Abstract — In the present study, a new stereophonic playback system was proposed, where the cross-talk signals would be reasonably cancelled at an arbitrary listener position. The system was composed of two major parts: the listener position tracking part and the sound rendering part. The position of the listener was estimated using acoustic signals from the listener (i.e., voice or hand-clapping signals). A direction of arrival (DOA) algorithm was adopted to estimate the directions of acoustic sources where the room reverberation effects were taken into consideration. A Cross-talk cancellation filter was designed using a free-field model. To determine the maximum tolerable shift of the listener position, a quantitative analysis of the channel separation ratio according to the displacement of the listener position was performed. Prototype hardware was implemented using a microprocessor board, a DSP board, a multi-channel ADC board and an analog frontend.

The results showed that the average mean square error between the true direction of a listener and the estimated direction was about 5 degrees. More than 80% of the tested subjects indicated that better stereo images were obtained by the proposed system, compared with the non-processed signals.

Index Terms — Sound rendering, sweet spot, direction of arrival, real-time hardware.

I. INTRODUCTION

When two loudspeakers are used to deliver binaural signals, the signals meant for the left ear are perceived by the right ear vice versa. This undesirable effect can be eliminated with a cross-talk cancellation filter [1]-[3]. Cross-talk cancellation is achieved under the assumption that the impulse responses from the loudspeakers to the listener's ears are known. These impulse responses depend on the listener's head location, therefore filters are designed with an estimated listener location. This means that good performance of the cross-talk canceller is guaranteed only for the limited area that is referred to as the “sweet spot”.

Recent advances in digital signal processing technology have led to the development of a stereophonic playback system that produces a good stereophonic illusion regardless of the listener's location [4]-[7]. In such a system, there are two major parts: automatic listener position tracking and sound rendering (or adjustment of the sweet spot) according to the estimated listener position. For tracking the listener position, several methods have been proposed to date. A vision-based listener-tracking system for accurate spatial sound reproduction was proposed by Kyriakakis et al. [4]. In this system, a number of procedures were involved to estimate the listener's head position including motion detection, the identification of the moving objects, and the identification of skin-color regions. With this method, however, huge computational loads of real-time image processing cannot be avoided. Moreover, the accuracy of head tracking was highly influenced by the illumination conditions of the underlying environment. Georgiou et al. [5] proposed a head detection algorithm for tracking the location of the listener's ears in real time using a laser scanner. In this approach, the accuracy of head detection and tracking was very high ( ≤ 15mm ). However, audio systems that used this method are expensive, since such an approach requires special equipment (e.g., a 2D laser radar system). Another method to estimate the location of the listener has been suggested by Kim et al. [6], where infrared and ultrasonic sensors were jointly used to estimate the listener's position along an x-y coordinate. This idea was successfully applied to digital television at low cost. With this system, however, the listener position cannot be estimated without a remote control.

In the present work, only an acoustic signal from a listener was used to estimate the listener position. Hence, with no additional equipment, such as a camera, a laser-tracker or a remote control, the listener position can be estimated by producing only an acoustic signals (e.g., a voice or hand-clapping). With this type of acoustic-based listener tracking, detection of the listener position was achieved by estimating the direction of arrival (DOA) from the acoustic source using two horizontally spaced microphones. Several practical issues associated with DOA estimation were discussed in the present study, including removal of the adverse effects caused by room reverberation.

The problem of adjusting the sweet spot to the listener's position can be formulated in terms of how to design the cross-talk cancellation filter for a given listener's location. Some of the information that is necessary for the design of a cross-talk cancellation filter includes a pair of the directional-dependent transfer functions that characterize the propagation of the waves through air from loudspeakers to listening points. The head-related transfer function (HRTF), which is the directional-dependent acoustical transfer
function from a sound source to a listener's eardrum has often been used for cross-talk cancellation [2], [3], [5], [6], [8], [9]. Differences between individual HRTFs, due to differences in ear shape and geometry, are known to strongly distort perception when non-individualized HRTFs are used for rendering [10]. Hence, personalization of the HRTF for a particular individual is a necessary condition for achieving good performance of cross-talk cancellation. However, the direct measure of HRTF is very time-consuming and requires a set of specific equipment. This indicates that, although the sweet spot can be accurately adjusted by using the HRTFs, such a system cannot be easily, thus inexpensively, implemented.

Instead of using a personalized HRTF, a generic HRTF (e.g. KEMAR HRTF by MIT media lab [11]) was used for cross-talk cancellation [2], [3]. In the present work, a free-field model was employed to approximate the path responses from the loudspeakers to the listening points. Since this approach is very simple and not user-specific, real-time hardware can be easily implemented. Moreover, this approach has the advantage of providing a simple analysis for the relationship between the performance of cross-talk cancellation and the displacement of the listener's position. As a result, a maximum tolerable shift of the listening position was provided in the present work.

The issue of real-time hardware implementation was also discussed in the present work. A microprocessor-controlled graphic user interface (GUI) and digital signal processor (DSP)-based real-time digital signal processing modules are the core parts of the developed audio system. In addition to the core parts, an MP-3 audio player was also included where the MP-3 audio files were downloaded through a wireless LAN module. A functional block diagram was also presented where the unique functionality of each module, as well as information communicated between the modules, was explained.

This paper is organized as follows. Section 2 presents a brief overview of the proposed system. In Section 3, the method for finding the listener position is presented. Section 4 provides the procedure for adjusting the sweet spot to the listener position. An overview of the prototype hardware system including a functional block diagram is presented in Section 5. Experimental results are presented in Section 6. Finally, future works are summarized in Section 7.

II. AN OVERVIEW OF THE PROPOSED AUDIO SYSTEM

In a binaural reproduction system using two loudspeakers, there are four paths between the loudspeakers and the listener's ears, as shown in Fig. 1. Each path can be characterized by the frequency responses $H_{LL}(f)$, $H_{LR}(f)$, $H_{RL}(f)$, and $H_{RR}(f)$. Ideally, the left channel signal is delivered to the left ear, and vice versa. However, since $H_{LR}(f)$, $H_{RL}(f) \neq 0$, under normal listening conditions, the cross-talk signals are perceived by each ear. If $H_{LL}(f)$, $H_{LR}(f)$, $H_{RL}(f)$, and $H_{RR}(f)$ are known, the cross-talk signals can be removed by using an adequately designed filter (cross-talk cancellation filter).

Note that the path responses $H_{LL}(f)$, $H_{LR}(f)$, $H_{RL}(f)$, and $H_{RR}(f)$ change when the listener moves (or, equivalently, the look direction of the user ($\theta$) is changed). Hence, the performance of cross-talk cancellation will degrade unless the parameters of the cross-talk cancellation filter are modified accordingly. If the system is able to track the listener position and to adjust the parameters of the cross-talk cancellation filter according to the listener position, the system always provides a given audio signal sound as if it had been played at the sweet spot. This scenario, which is the basic scheme of the proposed audio system, is depicted in Fig. 2. Listener tracking and cross-talk cancellation are the key functions to implement such an audio system. These two functions are described in more detail, as follows.

![Fig. 1. A binaural reproduction system using the two loudspeakers.](image1)

![Fig. 2. A block diagram of an audio system that can adjust the sweet spot to a listener position.](image2)
III. LISTENER TRACKING USING A DOA ALGORITHM

Several methods have been proposed to determine the direction, or position of a talker (or listener) in microphone array systems. Among these methods, the one using time delay estimation (TDE) between two direct signal paths from different microphones has the advantage of requiring a lesser number of microphones and is easier to realize [12]. In particular, the Generalized Cross-Correlation (GCC) method [13], where time delay estimation is achieved by maximizing the cross-correlation between a filtered version of the received signals, has been widely adopted due to its accuracy achieved with only a moderate degree of computational complexity. The underlying principle of the GCC-based time delay estimation is shown in Fig. 3. Let the two signals $x_1(t)$ and $x_2(t)$ denote the received signals from the two microphones that are a distance of $d$ apart. Assuming that the incoming signals are a plane wave and the distance from the source and each microphone is sufficiently longer than $d$, the relationship between the arrival angle $\theta$ relative to the sensor axis and the time delay between the two signals is as follows

$$\Delta t = \frac{\Delta x}{c} = \frac{d \sin \theta}{c}$$  \hspace{1cm} (1)

where $c$ is the velocity of sound at room temperature ($\approx 343$ m/s). Therefore, the arrival angle is obtained by

$$\theta = \sin^{-1}\left(\frac{\Delta t \cdot c}{d}\right)$$  \hspace{1cm} (2)

The above equation indicates that estimation of the time delay between the two signals is necessary to obtain the incident angle. In GCC-based time delay estimation, the estimated time delay is given by

$$\Delta t^* = \arg \max_{\Delta t} E[x_1'(t) x_2'(t - \Delta t)]$$  \hspace{1cm} (3)

where $E[\cdot]$ denotes expectation, and $x_1'(t)$ and $x_2'(t)$ denote filtered versions of $x_1(t)$ and $x_2(t)$, respectively:

$$x_1'(t) = x_1(t) * h_1(t)$$
$$x_2'(t) = x_2(t) * h_2(t)$$  \hspace{1cm} (4)

where $*$ denotes convolution. The impulse responses $h_1(t)$ and $h_2(t)$ reflect the acoustical paths from a sound source to the sensor locations of $x_1(t)$ and $x_2(t)$, respectively. In an echo-free environment, i.e. $h_1(t) = h_2(t) = \delta(t)$, the GCC can be approximated as $E[x_1(t) x_2(t - \Delta t)]$. For a real situation, however, the approximated GCC cannot guarantee a reliable estimate of time delay due to room reverberation. The example waveforms are shown in Fig. 4, where the two hand-clap signals were recorded in a small room (5mx4mx2m, WDH) and in an anechoic chamber (4mx3mx1.5m, WDH), respectively. For the signals recorded in a small room, the reverberation effects were noticeable. As shown by the right waveform in Fig. 4, it appears that the shape of the reverberation-free hand-clap signal is very close to the impulse function. Hence, the hand-clap signal recorded in the reverberant room can be approximated as the impulse response of the underlying room.

As shown in Fig. 4, the length of the reverberated signals were very long ($\approx 10000$ samples in case of 48kHz sampling rate). This means that if the room impulse response is represented by the finite impulse response (FIR) digital filter, a large number of filter taps should be employed. Since the memory size of the real-time system was limited, a shorter filter length was desirable. Hence, an infinite impulse response (IIR) digital filter was employed to represent the room impulse response. In a previous work on modelling room reverberation [14], an auto-regressive (AR) model was also employed for approximating room transfer function. Hence, the hand-clap signals recorded at the $i$-th sensor in the reverberation room could be given by

$$x_i(t) = -\sum_{p=1}^{P} a_i(p) x_i(t - p) + e_i(t), \hspace{1cm} 1 \leq t \leq T$$  \hspace{1cm} (5)

where $a_i(p)$ denotes the $p$-th AR coefficient for the $i$-th channel signal, which is generally known as the linear
prediction (LP) coefficient. $P$ is the order of the AR model and $e_i(t)$ is the residual signal. The AR coefficients $\{a(p), 1 \leq p \leq P\}$ were obtained by minimizing the sum of the squared residual signals. Dereverberation was achieved by inverse filtering using an estimated room impulse response, as follows

$$e_i(t) = x_i(t) + \sum_{p=1}^{P} a(p) x_i(t-p)$$  \hspace{1cm} (6)

The GCC in (3) was computed using $e_i(t)$, instead of $x_i(t)$. Note that the transfer function due to the acoustics of a room generally do not change considerably with time, but do vary with the spatial locations of the sound source and sensors [14]. Hence, the room impulse response was separately estimated for each channel.

Two more techniques were employed in the present work to improve the accuracy of direction estimation. First, an oversampling technique was adopted to increase the time-resolution of the time-delay estimation. In this work, the ratio of oversampling was two, and a Fourier Transform (FT)-based interpolation method was employed. Second, direction estimation was carried out on the eight independent signals from the horizontal linear array using the signal-arrival time of the GCCs. The experimental results showed that a high directional resolution was not a necessary condition for the reliable adjustment of the sweet spot. Hence, the number of the possible directions was limited to 33 in the present study. This number proved to be sufficient for the purpose of adjusting the sweet spot.

**IV. ADJUSTING THE SWEET SPOT**

In the present study, the sweet spot was adjusted by updating the parameters of the cross-talk cancellation filter for a given listener location. For a stereophonic reproduction environment, as shown in Fig. 5, the frequency domain representation of the signals observed at either ear are given by

$$X_L(f) = H_{LL}(f)S_L(f) + H_{RL}(f)S_R(f)$$
$$X_R(f) = H_{LR}(f)S_L(f) + H_{RR}(f)S_R(f)$$  \hspace{1cm} (7)

where $S_L(f)$ and $S_R(f)$ are the input signals to the left and right signals, respectively. $H_{LL}$, $H_{RL}$, $H_{LR}$, and $H_{RR}$ are the frequency responses for each path, which are shown in Fig. 1. Equation (7) can be represented by a matrix form as follows,

$$X = HS$$  \hspace{1cm} (8)

where $X$, $H$, and $S$ are the observation matrix, the transfer matrix, and the signal matrix, respectively, which are given by

$$X = [X_L(f) X_R(f)]^T$$
$$S = [S_L(f) S_R(f)]^T$$
$$H = \begin{bmatrix} H_{LL}(f) & H_{RL}(f) \\ H_{LR}(f) & H_{RR}(f) \end{bmatrix}$$  \hspace{1cm} (9)

Cross-talk cancellation is performed by multiplying $S$ by the cross-talk cancellation matrix $C$. The signals observed at either ear are then given by

$$\hat{X} = HC$$  \hspace{1cm} (10)

The cross-talk cancellation matrix $C$ is given by the inverse of the matrix $H$. In this case, $\hat{X} = S$. This means that cross-talk cancellation requires knowledge of the frequency responses from the loudspeakers to the listener's ears. In the present study, a free-field model [2], [3], [15] was employed to approximate the frequency responses for each path. Accordingly, $H$ was given by

$$H = \frac{\rho_0}{4\pi} \begin{bmatrix} e^{-j2\kappa d_{LL}/l_{LL}} & e^{-j2\kappa d_{RL}/l_{RL}} \\ e^{-j2\kappa d_{LR}/l_{LR}} & e^{-j2\kappa d_{RR}/l_{RR}} \end{bmatrix}$$  \hspace{1cm} (11)

where $\kappa$ and $\rho_0$ represent the wave number and the density. Note that $\kappa = 2\pi f/c_0$ where $c_0$ is the speed of sound. In this study, it was assumed that $\rho_0 = 1.21$ kg/m$^3$, $c_0 = 343$ m/s. $l_{LL}$, $l_{LR}$, $l_{RL}$, and $l_{RR}$ are the lengths from the loudspeakers to the ears, which were calculated as

$$l_{LL} = \sqrt{\frac{d_x^2 + y_x^2 + y_u^2}{2} + y_u^2}$$
$$l_{LR} = \sqrt{\frac{d_x^2 + y_x^2 + y_u^2}{2} + y_u^2}$$
$$l_{RL} = \sqrt{\frac{d_x^2 - y_x^2 + y_u^2}{2} + y_u^2}$$
$$l_{RR} = \sqrt{\frac{d_x^2 - y_x^2 + y_u^2}{2} + y_u^2}$$  \hspace{1cm} (12)

where $y_u$, $d_x$, $r$, and $\theta_u$ denote the distance of the listener in the y-direction, the distance between the two loudspeakers, the radius of the listener head, and the look direction of the listener, respectively.
listener. In the present study, the radius of the listener head was assumed to be 0.07 m [2].

Note that $d_s$ and $y_u$ should be given to obtain the matrix $C$. In a real situation, the distance between the two loudspeakers are almost always fixed. Hence, $d_s$ can be set by a user prior to using the audio system. However, it is very questionable that such a pre-setting method can also adopted for $y_u$, since the location of the user is not always fixed. In the present study, for a given $d_s$, how the performance of cross-talk cancellation was affected by displacement from a given $y_u$ was investigated. The performance of cross-talk cancellation was evaluated using the channel separation ratios (CSR) [1]-[3] for each ear, which are given by

$$CSR_l = 10 \log_{10} \left[ \frac{|G_{11}(f)|^2}{G_{12}(f)} \right] df$$

$$CSR_r = 10 \log_{10} \left[ \frac{|G_{21}(f)|^2}{G_{22}(f)} \right] df$$

where

$$G = \begin{bmatrix} G_{11}(f) & G_{12}(f) \\ G_{21}(f) & G_{22}(f) \end{bmatrix} = HC$$ (14)

In the present study, the integral interval $B$ was [20Hz, 1.5kHz] which is the frequency range that is free from the head-shadowing effects [16]-[18]. Note that the CSR is $\infty$, when $C = H^{-1}$.

Let $H(y_u)$ be the transfer matrix when the location of the listener in the $y$-direction is set at $y_u$. Now, if the listener moves $y_u$ to $y_u + \Delta y_u$, the matrix $G$ is given by

$$G(\Delta y_u) = H(y_u + \Delta y_u)H^{-1}(y_u)$$ (15)

In this case, although the real listener position is $y_u + \Delta y_u$, the cross-talk cancellation is performed for the position $y_u$. Fig. 6 shows the contour plots of CSRs at the left and right ears, in the case of $d_s = 4m$ and $y_u = 4m$. The previous study indicated that a choice of 10 dB for cross-talk cancellation was a reasonable estimate of the amount needed in order to ensure a desired subjective perception [3]. The contour plots of CSRs indicate that the performance of cross-talk cancellation was not seriously degraded by a large displacement of the user position in the $y$-direction, when the look direction of the user was relatively small ($< 20^\circ$). As the absolute angle of the user direction was increased, the width of the area with more than 10 dB CSR was reduced. However, when the absolute look direction of the listener was less than 45 $^\circ$, the tolerable shift in the $y$-direction was about 1m. The authors feel that these settings are reasonable for practical use.

Consequently, the pre-set method was also adopted for determination of $y_u$ in the present study.

V. IMPLEMENTATION OF THE PROTOTYPE SYSTEM

A block diagram of the prototype hardware is shown in Fig. 7. The major parts of the prototype hardware are the microprocessor module, the digital signal processor (DSP) module, the LCD/touch panel module, and the 8-channel microphone amplifiers/ADC modules. In the microprocessor module, a 32-bit RISC (266MHz version) was used, which included 64MB RAM memory and 64MB NAND flash memory. The memory space of RAM was partitioned into an operating system block (32MB) and an application program/data block (32MB). Three universal serial bus (USB) ports were available, which were connected to the USB wireless LAN module, the host computer (for development), and an external NAND flash memory (for storing the MP3 files). Audio signals were produced from a general purpose stereo audio codec, which produced input signals for the DSP module. A 4.3in. LCD/touch panel was used for user communication.

The DSP module was composed of a 32-bit floating-point DSP (225MHz version), a 16-bit stereo codec, a 16MB SDRAM and a 512KB Flash memory. The audio input of the DSP module was selected from the two external audio inputs and the audio output of the microprocessor module. The output of the DSP module was volume-controlled and could be connected to the general-power amplifier. The DSP module communicated with the microprocessor module through an inter-integrated circuit (I2C).

Eight electret condenser microphones (ECM) were used to find the user position, which was connected to the eight-channel microphone amplifier and converted to a digital signal using the ADC board. The ADC board consisted of a one-chip,
eight-channel delta-sigma ADC, an 8Mb high-speed SRAM and CPLD. The digital data was transmitted over an external memory bus to the DSP board. The incoming microphone signal was sampled at 48kHz with 16-bit precision. The memory size of the ADC module was 1MB. Hence, about 1.3 seconds of the 8-channel audio data could be stored.

The functional block diagram of the proposed audio system is shown in Fig. 8. Basically, the microprocessor module had a UI function, overall system control, and MP-3 decoding. An application program was developed using an embedded C++ language (version 4.0). Since the prototype system was self-playing, it was not always necessary to prepare an external audio player. The MP3 files could be downloaded from an external PC through a wireless LAN.

In the DSP module, a segment containing a valid hand-clapping signal was first extracted. This was achieved by comparing the short-time energy values with a given threshold. For the extracted segment, the LP analysis was carried out to remove the reverberation effects. The dereverberated signal was then oversampled by a factor of 2. The DOA was estimated from the oversampled signals. The estimated arrival angle was transferred to the microprocessor module to display the current position of the listener. The parameters for adjusting the sweet spot were computed in the DSP module, using the estimated DOA, the distance between the speakers and the distance from the listener. Note that the distance between the speakers and the distance from the listener were transferred from the microprocessor module that was set through the UI. The sampling frequency of the audio signals was 44.1kHz. The entire length of the program codes for DOA estimation and sweet spot control was about 171KB. Hence, the DSP program ran in the internal RAM of the DSP. Hence, the sweet spot adjusting program ran in the internal RAM of the DSP.
A photo of the front of the LCD/touch panel is shown in Fig. 9. In this photo, "anywhere mode" indicates a function of adjusting the sweet spot to the listener position. By touching the anywhere button, the user can select either of two modes: a normal playing mode and a sweet spot adjusting mode. The front and internal views of the prototype audio system developed herein are shown in Figures 10 and 11, respectively. A 4.3 in LCD/touch panel was placed at the center of the front panel, and the volume control and input select knobs were placed at the left and the right sides of the panel, respectively. A hairline-decorated aluminum alloy flat panel was used for making the front panel.

**VI. RESULTS**

**A. Evaluation of DOA estimation**

The performance of user direction finding method was evaluated by the ratio between the number of the correctly identified stimuli and that of the total stimuli for each direction. A total of fifteen directions were evaluated. The hand-clapping signals were produced by one subject. The signals were sampled in two environments, including an anechoic chamber (4mx3mx1.5m, WDH) and a home living room (8mx6mx2.5m, WDH). For each direction, the hand-clapping signals were recorded five times. Hence, a total of 75 stimuli were evaluated. The distance from the center of the microphone array to the subject was a constant 2m. The length of the frame for DOA estimation was 21.3msec (1024 samples). The order of the AR model was 24, which was determined from the experimental results.

Fig. 12 shows the plots of the estimated direction versus the real directions for the echo-free signals, the signals recorded in the living room, and the signals dereverberated by the AR model. The correlation and average root mean squared error (RMSE) between the estimated and real direction are also shown at the top of each figure. The results showed that the highest accuracy was achieved when DOA estimation was carried out on the echo-free signals. In this case, a correlation of 0.9936 and a RMSE of 4.9509 were obtained. As shown in Fig. 12, the deviations of the estimated direction from the real direction were almost uniform across directions.

As for the signals recorded under the reverberation environment, the performance was degraded compared with the echo-free signals. The deviations of the estimated direction were also increased. When the reverberation effects were removed by the LP model, a correlation of 0.9933 and a RMSE of 5.0771 were obtained. These were very close to those obtained from the echo-free signals. In the range of \([-10^\circ \text{ to } 30^\circ]\), the estimated directions were exactly the same as the real ones, as shown in Fig. 12. Although larger errors were observed in the case of \(\theta = \pm 40^\circ\), the LP-based reverberation removal proved to be an effective method in terms of the accuracy of direction estimation.
To this point, all results were obtained in otherwise silent environments. However, in most practical situations, the direction location should be accomplished while the music is playing. In this case, the music signal could be regarded as a noise signal that will affect the performance of direction estimation. Fortunately, the experimental results showed that the performance of the listener tracking was not seriously degraded when a moderate level of background signal was presented. This was due to the fact that the levels of hand-clapping signals were relatively higher than that of the background signals. When the level of the background signals was very high, the estimated listener direction had large errors. A more robust estimation of the listener position would be achieved by selective filtering of the hand-clapping signals from the noise signals or, equally, by removal of the background signals from the incoming signals. An adaptive filter approach would be useful for this purpose.

B. Subjective evaluation in sweet spot adjustment

The subject evaluation was carried out for the 14 directions presented in Fig. 13. The listening test was performed in a home living room (8mx6mx2.5m, WDH). Twenty subjects participated in this test. All subjects had normal hearing ability. For each direction, each subject listened to five types of programming including classical music, popular music, piano solo performances, Korean popular music, and a male voice. Hence, the total number of stimuli for each direction was 100. To alleviate the effects of listener fatigue, the listening tests were divided into 4 sessions. Each subject was presented with 25 stimuli in each session. The procedure for subject evaluation was as follows: 1) The subjects listened to the programs using equalized headphones-these stimuli were regarded as reference signals that have no cross-talk signal; 2) subjects moved to the position where the subject evaluation was carried out and the stimuli (not processed) were presented through the loudspeakers; 3) the sweet spot-adjusted stimuli were presented at the same position, through the same loudspeakers-as in point #2-then, subjects were asked to judge whether the third stimulus was perceived closer to the first stimulus compared with the second stimulus. Each subject was allowed to listen to the stimuli as many times as needed before the decision. Note that the sweet spot was adjusted to the estimated direction of the listener. We assumed that if the third stimulus was judged as closer to the first stimulus by the subject, the sweet spot was correctly adjusted. The ratio between the number of correctly adjusted stimuli and the total number of stimuli was computed for evaluation. In this listening test, a high-fidelity monitor speaker was used and the distance between the two speakers was set at 2.5m.

The results are summarized in Fig. 13; the radius of each circle for each direction was directly proportional to the correct adjustment ratio. The overall adjustment ratios computed from a total of 100 stimuli for each direction are also displayed in each figure. The overall correct adjustment ratio was 80%. All subjects who participated in the listening test clearly indicated that the sweet spot-adjusted stimuli sounded different than the original stimuli (not adjusted) and that the proposed system yielded preferred sound quality for the stereophonic playback of music, in terms of stereophonic illusion. The correct adjustment ratios were relatively low when the subjects were located near the symmetry axis between the two speakers (azimuth angles of ±10°). This was mainly due to the fact that for these directions, the differences in stereo image were not clearly perceived by the subjects. We also observed that the performance was slightly degraded when the subjects were located near each speaker. At these positions, the precedence effects strongly affected the subjects' decisions. Some subjects indicated that the sweet spot adjusted stimuli (the third stimuli) sounded clearly different from the unadjusted stimuli (the second stimuli). However, the third stimuli clearly were not close to the first stimuli.

VII. Future Works

In the proposed system, the listener's position was estimated as a way to find the look direction of the listener. Although this method yielded reasonable performance in terms of adjusting the sweet spot, some limitations were revealed. First, a large displacement for the listener position in the y-direction did not allow adjustment of the sweet spot, especially when the absolute look direction was large. This indicated that a method of tracking the listener position in both the x- and y-directions would be desirable. With such a method, however, the microphones should be installed outside the system. Using external microphones is less comfortable for users, since the too many connection cables are necessary. Hence, our future work will focus on issues associated with way to obtain a listener position in both the x- and y-directions without using microphones installed outside the system.

REFERENCES


