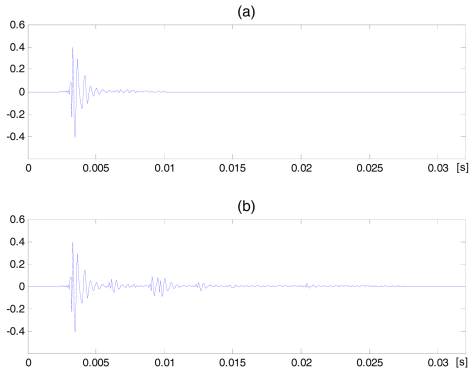

Fig. 6. Room impulse responses with and without room modeling. (a) HRTF and (b) estimated room impulse response based on room model.

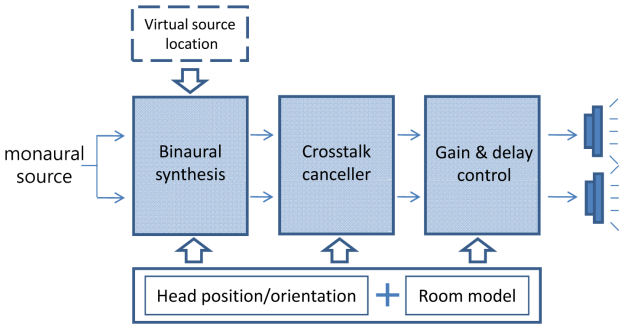

Fig. 7. Block diagram of the complete dynamic binaural audio system with head tracking and room modeling.

C. Complete System

The block diagram of the complete dynamic binaural audio system with a head tracking is shown in Fig. 7. It consists of three modules: the binaural synthesizer, the crosstalk canceller, and the gain and delay control module. These three modules continuously update their filter coefficients according to the newest information of the head position, the virtual source location, and the room model.

The system is updated as follows. The tracking module runs at 30 frames per second. Once motion is detected, the filter will be updated in a separated thread, during which the old filter is still used to output audio signals to avoid interruption. The delay incurred by the filter update process is around $100\,ms$. Such a delay is unnoticeable when the listener moves with normal speed. If the user moves very fast, the delay may be noticeable. However, the artifact will not last longer than the typical filter update process time ($100\,ms$). In practice, none of the test subjects have complained about discomfort due to update delays.

V. Performance Evaluation

In theory, it is obvious that accounting for the listener's head movements and the room impulse response will improve the results of 3D audio spatialization. The main question we would like to answer here is whether the proposed RIR estimation based on the head tracking and the room model is good enough to be useful for enhancing 3D sound perception in practical applications. This section presents simulation results on the crosstalk cancellation performance of our system by measuring the channel separation values.

A. Channel Separation Performance

The performance of the crosstalk canceller using the proposed room model based binaural system highly depends on the accuracy of the estimated room model. Therefore, a key question for the proposed interactive 3D audio system is whether the estimated RIR based on the room model is accurate enough to yield reasonable results. This subsection presents two computer simulation results. First, the performance of the proposed algorithm is compared to the conventional system in terms of channel separation. Second, the robustness issue due to the estimation errors of the proposed room model is described.

The *channel separation* is defined by the ratio between the power of ipsilateral path and contralateral path when only one of the binaural inputs exists. For instance, the left channel separation J_L is defined as

$$J_L = E\left\{20\log_{10}\frac{|C_{LL}H_{LL}+C_{RL}H_{LR}|}{|C_{LR}H_{LL}+C_{RR}H_{LR}|}\right\}\ [dB], \qquad (9)$$

where the expectation $E\{\cdot\}$ is across frequencies. If the crosstalk canceller works perfectly, the left binaural synthesizer output signal x_L must be perfectly reconstructed at the listener's left ear ($C_{LL}H_{LL}+C_{RL}H_{LR}=1$), while the listener's right ear hears nothing ($C_{LR}H_{LL}+C_{RR}H_{LR}=0$). Consequently, the left channel separation J_L becomes infinity.

The RIRs between the loudspeakers and the listener must be properly generated in order to substitute the RIRs measured in a real test room. Usually one can use the well-known image method to simulate a reverberant environment. However, in our simulation, RIRs alone are inadequate to simulate the binaural audio effect, because they do not take into account the "shadow effect" by the head and shoulder shapes. For that reason, we conduct the simulation based on the following model:

$$\mathbf{C} = \begin{bmatrix} \sum_{k=0}^{N'} \frac{\beta_k z^{-\Delta_{Lk}}}{r_{Lk}} C_{LL}(\theta_k) & \sum_{k=0}^{N'} \frac{\beta_k z^{-\Delta_{Rk}}}{r_{Rk}} C_{RL}(\theta_k) \\ \sum_{k=0}^{N'} \frac{\beta_k z^{-\Delta_{Lk}}}{r_{Lk}} C_{LR}(\theta_k) & \sum_{k=0}^{N'} \frac{\beta_k z^{-\Delta_{Rk}}}{r_{Rk}} C_{RR}(\theta_k) \end{bmatrix}. \qquad (10)$$

Note the above equation is very similar to Eq. (8) except that all high order reflections are considered during the simulation. In other words, we have $N' > N$, where N' is the total number of reflections (empirically between 110 and 130 in our setups) and N is the number of walls in the room. Some example RIRs generated with the above model are shown in Fig. 8. The room size is about $5.6\times2.5\times3\ m^3$, and the listener's center position was located at $3.5\,m$ away from the left wall and $1.2\,m$ away from the front wall. The loudspeakers are located at the front side of $-30°$ and $30°$ with a distance of $0.6\,m$ from the center listening location. The short room impulse response length is chosen to speed up the filtering process. Since we are most interested in the early reflections, we found 512 samples at 16 kHz sampling rate to be a good tradeoff between speed and accuracy.

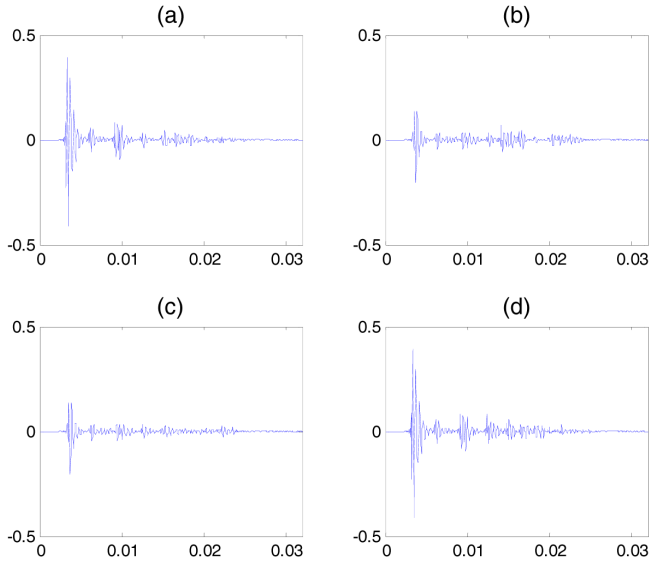

Fig. 8. Room impulse responses generated for computer simulation by Eq.(10). (a) C_{LL}, (b) C_{LR}, (c) C_{RL}, and (d) C_{RR}. x-axis and y-axis represent ms and amplitude, respectively.

TABLE I
Channel Separation by Different Crosstalk Canceller

[dB]	no CTX	conventional	proposed	perfect
J_L	6.7490	9.4878	11.9102	13.2366
J_R	8.0441	10.9441	13.3897	13.7461

Table I shows the results of the channel separations of left and right binaural signals when various crosstalk cancellers are used for binaural audio system. The "no CTX" column represents the results when binaural synthesized signals are directly transmitted to the listener without passing through the crosstalk canceller. The "conventional" and "proposed" columns represent the results by the conventional crosstalk canceller which does not consider reflections and the proposed room modeling based crosstalk canceller, respectively. Note for the proposed method, only the first reflections were used (see Section IV-B). The "perfect" column shows the results when the crosstalk canceller works ideally under the assumption that the exact RIRs are known. All four algorithms were adaptively obtained by the LMS method as in [10] and [31].

The results showed that J_R had slightly higher scores than J_L. This is caused by the geometry of the room assumed for the simulation, which was not centered around the listener's position. The right wall is much closer to the listener than the left wall. Nevertheless, on both sides the proposed algorithm always showed 2-3dB improvement compared to the conventional method.

We next examine the impact of the distance between the lis-

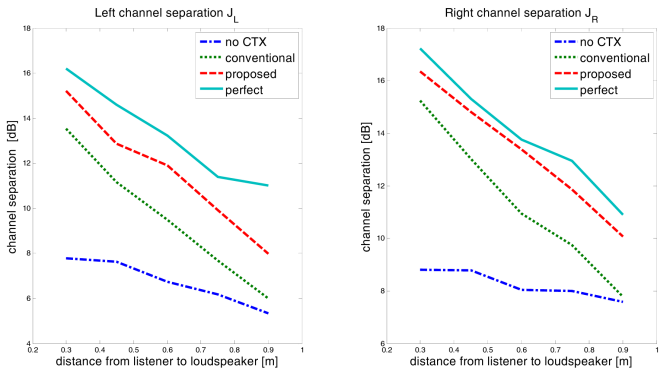

Fig. 9. Channel separations according to the distances between listener and loudspeakers. (a) Left channel separation. (b) Right channel separation.

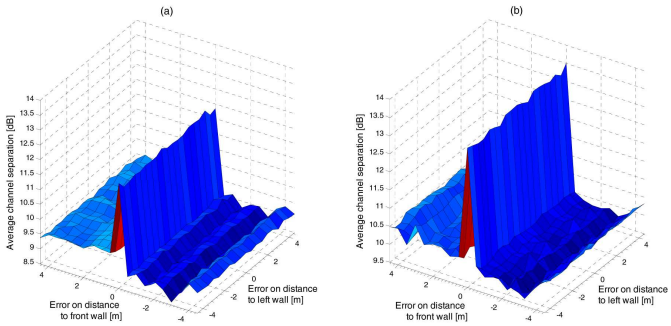

Fig. 10. Channel separations as a function of the error of room model estimation (distance to wall). (a) average channel separation J_L and (b) average channel separation J_R.

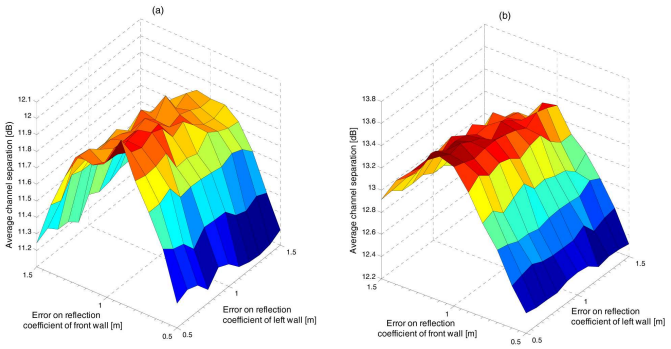

Fig. 11. Channel separations as a function of the error of room model estimation (reflection coefficient of wall). (a) average channel separation J_L and (b) average channel separation J_R.

tener and the loudspeakers on the performance of the crosstalk canceller. Fig. 9 shows the channel separation performance of the crosstalk canceller when varying the distance from the listener to the loudspeakers. In the simulation, the listener is located at a fixed position ($3.5\,m$ away from the left wall and $1.2\,m$ away from the front wall). The two loudspeakers are at the front side of -30° and 30° with distance $0.3 - 0.9\,m$ from the listening position. From Fig. 9, it can be seen that the channel separation values of all methods monotonically decrease when the loudspeakers move away from the listener. These are intuitive results. When the listener is close to the loudspeakers, the left and right channels are highly separated as the direct path dominates. As the distance increases, the room reverberation degrades the performance of all algorithms. On the other hand, the proposed method always showed 2-3dB higher channel separation values than the conventional algorithm. It is clear that the proposed system is effective to equalize the acoustic channel between the loudspeakers and the listener when a certain degree of room modeling is performed.

B. Error Analysis on the Room Model Estimation

In the second experiment we study the influence of room modeling errors to the performance of the crosstalk canceller. The errors in wall distance and reflection coefficients will change the time of arrival and amplitude of the reflections, hence the accuracy of the room model is strongly linked with the precision of the RIR estimates. To model these imprecisions, we include two perturbation values β and Δ into C_{mn} as follows:

$$C_{mn} = \sum_{k=0}^{N} \frac{\beta_k \cdot \beta}{r_{mk}} z^{-\Delta_{mk}-\Delta} C_{mn}(\theta_k), \tag{11}$$

where β is a scale error for the reflection coefficient and Δ represents a delay error caused by the distance to each wall: $\Delta = \Delta d_k / cf$. Here, Δd_k and c represent errors in distance to the k^{th} wall and the speed of sound, respectively. f is the audio sampling rate, which is 16 kHz.

Fig. 10 shows the results of the channel separations of left and right binaural signals by the proposed crosstalk canceller when the room model estimation has error on the distance to each wall while it has perfect knowledge of reflection coefficient of each wall. The simulated room is the same as previously described, thus the front wall is the closest wall to the listener and the left wall is the farthest wall. Intervals for X and Y axis are $0.5\,m$.

The performance of the crosstalk canceller is much more sensitive to the error on distance to the front wall than that to the left wall. The channel separation values are severely degraded by small errors on the distance to the front wall, while it remains flat with respect to the errors on the distance to the left wall. This is expected. As Kyriakakis claimed in [18], "the main impact of reverberation on sound quality in immersive audio systems is due to discrete early reflections, especially early reflections which arrive to the listener less than $15\,ms$". In $15\,ms$ the sound will travel merely $5.1\,m$, therefore the sound reflected by front wall will dominate the channel separation performance. Note that the figure shows that a small error of $0.5\,m$ in the estimate of the front wall distance can cause the channel separation value to drop to about 9 dB, which is comparable to the results of the conventional system in Table I. Fortunately, in previous works such as [27], the room model can usually be estimated with a precision of about $0.02\,m$.

Fig. 11 shows the results of the channel separations of left and right binaural signals by the proposed crosstalk canceller when the room model estimation has error on the reflection coefficient of each wall while the exact distance to each wall

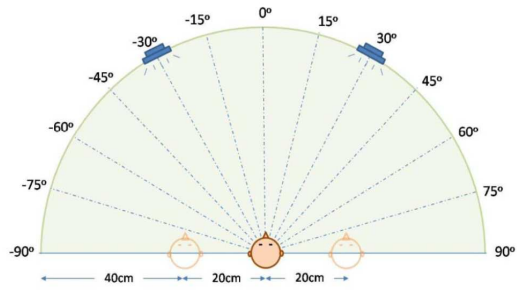

Fig. 12. Listening test configuration with two loudspeakers. Subject was tested at three different positions: center, $20cm$ to the left, and $20cm$ to the right

is known. Although the errors on the reflection coefficient of the front wall still dominate the performance of the crosstalk canceller, the performance does not fall as rapidly as in Fig. 10. Note that even for the worst case ($\beta = 0.5$), the channel separation values are generally above 11dB. This shows that the error on estimating the reflection coefficients of the walls does not influence much the performance of the crosstalk canceller.

VI. Subjective Experiments

To evaluate the effectiveness of the head tracking module in isolation and in combination with the room modeling based approach, we perform controlled subjective tests with real users. This Section summarizes the results of subjective listening tests.

A. Subjective Test Setup

In all our listening tests, the subjects were asked to identify the position from which the sound was originated (between $-90°$ and $90°$). The sounds were generated to be virtually located at a certain position around the listener using the binaural audio algorithm. As shown in Fig. 12, the virtual sound images were rendered at 10 prespecified locations which are randomly selected from 11 candidates $(-90°, -75°, -60°, -45°, -30°, 0°, 15°, 45°, 60°, 75°,$ and $90°)$ on the front horizontal plane with a distance of $0.6\,m$ from the center listening location. The loudspeakers are located at the front side of $-30°$ and $30°$.

All subjects were asked to report their listening results on an answer sheet by indicating the sound source directions freely on a semi-circle. We allowed the listeners to indicate their perceived location with arbitrary accuracy. Most subjects indicated their perceived locations with an increment of 5 degrees. The presentation of the signals and logging of the answers were controlled by the listener. Sound samples were played randomly and repetitions were allowed in every test. The test stimulus consisted of 5 sub-stimulus with $150\,ms$ silent interval. Each sub-stimulus had a pink noise with a sampling rate of 16 kHz and played 5 times in $25\,ms$ duration with $50\,ms$ silent interval.

The tests were conducted with nine participants. Each subject was tested at three different positions: center, $20cm$ to the left, and $20cm$ to the right (Fig. 12). No specific instructions were given to the subjects regarding the orientation of their heads. These three positions were used to tabulate the results, though the subjects might make slight movement during the test. The subjects' head positions and orientations were continuously tracked by a 3D face model based tracker as described in Section III-A, though in the traditional scheme such information was discarded. The acoustic transfer matrix $\mathbf{C}$ were computed based on Eq. (4) and (8).

The results were evaluated by comparing the listener's results with the ground truth information or the "desired" target position. All listening tests were conducted in a normal laboratory room, whose size was about $5.6 \times 2.5 \times 3\ m^3$. The listener's center position was located at $3.5\,m$ away from the left 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 tape, with accuracy around $1\,cm$. The reflection coefficients of the walls were all set to be 0.5, which was a crude approximation. Note that by using a method like the one proposed in [27], one can obtain a reasonably accurate estimate of the reflection coefficients. The room reverberation time RT_{60} of the listening room was approximately $200\,ms$, calculated by utilizing the Sabine's equation [34].

B. Subjective Test Results

The average and standard deviation of azimuth identified by the nine tested subjects are plotted in Figs. 13-15. The squares ("proposed") represent the results of the proposed room model based binaural audio system, the circles ("no room model") show results of the system using head tracking but no reverberation compensation, and the triangles ("no head tracking") shows the results for a conventional system, i.e., without head tracking or consideration of room reverberation. The horizontal axes denote ground truth angles (i.e., "target angles"), and the vertical axes represent the angles identified by the listener. The "X"s ("reference") depict the ground truth, i.e., perfect results. Identified angles closer to the reference (i.e., the "X") are better. For better visualization, the results of the system without room model are plotted with a small offset to the left, and those of the conventional system are plotted to the right.

Fig. 13 shows the results when the listener was at the center position. Virtual sources between $-30°$ and $30°$ degree were almost always identified correctly. This is expected, because they were inside the range of the two loudspeakers, and the user is at the sweet spot. When the virtual sources were outside of the range of the two loudspeakers, the performance of the system with only head tracking module and the conventional system dropped. However, the proposed system showed much better accuracy in localizing the virtual sources compared to others. This demonstrates the effectiveness of the proposed approach of room modeling based crosstalk cancellation. Note also that "no room model" and "no head tracking" show similar performance. This is expected as well, since the listeners were asked to stay at the center position, which happened to be the sweet spot for the conventional system.

Even with room modeling, listeners still could not achieve perfect localization for virtual sources outside the loudspeaker

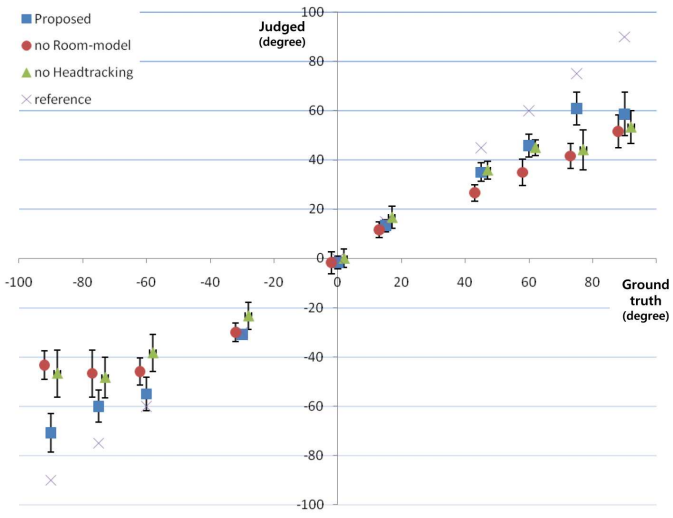

Fig. 13. Results when the listener is at center. 9 subjects were tested in a normal laboratory room, whose size is about $5.6 \times 2.5 \times 3\ m^3$.

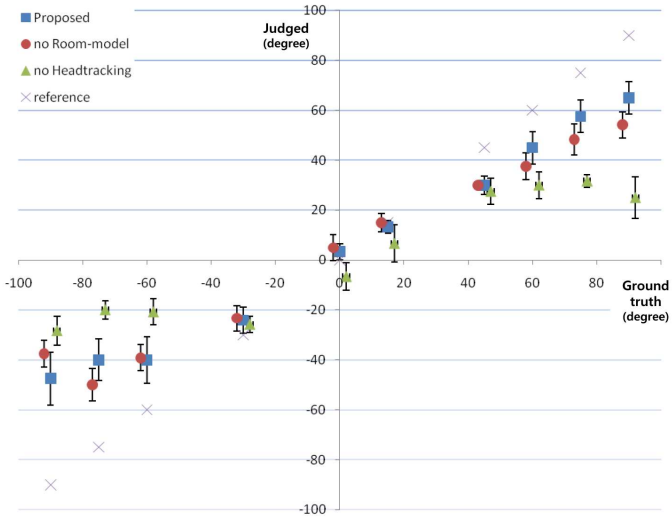

Fig. 14. Results when the listener is at 20cm left. 9 subjects were tested in a normal laboratory room, whose size is about $5.6 \times 2.5 \times 3\ m^3$.

range. There are a number of reasons for this besides the non-perfect reverberation modeling. Among many contributing facts, we mention 1) there may be small offset or errors between the estimated listener position and actual position, 2) the HRTFs used in the 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.

Fig. 14 shows the results when the listener was at $20\,cm$ to the left from the center position. While the circles are still plotted similarly with previous results obtained at the center position, the triangles are now constrained between $-30°$ and $30°$. Since the subjects were away from the sweet spot, they identified the virtual source localized outside of the loudspeakers as somewhere between $-30°$ and $30°$. Even for the virtual sources located inside of the loudspeakers, the performance of the conventional system degraded. The triangles for $0°$ and $15°$ are at much lower angles than the ground truth, because the virtual source reproduced without head tracking follows the listeners' movement to the left. In contrast, the systems with head tracking showed more robust performance than the conventional one without head tracking.

The proposed system shows much better performance compared to others in all aspects. Although both the proposed system and the system without room modeling were designed under the assumption that the listeners' positions were known, the results were very different from the previous results obtained at the center position. This is understandable, since the acoustic paths between the loudspeakers and the ears have been altered significantly. Note that the virtual sources located beyond $30°$ are identified more clearly compared to the ones beyond $-30°$. It was much easier to reproduce the virtual source on the right side than the ones on the left, since the listeners were much closer to left speaker. The proposed room modeling based method still outperforms the system without room modeling, although the margin is much smaller compared with Fig. 13. We believe the small margin can be

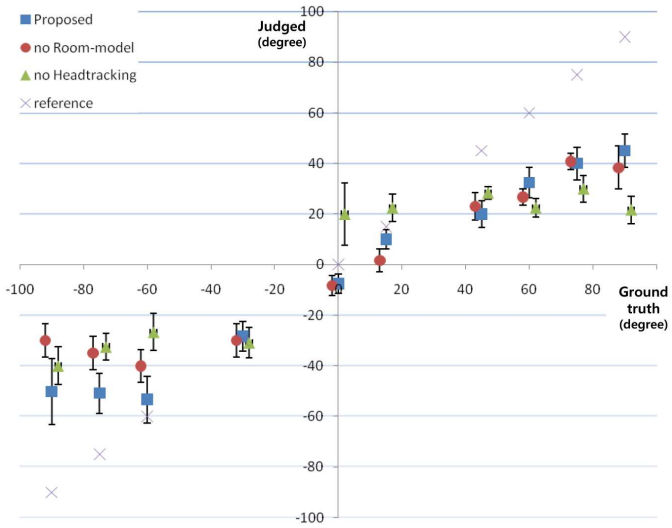

Fig. 15. Results when the listener is at 20cm right. 9 subjects were tested in a normal laboratory room, whose size is about $5.6 \times 2.5 \times 3\ m^3$.

attributed to the general difficulty in the listeners' localization capability when the loudspeakers are asymmetric [35].

Fig. 15 shows the results when the listener was at $20\,cm$ to the right from the center position. The overall trend is similar to the previous results obtained from the left position. The proposed method performs best, followed by the method with only head tracking, with the conventional system performing the worst. The result is not exactly flipped over the previous results, however, as the geometry of the room used in this test was not symmetric to the center of the listener's location (the right wall is much closer to the listeners than the left wall).

We conducted Student's t-tests to assess the statistical significance of the results. The absolute values of difference between the ground-truth and the judged azimuth $\left|\text{Reference}_i - \text{Judged}_{i,n}\right|$ are compared, where i and n are azimuth and subject index, respectively. The t-test score (p-value) of the event that the proposed algorithm is better than the system without room modeling is 0.000023 (both

TABLE II
CHANNEL SEPARATION BY DIFFERENT CROSSTALK CANCELLER

p-value	Center	Left	Right
RM vs noRM	0.000021	0.080848	0.028106
noRM vs noHT	0.346595	0.000079	0.081687

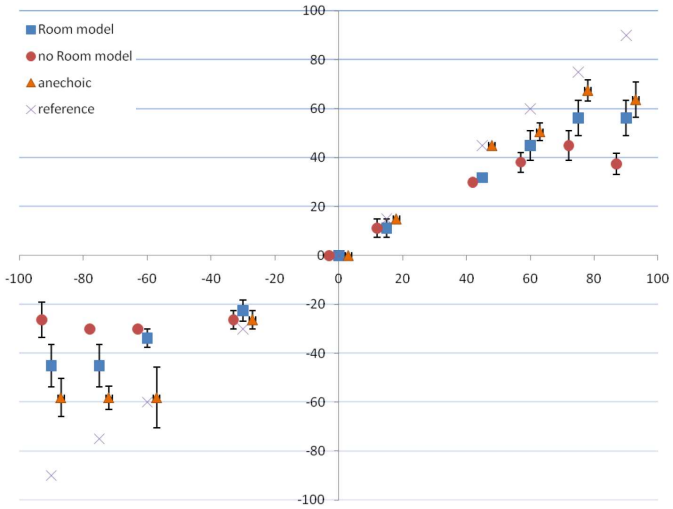

Fig. 16. Results of 4 "good" listeners among the 9 subjects, who were tested in both of the normal laboratory room and an anechoic chamber.

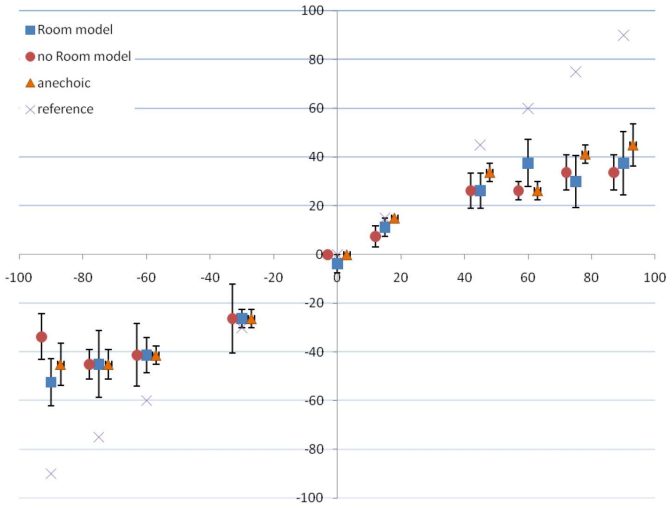

Fig. 17. Results of 4 "bad" listeners among the 9 subjects, who were tested in both of the normal laboratory room and an anechoic chamber.

has head tracking), which shows the effectiveness of room modeling. The t-test score of the event that the system without room modeling (but with head tracking) is better than the conventional system (no head tracking) is 0.0019, which shows the effectiveness of head tracking. Overall, both results are statistically very significant.

Table II shows the p-values for different listeners' positions. "RM" represents the results of the proposed room model based binaural audio system with head tracking module, "noRM" shows results of the system using head tracking but no reverberation compensation. "noHT" shows the results for a conventional system, i.e., without head tracking or consideration of room reverberation. The results show that the proposed room modeling (RM) method with head tracking is very good at center position and is fairly good at other positions. although there is no significant difference between "noRM" and "noHT" at the center position (sweet spot), "noRM" shows much better performance at other positions.

We further conducted experiments to examine if there is significant difference in listening capability among the test subjects. In this comparison the listening tests were conducted in both the laboratory room and an anechoic chamber. The size of the anechoic chamber is about $3 \times 3 \times 3\ m^3$, the listeners were located at the middle of it. The same 9 subjects participated in the test. The head tracking module was not invoked, and we assume the listener's head is at the center position (Fig. 12). By calculating the mean square error with the ground truth information, we selected 4 "good" listeners who have the smallest errors, and 4 "bad" subjects who have the largest errors. The results are shown in Fig. 16 and Fig. 17, respectively. In both figures, the squares ("Room model") represent the results of the room model based binaural audio system without head tracking, the circles ("no room model") show results of a conventional system without head tracking or consideration of room reverberation, and the triangles ("anechoic") shows the results of the conventional system tested in the anechoic room, which can be considered as an upper bound of the performance that the system can possibly reach.

Compared with Fig. 13, the overall performance degraded slightly, since head tracking is disabled and the subjects may move their head slightly during the test. In Fig. 16, the sound in the anechoic room was accurately identified, and it is confirmed that the room model based system outperforms the system without reverberation compensation significantly. However, in Fig. 17, the "bad" listeners performed poorly in all three tests and have no perception for virtual sources outside the loudspeaker range, even in the anechoic chamber. In this case there is not much advantage to use the proposed method with room modeling. This could be explained as follows. The HRTF used in our system is a generic one [25]. It may not fit the "bad" listeners' head and ear shape very well, thus the binaural synthesis stage failed to produce good spatialized sound for these subjects. It is our future work to find novel schemes to measure each subject's personal HRTF efficiently.

VII. CONCLUSION

In this paper, we explored a novel interactive 3D audio system using loudspeakers, which actively combines dynamic sweet spot with dynamic reverberation compensation. The reverberation compensation is based on a novel room modeling approach, which circumvents the need for the daunting task of RIR interpolation. With an accurate 3D face model based head tracking algorithm, the system can move the sweet spot to the position of the listener, as well as compensate for the user's head orientation. The proposed room modeling based approach can provide better estimations of the acoustic transfer functions between the loudspeakers and the listener, which leads to better crosstalk cancellation and audio spatialization.

REFERENCES

[1] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen and C. Zhang, "multi-view imaging and 3DTV," *Signal Processing Magazine*, special issue on Multi-View Imaging and 3DTV, vol. 24, no. 6, 2007.

[2] D. Cooper and J. Bauck, "prospects for transaural recording," *Journal of the Audio Engineering Society*, vol. 37, pp. 3-19, 1989.

[3] C.Kyriakakis, "fundamental and technological limitations of immersive audio systems," *Proceedings of the IEEE*, vol. 86, pp.941-951, 1998.

[4] A. Berkhout, D. de Vries, and P. Vogel, "acoustic control by wave field synthesis," *Journal of the Acoustical Society of America*, vol. 93, 1993.

[5] V. Pullki, "spatial sound generation and perception by amplitude panning techniques," Ph.D. thesis, *Helsinki University of Technology*, 2001.

[6] D. Malham and A. Myatt, "3-D sound spatialization using ambisonic techniques," *Journal of the Computer Music*, vol. 19, no. 4, pp. 58-70, 1995.

[7] V. Pullki, "virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, pp. 456-466, 1997.

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

[9] J. Bauck and D. Cooper, "generalized transaural stereo and applications," *Journal of the Audio Engineering Society*, vol. 44, pp. 683-705, 1996.

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

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

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

[13] J. Lopez and A. Gonzalez, "3-D audio with dynamic tracking for multimedia environtments," *2nd COST-G6 Workshop on Digital Audio Effects*, 1999.

[14] S. Kim, D. Kong and S. Jang, "adaptive virtual surround sound rendering system for an arbitrary listening position," *Journal of the Audio Engineering Society*, vol. 56, no. 4, 2008.

[15] T. Lentz and G. Behler, "dynamic crosstalk cancellation for binaural synthesis in virtual reality environments," *Journal of the Audio Engineering Society*, vol. 54, Issue 4, pp. 283-294, 2006.

[16] P. Georgiou, A. Mouchtaris, I. Roumeliotis and C. Kyriakakis, "immersive sound rendering using laser-based tracking," *Proceedings of the 109th Convention of the Audio Engineering Society*, Paper 5227, 2000.

[17] T. Lentz and O. Schmitz, "realisation of an adaptive cross-talk cancellation system for a moving listener," *21st Audio Engineering Society Conference on Architectural Acoustics and Sound Reinforcement*, 2002.

[18] C. Kyriakakis and T. Holman, "immersive audio for the desktop," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 6, pp. 3753-3756, 1998.

[19] C. Kyriakakis and T. Holman, "video-based head tracking for improvements in multichannel loudspeaker audio," *Proceedings of the 105th Convention of the Audio Engineering Society*, San Francisco, CA, 1998.

[20] J. J. Lopez, A. Gonzalez and F. O. Bustamante, "measurement of crosstalk cancellation and equalization zones in 3-D sound reproduction under real listening conditions," *Audio Engineering Society 16th International Conference*, 1999.

[21] Q. Cai, A. Sankaranarayanan, Q. Zhang, Z. Zhang, and Z. Liu, "real time head pose tracking from multiple cameras with a generic model," *IEEE Workshop on Analysis and Modeling of Faces and Gestures*, 2010.

[22] C. Zhang and P. Viola, "multiple-instance pruning for learning efficient cascade detectors," *Proceedings of the Neural Information Processing Systems*, 2007.

[23] Y. Zhou, L. Gu, and H. J. Zhang, "Bayesian tangent shape model: estimating shape and pose parameters via Bayesian inference," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[24] M. Song, C. Zhang, D. Florencio and Hong-Goo Kang, "personal 3D audio system with loudspeakers," *IEEE International Workshop on Hot Topics in 3D*, in conjunction with the IEEE International Conference on Multimedia & Expo, 2010.

[25] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," Tech. Rep. 280, MIT Media Lab Perceptual Computing, 1994. http://sound.media.mit.edu/KEMAR.html.

[26] M. Song, C. Zhang, D. Florencio and Hong-Goo Kang, "enhancing loudspeaker-based 3D audio with room modeling," *IEEE International Workshop on Multimedia Signal Processing*, 2010.

[27] D. Ba, F. Ribeiro, C. Zhang and D. Florencio, "L1 regularized room modeling with compact microphone arrays," *International Conference on Acoustics, Speech and Signal Processing*, 2010.

[28] D. Kimber, C. Chen, E. Rieffel, J. Shingu and J. Vaughan, "marking up a world: visual markup for creating and manipulating virtual models," *International Conference on Immersive Telecommunications*, 2009.

[29] F. Ribeiro, C. Zhang, D. Florencio and D. Ba "using reverberation to improve range and elevation discrimination for small array sound source localization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1781-1792, 2010.

[30] F. Ribeiro, D. Ba, C. Zhang and D. Florencio "turning enemies into friends: using reflections to improve sound source localization," *IEEE International Conference on Multimedia & Expo*, 2010.

[31] J. Lim and C. Kyriakakis, "multirate adaptive filtering for immersive audio," *International Conference on Acoustics, Speech and Signal Processing*, 2001.

[32] F. E. Toole, "loudspeaker measurements and their relationship to listener preferences," *Journal of the Audio Engineering Society*, vol. 34, pp. 227-235, 1986.

[33] J. Allen and D. Berkley, "image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943-950, 1979.

[34] L. Beranek, "concert and opera halls. how they sound," *Journal of the Acoustical Society of America*, 1996.

[35] K. Foo, M. Hawksford and M. Hollier, "three dimensional sound localisation with multiple loudspeakers using a pair-wise association paradigm with embedded HRTFs," *Proceedings of the 104th Convention of the Audio Engineering Society*, 1998.