PERSONAL 3D AUDIO SYSTEM WITH LOUDSPEAKERS

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

[#] Department of Electrical and Electronic, Yonsei University
[*] Microsoft Research
[1] earth112@dsp.yonsei.ac.kr, [2] chazhang@microsoft.com,
[3] dinei@microsoft.com, and [4] hgkang@yonsei.ac.kr

ABSTRACT

Traditional 3D audio systems often have a limited sweet spot for the user to perceive 3D effects successfully. In this paper, we present a personal 3D audio system with loudspeakers that has unlimited sweet spots. The idea is to have a camera track the user's head movement, and recompute the crosstalk canceller filters accordingly. As far as the authors are aware of, our system is the first non-intrusive 3D audio system that adapts to both the head position and orientation with six degrees of freedom. The effectiveness of the proposed system is demonstrated with subjective listening tests comparing our system against traditional non-adaptive systems.

Keywords— binaural, immersive, 3D audio, head tracking

1. INTRODUCTION

A three-dimensional audio system renders sound images around a listener by using either headphones or loudspeakers [1]. In the case of a headphone-based 3D audio system, the 3D cues to localize a virtual source can be perfectly reproduced at the listener's ear drums, because the headphone isolates the listener from external sounds and room reverberations. In contrast, with loudspeakers, the sound signal from both speakers will be heard by both ears, which creates challenges for generating 3D effects.

One simple yet effective technique for loudspeaker-based 3D audio is *amplitude panning* [2]. Amplitude panning relies on the fact that human can perceive sound directions effectively based on the level difference between the ear drums. It renders the virtual sound source at different locations by adaptively controlling the output amplitude of the loudspeakers. Unfortunately, amplitude panning cannot reproduce virtual sources outside the region of loudspeakers, which limits its applications in desktop scenarios where usually only two loudspeakers are available.

An alternative solution is to generate the virtual sound sources based on synthetic head related transfer functions (HRTF) [3] through crosstalk cancellation. Crosstalk cancellation uses the knowledge of HRTF and attempts to cancel the crosstalk between the left loudspeaker and the right ear and between the right loudspeaker and the left ear. Since HRTF faithfully records the transfer function between sound sources and human ears, the virtual sound source can be placed beyond the loudspeakers' boundaries. On the other hand, HRTF varies due to changes in head positions and orientations, thus such HRTF-based 3D audio systems work only when the user is in a small zone called "sweet spot".

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

In order to overcome the small sweet spot problem, researchers have proposed to use a head tracking module to facilitate 3D audio generation [4, 5, 6, 7]. The listener's head movement is tracked to adaptively control the crosstalk canceller in order to steer the sweet spot towards the user's head position/orientation. For instance, in [8, 9], the listener's head movement was tracked using electromagnetic trackers, although such devices are expensive and discomfortable to wear. A non-intrusive and more attractive method is to track the head movement with webcams and face tracking techniques [5, 10, 11]. Nevertheless, due to the limited computational resources and incapable face tracking techniques at that time, these early works cannot fully evaluate the effectiveness of tracking based 3D audio generation. For instance, none of the above work considered the listener's movement beyond their 2D motion parallel to the webcam's imaging plane, and none of them provided any evaluation results on how well their systems performed.

In this paper, we combine a 3D model based face tracker with dynamic binaural synthesis and dynamic crosstalk cancellation to build a true personal 3D audio system. The basic hardware

978-1-4244-7493-6/10/$26.00

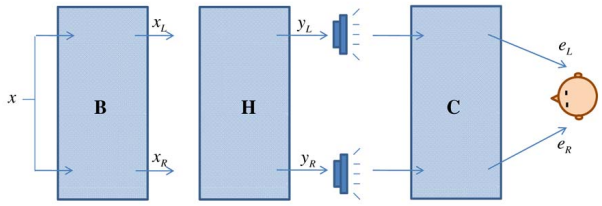

Fig. 2. Schematic of binaural audio system with loudspeakers.

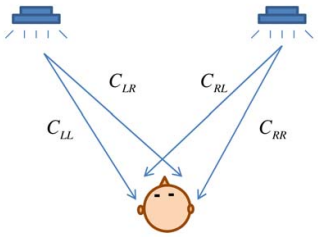

Fig. 3. acoustic path between two loudspeakers and listener's ears

setup is shown in Figure 1. The webcam-based 3D face tracker provides accurate head position and orientation information to the binaural audio system, which uses the information to adaptively synthesize the target audio to be played by the loudspeakers. The system runs in real-time on a dual-core 3GHz machine, which serves the listener with realistic 3D auditory experiences.

In addition, we conducted subjective listening tests to evaluate the effectiveness of head tracking for 3D audio synthesis. Subjects were asked to identify the virtual sound source locations at different head positions. The results were compared with the ground truth information to measure the impact of head tracking on human localization accuracy. Results of the subjective tests showed clear advantage of the proposed system when compared with traditional 3D audio systems without head tracking based adaption.

The rest of the paper is organized as follows. Section 2 introduces conventional binaural audio systems. The proposed personal 3D audio system with head tracking is described in Section 3. Experimental results and conclusions are presented in Section 4 and Section 5, respectively.

2. CONVENTIONAL BINAURAL AUDIO SYSTEM

The block diagram of a typical binaural audio playback system with two loudspeakers is depicted in Figure 2. Component $\mathbf{C}$ represents the physical transmission path or the acoustic channel between the loudspeakers and the listener's ears, which is usually assumed as known. 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 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. The crosstalk canceller, subsequently, aims to equalize the effect of the transmission path $\mathbf{C}$ [12][13].

2.1. Binaural synthesis

The binaural synthesizer $\mathbf{B}$ synthesizes one or multiple virtual sound images at different locations around the listener using 3D audio cues. Among many binaural cues for the human auditory system to localize sounds in 3D such as the interaural time difference (ITD) and the interaural intensity difference (IID), we explore the use of HRTF, which is the Fourier transform of the head-related impulse response (HRIR). Since HRTF captures most of the physical cues that human relies on for source localization. Once the HRTFs of the ears are known, it is possible to synthesize accurate binaural signals from a monaural source [4]. For instance, one can filter the monaural input signal with the impulse response of the HRTF for a given 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 output of binaural synthesis x_L and x_R are the signals that should be reproduced at the listener's ear drums.

2.2. Crosstalk Cancellation

The acoustic paths between the loudspeakers and the listener's ears (Figure 3) are described by an 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 to the right ear. For headphone applications, the acoustic channels are completely separated, because the sound signal from the left speaker goes only to the left ear, and the right signal goes only to the right ear. Therefore, the listener feels perfect 3D auditory experience. In loudspeaker applications, however, the paths from the contralateral speakers such as C_{RL} and C_{LR}, often referred as the "crosstalks", can destroy the 3D cues of binaural signals. The crosstalk canceller plays an essential role in equalizing 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 determinant of the matrix $\mathbf{C}$. Note that it is not easy to calculate the inverse filter $\frac{1}{D}$ due to unstability, because acoustic transfer functions including HRTFs generally are

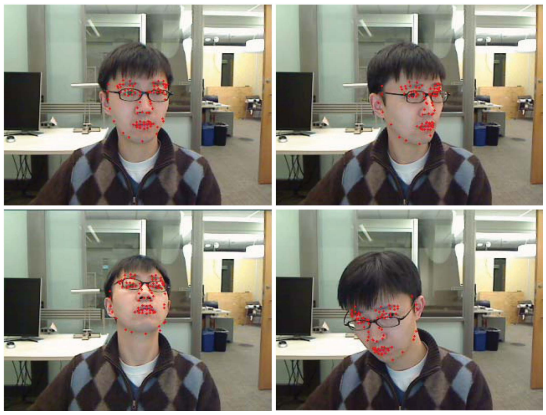

Fig. 4. The tracker adopted in our system tracks the head position and orientation with high accuracy.

non-minimum phase filters. In practice, the crosstalk canceller $\mathbf{H}$ can be adaptively obtained by a least mean square (LMS) method [14][15].

3. PERSONAL 3D AUDIO SYSTEM WITH HEAD TRACKING

The conventional binaural audio system works well if the listener stays at the position (usually along the perpendicular bisector of the two loudspeakers) corresponding to the presumed binaural synthesizer $\mathbf{B}$ and acoustic transfer matrix $\mathbf{C}$. However, once the listener moves away from the sweet spot, the system performance degrades rapidly. If the system intends to keep the virtual sound source at the same location, when the head moves, the binaural synthesizer shall update its matrix $\mathbf{B}$ to reflect the movement. In addition, the acoustic transfer matrix $\mathbf{C}$ needs to be updated too, which leads to a varying crosstalk canceller matrix $\mathbf{H}$. The updates of $\mathbf{B}$ and $\mathbf{H}$ were referred as "dynamic binaural synthesis" and "dynamic crosstalk canceller", respectively [7].

In this paper, we propose to build a personal 3D audio system with a 3D model based head tracker. The hardware setup is shown in Figure 1. The working flow of the dynamic 3D audio system is as follows. First, the position and orientation of the listener's head is 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. We describe each processing step of the system in detail below.

3.1. Head Tracking

We adopt a 3D face model based head tracker similar to the one in [16]. Given the input video frames from the webcam, a face detector [17] is first applied to find faces in the scene. A face alignment algorithm [18] 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 [16] for more technical details. A few examples of the tracked faces are shown in Figure 4.

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 number.

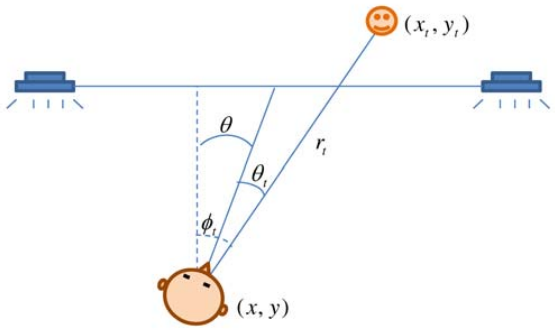

Fig. 5. Dynamic binaural synthesis.

3.2. Dynamic Binaural Synthesis

Given the head tracking information, the dynamic binaural synthesizer renders the virtual sound sources at specified locations. In order to avoid changing of the virtual source position due to head movement, the synthesizer matrix $\mathbf{B}$ needs to be adaptive. A simplified 2D configuration of the synthesizer is shown in Figure 5. The position (x, y) and rotation θ of the listener is first tracked. By calculating azimuth θ_t and distance r_t to the position that the virtual source should be located with respect to the tracked listener's position, the appropriate HRTF is recomputed. The filters of the dynamic binaural synthesizer $\mathbf{B}$ are updated, so that the virtual sources remain fixed as the listener moves rather than moving with the listener.

3.3. Dynamic Crosstalk Canceller

When the listener moves around, the acoustic transfer functions between the loudspeakers and the ears are changed. Figure 6 depicts a configuration for dynamic crosstalk cancellation. To determine the transfer function between the listener and the left speaker, the HRTF of azimuth θ_L is used. Similarly, for the transfer function between the listener and the right speaker, the HRTF of azimuth θ_R is chosen.

The listener's movement changes the distance between the listener and each loudspeaker, which results in level differences and varying time delays of the sounds from the loudspeakers to the listener's head position. The new time delays d_L and d_R

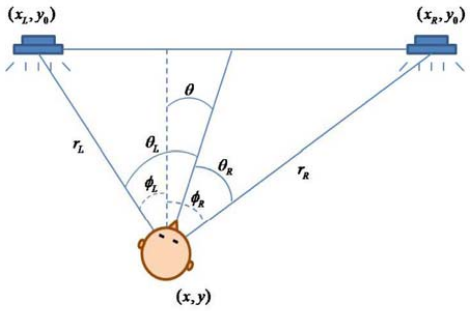

Fig. 6. Dynamic crosstalk canceller.

can be calculated based on r_L, r_R and the sound speed. And the level can be adjusted by considering the spherical wave attenuation for the specific distances r_L and r_R. For instance, the acoustic transfer functions from the left speaker to the listener C_{LL} and C_{LR} need to be attenuated by $\frac{r_0}{r_L}$ and delayed by d_L, and the acoustic transfer functions from the right speaker to the listener C_{RL} and C_{RR} need to be attenuated by $\frac{r_0}{r_R}$ and delayed by d_R. Here r_0 is the distance between the loudspeakers and the listener in the conventional binaural audio system. The new acoustic transfer matrix $\mathbf{C}_d$ is thus 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 C_{LL}, C_{LR}, C_{RL} and C_{RR} are the transfer functions when the listener is at the perpendicular bisector of the loudspeakers. The delays d_L and d_R are computed as follows. If $r_L \leq r_R$,

$$\begin{cases} d_L = \text{int}\left[\frac{(r_R - r_L) f_s}{c}\right] \\ d_R = 0 \end{cases}, \tag{5}$$

otherwise,

$$\begin{cases} d_L = 0 \\ d_R = \text{int}\left[\frac{(r_L - r_R) f_s}{c}\right] \end{cases}. \tag{6}$$

where int$[\cdot]$, f_s, and c are the integer operator, the sampling frequency and the velocity of sound wave, respectively.

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

$$\begin{aligned} \mathbf{H}_d = \mathbf{C}_d{}^{-1} &= \begin{bmatrix} \frac{r_0}{r_L} z^{-d_L} C_{LL} & \frac{r_0}{r_R} z^{-d_R} C_{RL} \\ \frac{r_0}{r_L} z^{-d_L} C_{LR} & \frac{r_0}{r_R} z^{-d_R} C_{RR} \end{bmatrix}^{-1} \\ &= \frac{1}{r_0} \begin{bmatrix} r_L z^{d_L} & 0 \\ 0 & r_R z^{d_R} \end{bmatrix} \begin{bmatrix} C_{LL} & C_{RL} \\ C_{LR} & C_{RR} \end{bmatrix}^{-1}. \end{aligned} \tag{7}$$

As seen in Eq. (7), $\mathbf{H}_d$ can be separated as two modules. The latter matrix represents the conventional crosstalk canceller. And the former 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.

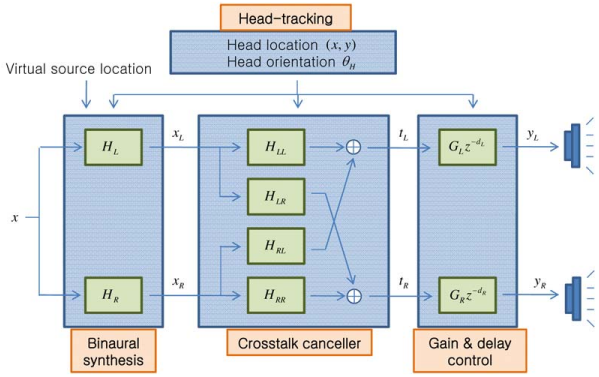

Fig. 7. Block diagram of the complete dynamic binaural audio system.

3.4. The Complete Personal 3D Audio System

To summarize this section, we show the block diagram of the complete dynamic binaural audio system with head tracking in Figure 7. There are three audio related modules in the system: the binaural synthesizer, the crosstalk canceller, and the gain and delay control. These three modules keep updating their filters every time the listener's movement is detected by the head tracking module.

4. EXPERIMENTAL RESULTS

We conducted subjective listening tests to evaluate the performance of the proposed personal 3D audio system with head tracker shown in Figure 7. The results are compared with a conventional binaural audio system without head tracking based adaptation.

4.1. 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 8. The two loudspeakers were located at $\pm 30°$, respectively. The virtual sound images were rendered at 10 prespecified 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 results on an answer sheet. 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 original monaural stimulus consisted 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.

A total of 9 subjects participated the subjective study. Each subject was tested at 3 different positions: center, 20 cm to the

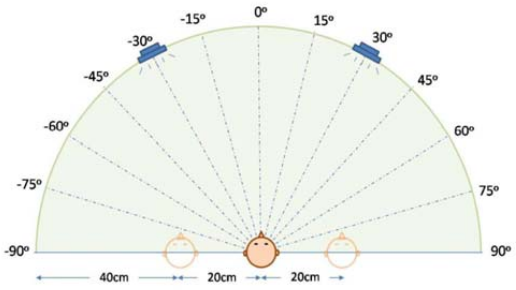

Fig. 8. The listening test configuration.

left, and 20 cm to the right (Figure 8). No specific instructions were given to the subjects regarding the orientation of their heads. The conventional binaural audio system and the proposed head tracking based dynamic binaural audio system were evaluated by comparing the listener's results with the ground truth information. 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 3.5 m away from the left wall and 1.2 m away from the front wall.

4.2. Test Results

The average and standard deviation of the azimuth angles identified by the 9 tested subjects are plotted in Figure 9-11. The diamonds represent the results of the proposed dynamic binaural audio system with a head tracking module, and the squares show the results of the conventional system that does not consider the listener's movement. The x-axis represents the ground truth angles and the y-axis represents the angles identified by the subjects. The ground truth or reference angles are also marked in the figures with cross marks. The system with judged angles closer to the reference is better.

Figure 9 shows the results when the listeners were at the center position. The virtual source between $-30°$ and $30°$ were mostly correctly identified. This is the easy case, because the virtual sound images were within the range of the two loudspeakers. In contrast, when the virtual sources were outside the range of the two loudspeakers, there were big mismatches between the ground truth and what the listeners perceived. One explanation to this phenomenon is that the HRTFs used in both systems were not personalized, hence they do not fit perfectly on each listener's head and ear shape. Another observation is that the results of the proposed system with head tracking and the conventional system were very similar. This is expected, since the listeners were asked to stay at the center position, which happened to be the sweet spot for the conventional system.

Figure 10 shows the results when the listeners were at 20 cm to the left from the center position. While the diamonds were similar to the previous results obtained at the center position, the squares were limited between $-30°$ and $30°$ of the y-axis. Since the subjects were away from the sweet spot, they identified the virtual source localized outside of the loudspeakers as

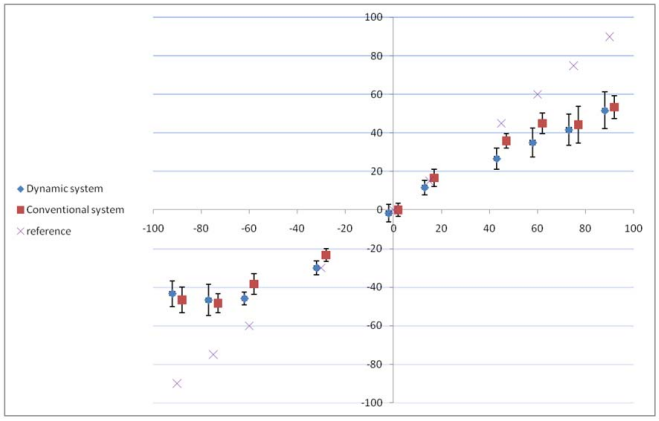

Fig. 9. Results when the listener is at center.

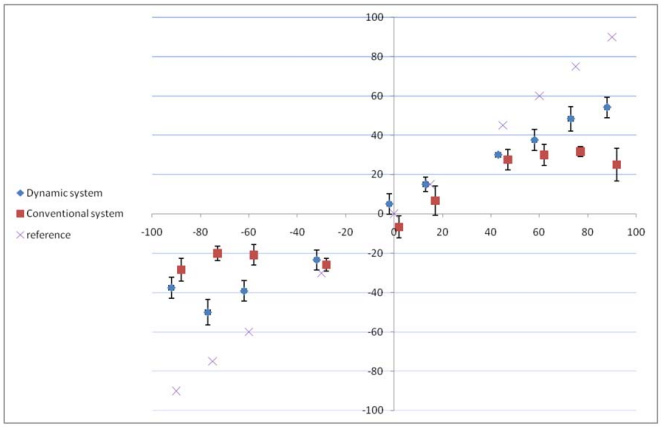

Fig. 10. Results when the listener is at 20cm left.

somewhere between $-30°$ and $30°$ when the conventional system was used. Even for the virtual sources between the two loudspeakers, the performance of the conventional system degraded. The squares for $0°$ and $15°$ were 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 proposed system with head tracking showed more robust performance than the conventional one in all aspects. Note the virtual sources located greater than 30 degree were identified more clearly compared to the ones less than -30 degree. Since the listeners were much closer to left speaker, it was much easier to reproduce the virtual sources on the right side than the ones on the left.

Figure 11 shows the results when the listeners were at 20 cm to the right from the center position. The overall trend is similar to Figure 10, i.e., the proposed system with head tracking still shows better performance than the conventional system. However, the results were not an flipped version of the previous results. We suspect this may have been caused by the geometry of the room used in this test, which was not symmetric centering around the listener's position (the right wall is much closer to the listeners than the left wall).

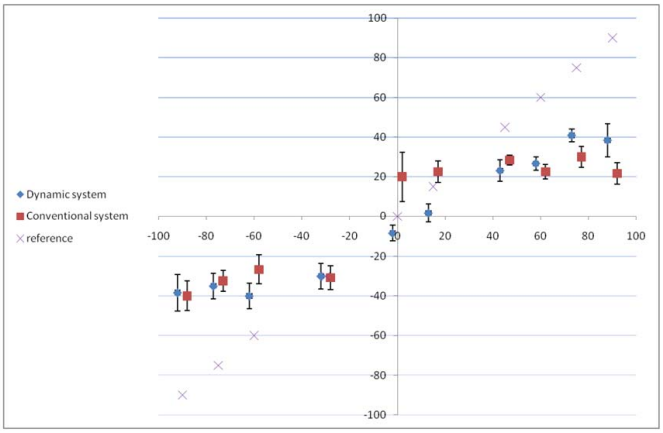

Fig. 11. Results when the listener is at 20cm right.

We further conducted the student t-test to assess whether mean 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 proposed algorithm is better than the conventional system is merely 0.19%, which shows that the difference is indeed statistically significant.

5. CONCLUSIONS

In this paper, we built a personal 3D audio system with head tracking using loudspeakers. By updating filters during dynamic binaural synthesis and dynamic crosstalk cancellation based on the movement of the listener, our system can steer the sweet spot to the position of the listener in real-time. An subjective study was conducted to compare the proposed system with the conventional system that does not monitor the listener's movement, and showed statistically significant improvements.

6. REFERENCES

[1] C.Kyriakakis, "Fundamental and technological limitations of immersive audio systems," Proc. IEEE, vol. 86, pp.941?951, 1998.

[2] V. Pullki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," J. Audio Eng. Soc., vol. 45, pp. 456-466, 1997.

[3] 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.

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

[5] J. Lopez and A. Gonzalez, "3-D Audio With Dynamic Tracking For Multimedia Environtments," 2nd COST-G6 Workshop on Digital Audio Effects, 1999.

[6] 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.

[7] T. Lentz, G. Behler, "Dynamic Crosstalk Cancellation for Binaural Synthesis in Virtual Reality Environments," J. Audio Eng. Soc., Vol. 54, Issue 4, pp. 283-294, 2006.

[8] 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.

[9] 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.

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

[11] C. Kyriakakis and T. Holman, "Video-based head tracking for improvements in multichannel loudspeaker audio," 105th Audio Engineering Society, San Francisco, CA, 1998.

[12] D. Cooper and J. Bauck, "Prospects for transaural recording," J. Audio Eng. Soc., vol. 37, pp. 3.19, 1989.

[13] J. Bauck and D. Cooper, "Generalized transaural stereo and applications," J. Audio Eng. Soc., vol. 44, pp. 683.705, 1996.

[14] 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.

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

[16] Q. Wang, W. Zhang, X. Tang and H.-Y. Shum, "Real-time Bayesian 3-D pose tracking," IEEE Trans. on CSVT, vo. 16, no. 12, Dec. 2006.

[17] C. Zhang and P. Viola, "Multiple-Instance Pruning for Learning Efficient Cascade Detectors," NIPS 2007.

[18] Y. Zhou, L. Gu, and H. J. Zhang, "Bayesian tangent shape model: Estimating shape and pose parameters via bayesian inference," in Proc. of CVPR, 2003.