A Spatial Audio System Using Multiple Microphones on a Rigid Sphere

Taejin Lee, Daeyoung Jang, Kyeongok Kang, Jinwoong Kim,
Dae-Gwon Jeong, and Hareo Hamada

The main purpose of a spatial audio system is to give a listener the same impression as if he/she were present in a recorded environment. A dummy head microphone is generally used for such purposes. Because of its human-like shape, we can obtain good spatial sound images. However, its shape is a restriction on its public use and it is difficult to convert a 2-channel recording into multi-channel signals for an efficient rendering over a multi-speaker arrangement. In order to solve the problems mentioned above, a spatial audio system is proposed that uses multiple microphones on a rigid sphere. The system has five microphones placed on special points of the rigid sphere, and it generates audio signals for headphone, stereo, stereo dipole, 4-channel, and 5-channel reproduction environments. Subjective localization experiments show that front/back confusion, which is a common limitation of spatial audio systems using the dummy head microphone, can be reduced dramatically in 4-channel and 5-channel reproduction environments and can be reduced slightly in a headphone reproduction

Keywords: 3D audio, sphere microphone, HRTF, dummy head microphone.

Manuscript received June 9, 2004; revised Mar. 10, 2005.

This work is the result of the collaboration research between ETRI, Tokyo Denki University, and DiMAGIC Co., Ltd. This work is being supported by the Ministry of Information and Communication, Korea, under the title of "SmarTV Technology Development."

Taejin Lee (phone: +82 42 860 5713, email: tjlee@etri.re.kr), Daeyoung Jang (email: dyjang@etri.re.kr), Kyeongok Kang (email: kokang@etri.re.kr), Jinwoong Kim (email: jwkim@etri.re.kr) are with Digital Broadcasting Research Division, ETRI, Daejeon, Korea.

Dae-Gwon Jeong (email: dgjeong@mail.hangkong.ac.kr) is with the Department of Avionics Engineering, Hankuk Aviation University, Goyang, Korea.

Hareo Hamada (email: hamada@sie.dendai.ac.jp) is with the School of Information Environment, Tokyo Denki University, Tokyo, Japan.

I. Introduction

An ideal spatial audio system would produce the illusion of hearing sounds as if a listener were actually present in a recorded environment. The sound is actually created by the headphone or loudspeakers, but a listener's perception is that the sounds are coming from all around him. While it is possible to use real human ears for binaural recording (using probe microphones inserted into the eardrums of a person), it is difficult to mount high-quality microphones in the ears, and the head movements and noise of the person can be obstructive. Dummy head microphones are models of human heads with pressure microphones in the ears that can be used for originating binaural signals suitable for measurement or reproduction [1], [2]. Binaural recording and reproduction systems using a dummy head have been used in various fields such as virtual reality systems, transaural reproduction systems for entertainment, and in the evaluation of many sound fields [3]-[5]. The idea behind the binaural technique is that the input to hearing consists of two signals—sound pressures at each of the eardrums. If these are recorded in the ears of a listener and reproduced exactly (usually through a headphone), then a complete auditory impression is recreated, including spatial aspects such as direction and distance of the sound sources [6]. A head related transfer function (HRTF) of the dummy head is mainly determined by geometric shapes of the pinna, head, and torso of the dummy head. Therefore, its HRTFs can be viewed as averaged and fixed characteristics for the sets of HRTF under consideration. Despite its high quality of spatial sound images, a dummy head may cause several problems such as an elevation of the front image, coloration, and reversals. Among these problems, a common limitation of the dummy head is

reversals: sound sources in the front hemisphere are often heard as if they were placed in the rear hemisphere [7]-[9].

The reproduction of binaural recordings through loudspeakers, referred to as "transaural audio" [10]-[12] overcomes the problem of inside-head localization of virtual acoustic images but not the problem of reversals: reproduction with two front loudspeakers also causes reversals of virtual acoustic images [13]. There are two kinds of reversals: back-to-front reversals generally occur in the case of frontal loudspeaker reproduction, whereas front-to-back reversals usually occur with headphone reproduction. The explanation for this ambiguity might be that commercial dummy heads do not replicate exactly individual shapes of the human head. Also, listeners use small head movements to distinguish between front and back images, while binaural recording is undertaken with a fixed dummy head position [14], [15]. The available commercial dummy heads have not been standardized yet, and consequently different dummy heads produce different performances [16]-[18].

Sometimes, human heads are approximated by using a rigid sphere, which simulates the shadowing effect of the head [19]. Recordings made using such approaches have been found to have reasonable loudspeaker compatibility as they do not have unusual equalization that results from pinna filtering. Binaural signals recorded by a dummy head will typically suffer two stages of pinna filtering when they are replayed through loudspeakers—once on recording and then again on reproduction—giving rise to distorted timbral characteristics. There are products which place two microphones at the opposite places on a rigid sphere to increase spatial images for stereo reproduction [20], [21]. However, because of a sphere's symmetrical characteristics, a sound source in the frontal hemisphere plane has the same sound image as one in the rear hemisphere plane which has the same distance and mirrored angle with the source in frontal hemisphere plane. So there exist more severe reversals. The previous study [22] on multiple microphones using a rigid sphere adopted four microphones and 4×4 inverse filters to reproduce sound images in a 4-channel environment. This system also reduced the front/back confusions but only considered the 4-channel reproduction environment.

Our goal is to provide a spatial audio system which can resolve the so-called reversal problem and produce audio signals adapting to 5-channel, 4-channel, stereo/stereo dipole, and headphone reproduction environments. The processing procedure of our system is shown in Fig. 1. For acquisition of acoustical sources, we placed five microphones on the horizontal plane of a rigid sphere. The positions of four side microphones and a center microphone are chosen to reflect a slight head movement and to increase frontal virtual images, respectively. A post processing module uses inverse filters for various kinds of reproduction environments. In this part, the audio signals acquired from the multiple microphones are processed via a matrix of linear filters in order to produce corresponding reproduction signals. The matrixes of linear filters are 5×5 for 5-channel, 4×4 for 4-channel, and 2×2 for stereo/stereo dipole reproduction environments generated using a method of frequency domain deconvolution with regularization (fast deconvolution) [23]. These filters are designed to ensure that the recorded signals are generated at the same microphone positions on the surface of the sphere placed at the position of a listener's head in reproduction environments. The post processed signals are those for 5-channel, 4-channel, headphone, stereo, and stereo dipole [24], [25] reproduction environments. The 5-channel reproduction environment follows the ITU 5.1 loudspeaker configuration [26] except the subwoofer for low frequency effect (LFE). The 4-channel reproduction also follows the ITU 5.1 configuration except for the center and subwoofer. The 3D audio reproduction parts reproduce the post-processed audio signals. For the generation of various inverse filters, we measured a rigid sphere's impulse responses in an anechoic chamber. Using these impulse responses, we generated virtual sources using mono sources for a localization experiment. The subjective localization experiments were performed in the anechoic chamber to validate the system's performance. The results of our experiments show that our system can resolve the front/back confusions in 5-channel and 4-channel reproduction environments and can get similar results in a headphone reproduction as a binaural recording.

A 3D audio acquisition method using multiple microphones on a rigid sphere and a post processing and reproduction method are described in sections II and III, respectively. Section IV describes the subjective localization experiments and results. Finally, conclusions and future work are presented in section V.

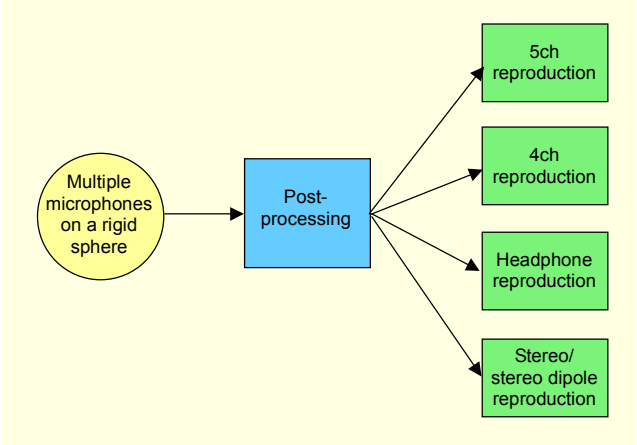


Fig. 1. The processing procedure of the spatial audio system.

II. 3D Audio Acquisition Using Multiple Microphones on a Rigid Sphere

1. Measurement of a Sphere's Impulse Responses

Impulse responses produced by a sphere were measured in order to calculate various inverse filters for post processing. The measurements were performed in Tokyo Denki University's anechoic chamber. The sphere was mounted 1.2 m above the floor, and a loudspeaker was positioned 1.4 m from the origin of the sphere in the manner of the KEMAR dummy HRTF measurements by Bill Gardner and others [27]. For the measurement of impulse responses, we used the AEIRM impulse response measurement system [28]. The impulse responses were obtained using TSP (Time Stretched Pulse) sequences. We measured a horizontal plane's impulse responses by rotating the sphere by 5 degrees and obtained 72 impulse responses. In order to compensate for the non-uniform responses of both the loudspeaker and microphone, a free-field impulse response was measured and deconvolved from the results of the sphere's impulse responses. Figure 2 shows examples of the measured impulse responses. As we can see in Fig. 2, two impulse responses with the same distance and mirrored angle are the same, so front/back confusion is common in recording using a rigid sphere with two microphones at the opposite place of a sphere's horizontal plane. To resolve this problem, our system uses five microphones on a rigid sphere's horizontal plane, consequently changing the responses of the mirrored direction.

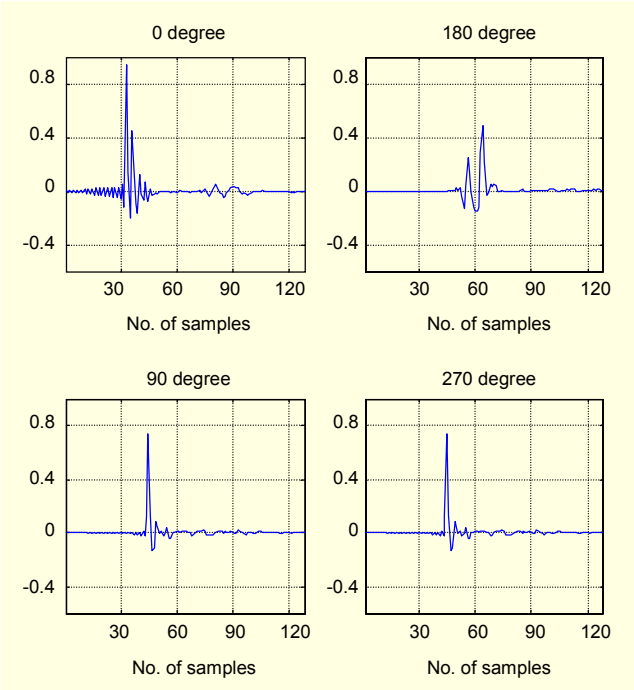


Fig. 2. Impulse responses of a rigid sphere (amplitude vs. no. of samples).

2. The Layout of Microphones on a Rigid Sphere.

The layout of five microphones on a horizontal plane of a rigid sphere is shown in Fig. 3. A center microphone (No. 1) is placed in front of the rigid sphere for increasing the frontal virtual image, and the left two microphones and right two microphones are placed at the left and right sides of the rigid sphere with a 30 degree gap in the horizontal plane. In Fig. 3, u_i indicates the i-th microphone's output signal.

For human hearing, the onset time of a sound will be different for each ear. This is referred to as the interaural time difference (ITD). Psychophysical experiments have shown that these localization cues are effective only in the range below 1.5 kHz. Another mechanism known as the interaural level difference (ILD) may be used. Since the head is a relatively dense medium, it will tend to cast an acoustical shadow on the ear contralateral to the sound sources. The attenuation has been measured to be just over 40 dB for frequencies above 3 kHz [29].

Generally, one rotates his head to detect the sound direction,

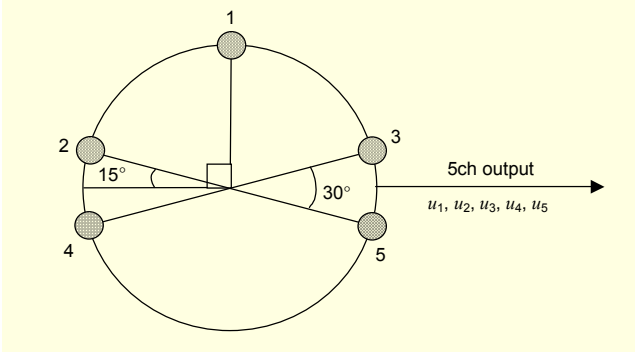


Fig. 3. Layout of microphones on a rigid sphere.

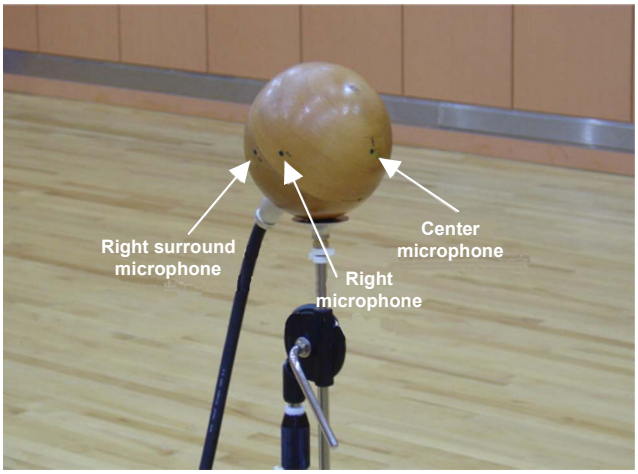


Fig. 4. The prototype of a sphere microphone.

since this consequently changes the ITD and ILD. In the median plane of the head, if sound comes from the same distance of the front/back positions, there is no difference in the ITD and ILD between left and right ears, so it is difficult to decide the direction in this case without rotating the head. Therefore, the positions of the four side microphones are chosen for reflecting a slight head movement, with an average of 30 degrees. The position of the four side microphones correspond to a slight head rotation of the listener in the reproduction environment. During the reproduction, the listener can rotate his head and get different sound sources through these microphones, which can help in determining the front/back sound sources. Figure 4 shows the prototype of a sphere microphone that consists of multiple microphones on a horizontal plane.

III. Post Processing and 3D Audio Reproduction

1. Fast Deconvolution

Binaural signals are often used in virtual reality systems to generate a three dimensional sound field. Generally, we acquire spatial sound sources using a dummy head and reproduce them through a headphone. Although the reproduction of 3D sound through the headphone works well, this way has some disadvantages such as in-head localization. To overcome these disadvantages, we use a pair of loudspeakers instead of a headphone. When reproducing 3D audio signals using loudspeakers, we have to cancel so-called crosstalk. Crosstalk cancellation involves the acoustical cancellation of an unwanted signal from the left loudspeaker to the right ear and vice versa. There are various methods for the design of a crosstalk cancellation filter [30]-[33]. In our system, we use the "Fast Deconvolution method using Regularization," which is suggested by Ole Kirkeby and others [23]. The fast deconvolution algorithm combines the well known principles of least squares inversion in the frequency domain [34] with the zeroth order regularization method [35], which is traditionally used when one is faced with an ill-conditioned inversion problem [36], [37].

The purpose of multi-channel deconvolution is to derive the appropriate signals to be played over a set of S loudspeakers, so that a desired sound field is reproduced at R target points in space as accurately as possible. In our system, we need 5×5, 4×4, and 2×2 inverse filters for 5-channel, 4-channel, and stereo or stereo dipole reproduction, respectively.

Figure 5 shows a signal processing flow of a multi-channel sound reproduction system. From Fig. 5, **u(z)** is a vector of T observed signals, **v(z)** is a vector of S source input signals, **w(z)** is a vector of R reproduced signals, **d(z)** is a vector of R desired signals, and **e(z)** is a vector of R performance error signals. We hope to reproduce **u(z)** in plant **C(z)** at R points as closely as possible by an optimal designing of inverse filter **H(z)**, which can compensate for plant transfer function **C(z)**.

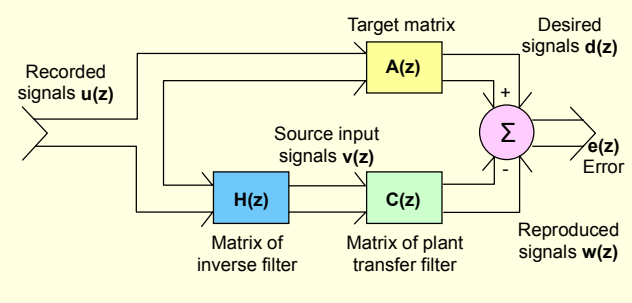


Fig. 5. The block diagram of a discrete-time multi-channel sound reproduction system.

Figure 5 contains three primary matrices that define the operation of this system; these are $R\times T$ target matrix **A(z)**, $R\times S$ plant matrix **C(z)**, and $S\times T$ inverse filter **H(z)**. The **C(z)** matrix is comprised of the impulse response of the system configuration as measured at each target point while using each loudspeaker independently. The output of the **A(z)** matrix is a vector containing the ideal or desired signals [14]. So **A(z)** represents the desired model of the system. Individually, the **C(z)** and **H(z)** matrices provide a limited amount of information; however, when they are convolved, the resulting matrix **w(z)** indicates the quality of the crosstalk cancellation. The fast deconvolution algorithm is for determining a matrix of causal inverse filters **H(z)** given **C(z)** and **A(z)**. The T inputs that comprise the **u(z)** vector are processed through the **H(z)** filter matrix and results in S sources in the vector **v(z)**. This vector contains the signals targeted for the loudspeaker.

In the special case where the desired signals **d(z)** are identical to the observed signals **u(z)**, the matrix **A(z)** is an identity matrix of order $R=T$, so the optimal inverse filters are given by

$$\mathbf{H}_I(k) = [\mathbf{C}^H(k)\mathbf{C}(k) + \beta\mathbf{I}]^{-1}\mathbf{C}^H(k), \qquad (1)$$

where k denotes the k-th frequency index (refer to [23] for more details).

From (1), β is a regularization parameter for control of the inverse filters. If β is too small, there will be sharp peaks in the frequency response of the inverse filter, and if β is too large, the deconvolution will not be very accurate. Fortunately, though, the exact value of β is usually not critical [23]. Ultimately, a subjective judgment is necessary in order to determine whether the value of β is acceptable. For

determining the value of β, a simple trial-and-error experiment is enough.

In our system, for 5×5 inverse filters, $\beta = 0.0001$; for 4×4 inverse filters, $\beta = 0.0005$; for stereo inverse filters, $\beta = 0.01$; and for stereo dipole, $\beta = 0.005$.

In 5×5 inverse filtering, we hope to reproduce $u_{1\sim5}$ signals at the rigid sphere's five points as closely as possible, as shown in Fig. 6. In 4×4 inverse filtering, we hope to reproduce $u_{2\sim5}$ signals at the rigid sphere's four points (same as 5×5 except no. 1 point shown in Fig. 6) as closely as possible. In 2×2 inverse filtering, we hope to reproduce 2-channel headphone signals at the rigid sphere's two points (left and right 90 degrees) as closely as possible.

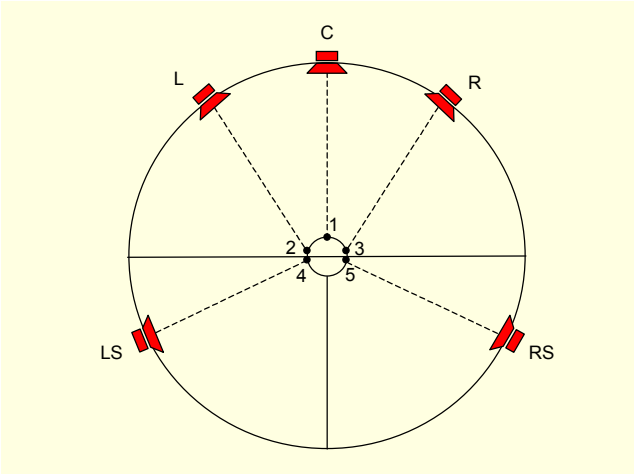


Fig. 6. 5-channel inverse filtering.

2. Inverse Filtering for Reproduction through Loudspeakers

The post processing part generates various reproduction signals using convolution between source input signals and filters. Figure 7 shows all of these post processing and notations of the resulted signals. In Fig. 7, u_i indicates i-th microphone's output signal (recorded signal). The v indicates various reproduction signals, which are generated using convolution between recorded signals and inverse filters. In our system, we have two options for the generation of headphone reproduction signals: one uses the 5-channel recorded signals $u_{1\text{-}5}$ and the other uses 5-channel reproduction signals (inverse filtered signals v_C^{5ch}, v_L^{5ch}, v_R^{5ch} v_{LS}^{5ch}, v_{RS}^{5ch}). So for the generation of stereo/stereo dipole reproduction signals, a user can choose one of them.

A. Inverse Filtering for 5-Channel Reproduction

The 5-channel reproduction configuration follows the ITU 5.1 standard except for the subwoofer for LFE. For the 5-channel reproduction, we convolve $u_{1\text{-}5}$ recorded signals with

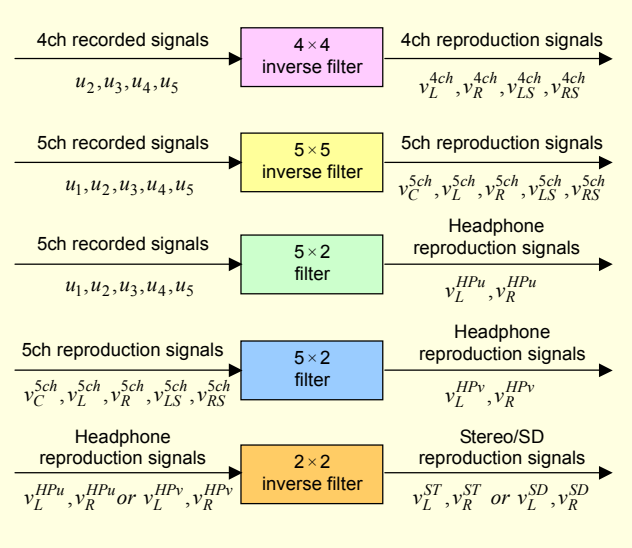


Fig. 7. Block diagram of post processing.

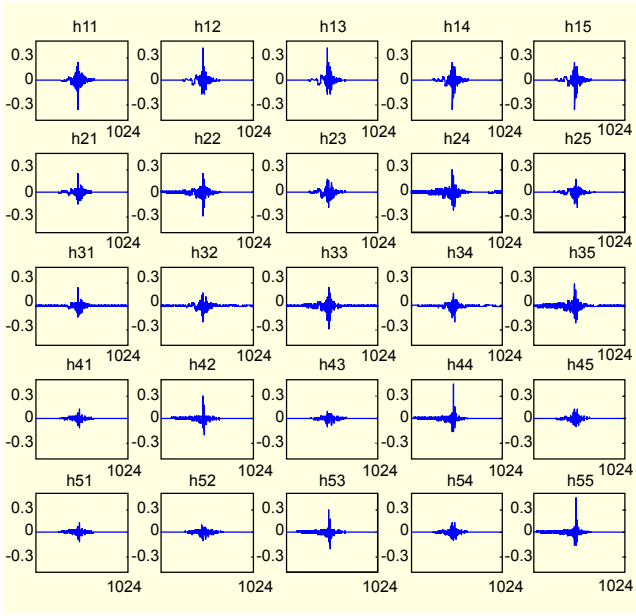


Fig. 8. Inverse filter coefficients for 5-channel reproduction (amplitude vs. no. of samples).

5×5 inverse filters and produce 5-channel reproduction signals (v_C^{5ch}, v_L^{5ch}, v_R^{5ch} v_{LS}^{5ch}, v_{RS}^{5ch}). Figures 8 and 9 show the 5×5 inverse filters and their frequency characteristics, respectively. The length of the inverse filter's response for 5-channel reproduction is 1024 coefficients. The h_{RS} indicates the inverse filter from S loudspeaker to R target point on the rigid sphere. The w_{RS} indicates a reproduced signal from S loudspeakers to R target point on the rigid sphere and can be generated through convolution between plant impulse response **c(n)** and h_{RS}. So the sum of R rows in Fig. 9 indicates the reproduced signals at R point on the rigid sphere.

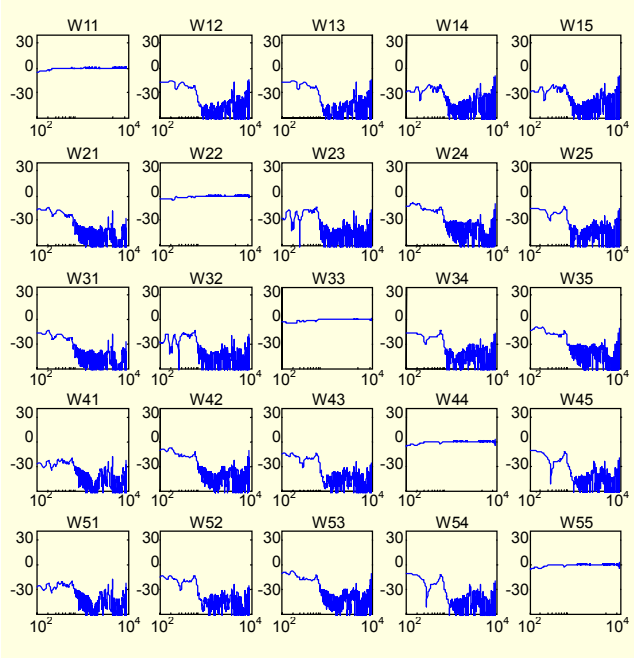


Fig. 9. Frequency characteristics of 5-channel inverse filters (the magnitude in dB vs. frequency in kHz).

B. Inverse Filtering for 4-channel Reproduction

The 4-channel reproduction configuration follows the ITU 5.1 standard except for the center and subwoofer for LFE. For the 4-channel reproduction, we convolve $u_{2\text{-}5}$ microphone signals with 4×4 inverse filters and produce 4-channel reproduction signals (v_L^{4ch}, v_R^{4ch} v_{LS}^{4ch}, v_{RS}^{4ch}). Figures 10 and 11 show 4×4 inverse filters and their frequency characteristics, respectively. The length of the inverse filter's response for 4-channel reproduction is 1024 coefficients.

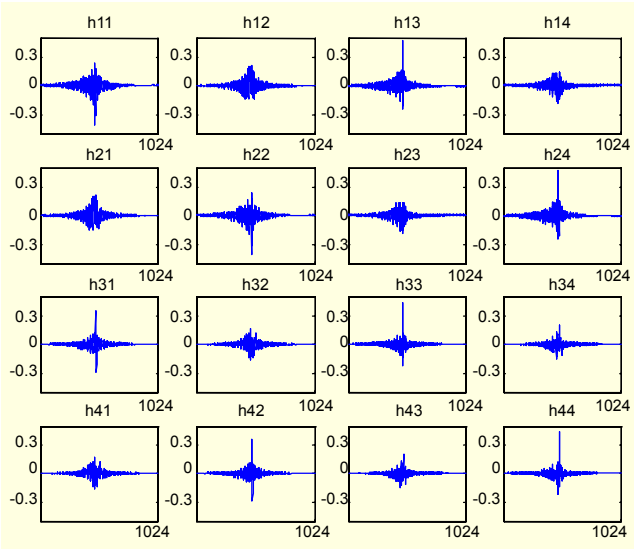


Fig. 10. Inverse filter coefficients for 4-channel reproduction (amplitude vs. no. of samples).

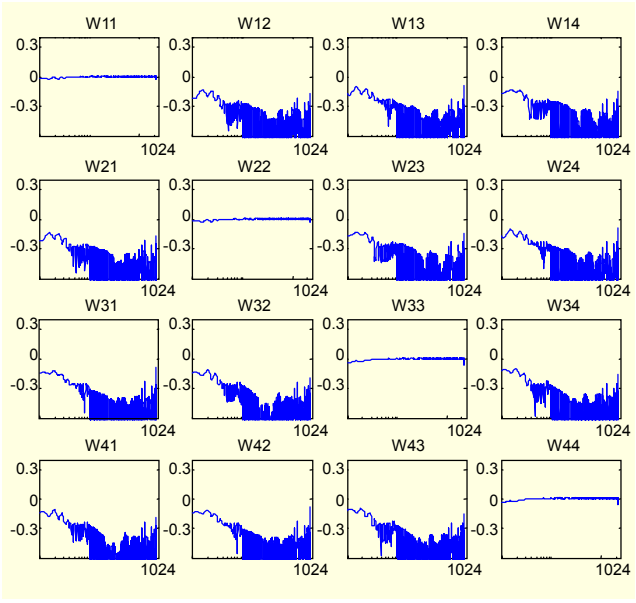


Fig. 11. Frequency characteristics of 4-channel inverse filters (the magnitude in dB vs. frequency in kHz).

C. Inverse Filtering for Stereo and Stereo Dipole Reproduction

For the stereo/stereo dipole reproduction environment, we convolve 2-channel headphone signals with 2×2 inverse filters and produce stereo or stereo dipole reproduction signals. Figure 12 shows the 2×2 inverse filters and their frequency characteristics of stereo dipole reproduction. Figure 13 shows 2×2 inverse filters and their frequency characteristics for stereo reproduction. The stereo/stereo dipole inverse filters are

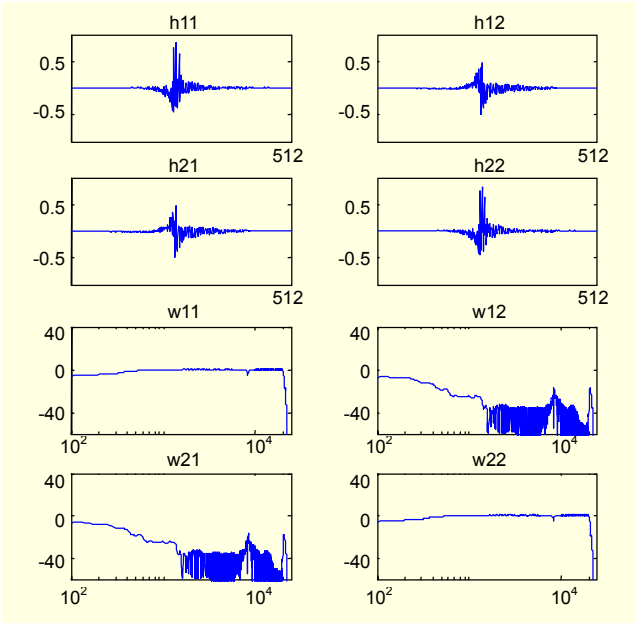


Fig. 12. Inverse filters for stereo dipole reproduction (amplitude vs. no. of samples) and their frequency characteristics (the magnitude in dB vs. frequency in kHz).

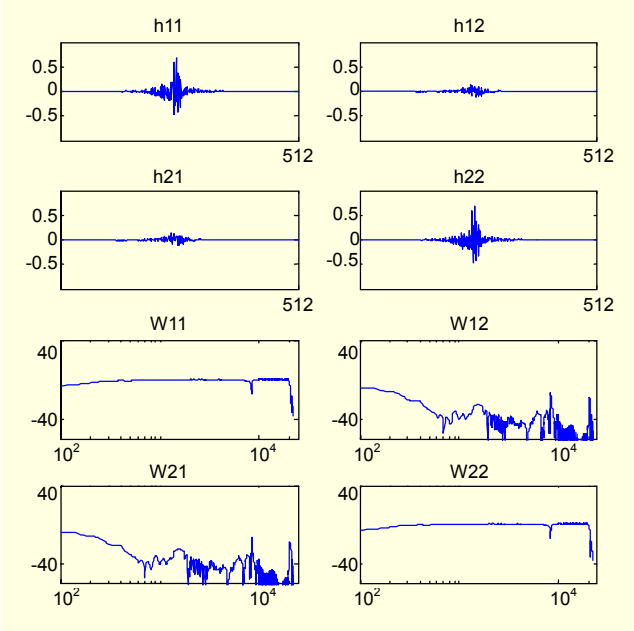


Fig. 13. Inverse filters for stereo reproduction (amplitude vs. no. of samples) and their frequency characteristics (the magnitude in dB vs. frequency in kHz).

generated using impulse responses between the stereo/stereo dipole loudspeaker position and LT (-90 degree) and RT (90 degree) point on a rigid sphere. The length of the inverse filter's response for stereo/stereo dipole reproduction is 512 coefficients.

The 5×5, 4×4, and 2×2 inverse filters are generated using impulse responses between loudspeakers and microphones on a rigid sphere. The frequency characteristic figure of these inverse filters shows that only diagonal elements (direct sources) are generated (flat frequency characteristic) and other elements (crosstalk sources) are eliminated.

3. Generation of Headphone Reproduction Signals

We have two methods for the generation of headphone reproduction signals. One uses 5-channel recorded signals and the other uses 5-channel reproduction signals, both mentioned before.

A. Generation of Headphone Reproduction Signals Using 5-channel Recorded Signals: $u_{1\sim5}$

This method is based on the conversion of 5-channel recorded signals, $u_{1\sim5}$. We converted the output signals of each microphone using conversion filters, which are generated using a sphere's impulse responses. When we measured impulse responses of the rigid sphere, we positioned a loudspeaker on a fixed point and rotated the rigid sphere's microphone direction on a horizontal plane as shown in Fig. 14. In this figure, SIR_n indicates the sphere's impulse response of n degrees between the loudspeaker and microphone on the rigid sphere. So, SIR_0 indicates the direct impulse response between the loudspeaker and microphone on the rigid sphere. The conversion filters are generated using SIR_0 and impulse responses of the rigid sphere.

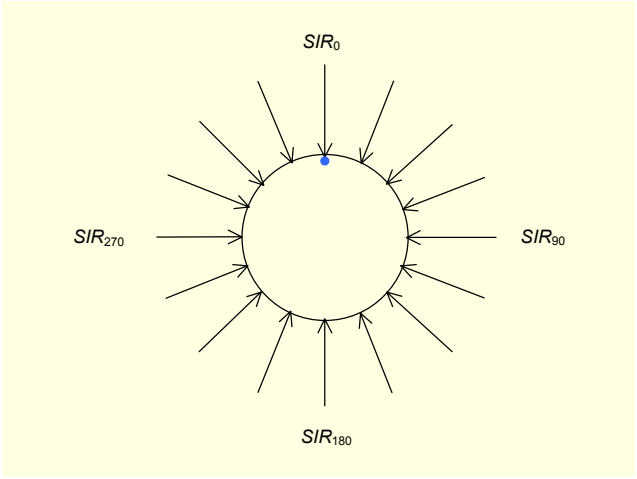


Fig. 14. Measurement of SIR (sphere impulse response).

Equation (2) shows the generation method of sphere conversion filter $SCF_{0\text{-}355}$, which is used for the conversion of 5-channel recorded signals to headphone reproduction signals, using the convolution between a sphere's impulse response and the inverse of SIR_0.

$$SCF_{0\sim355} = \mathrm{conv}(SIR_{0\sim355}, SIR_0^{-1}) \tag{2}$$

For the generation of 2-channel signals for headphone reproduction using 5-channel recorded signals, we convert each microphone's output signal to LT (-90 degree) and RT (90 degree) points using *SCFs* as shown in Fig. 15.

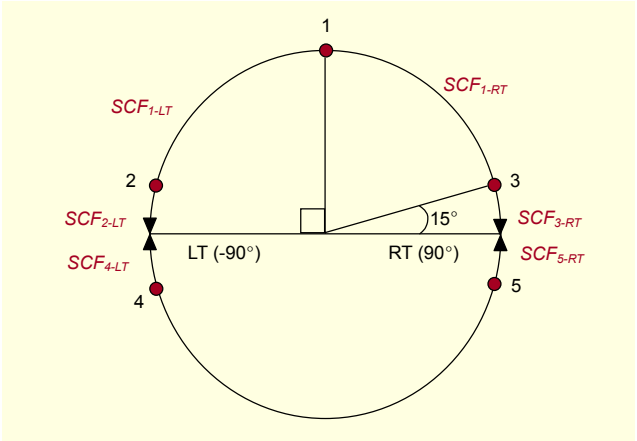


Fig. 15. The generation of headphone reproduction signals using 5-channel recorded signals ($u_{1\text{-}5}$).

Equation (3) shows the generation of headphone reproduction signals using u_{1-5} and *SCFs*. A headphone's left reproduction signal ($v_L^{HP\_u}$) is generated using the convolution between u_1, u_2, u_4 and $SCF_{1\text{-}LT}$, $SCF_{2\text{-}LT}$, $SCF_{4\text{-}LT}$. The headphone's right reproduction signal ($v_R^{HP\_u}$) is generated using the convolution between u_1, u_3, u_5 and $SCF_{1\text{-}RT}$, $SCF_{3\text{-}RT}$, $SCF_{5\text{-}RT}$.

$$\begin{aligned} v_L^{HP\_u} &= \mathrm{conv}(u_1, SCF_{1-LT}) + \mathrm{conv}(u_2, SCF_{2-LT}) \\ &\quad + \mathrm{conv}(u_4, SCF_{4-LT}) \\ v_R^{HP\_u} &= \mathrm{conv}(u_1, SCF_{1-RT}) + \mathrm{conv}(u_3, SCF_{3-RT}) \\ &\quad + \mathrm{conv}(u_5, SCF_{5-RT}) \end{aligned} \tag{3}$$

If we just add the left microphone signals (microphones 2 and 4) and the right microphone signals (microphone 3 and 5) for the generation of headphone reproduction signals without using *SCF*s, we feel more inside-head localizations, especially from the left- and right-sided sources (-90 and 90 degree sources).

B. *Generation of Headphone Reproduction Signals Using Inverse Filtered 5-channel Signals:* $v_{C,L,R,LS,RS}$

In this method, we use multi-channel inverse filtered signals $v_{C,\,L,\,R,\,LS,\,RS}$ for the generation of a headphone reproduction signal. For the generation of headphone reproduction signals using $v_{C,\,L,\,R,\,LS,\,RS}$, we use each impulse response between five loudspeakers and a rigid sphere's LT (left 90 degree) or RT (right 90 degree) points in the center of the reproduction environment as shown in Fig. 16. Equation (4) shows the generation method for headphone reproduction signals. In this equation, $SIR_{A\text{-}B}$ indicates the sphere impulse response from point *A* (loudspeakers, *C* for center, *L* and *R* for left and right, *LS* and *RS* for left and right surround) to point *B* (*LT*, *RT* points on a rigid sphere), and *conv* stands for convolution.

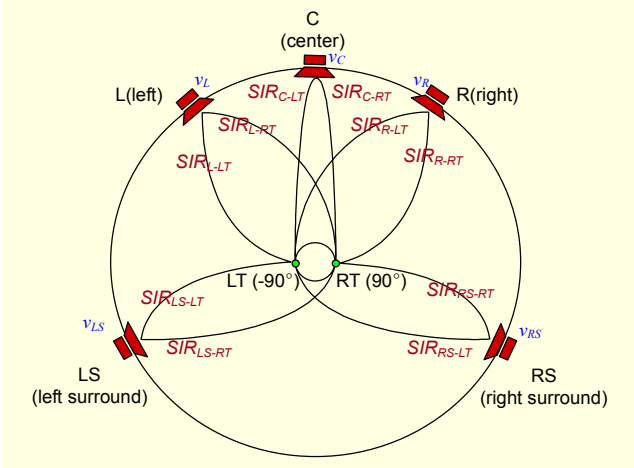


Fig. 16. Generation of headphone reproduction signals using inverse filtered 5-channel signals: $v_{C,\,L,\,R,\,LS,\,RS}$.

$$\begin{aligned} v_L^{HP\_v} &= \mathrm{conv}(v_C^{5ch}, SIR_{C-LT}) + \mathrm{conv}(v_L^{5ch}, SIR_{L-LT}) \\ &\quad + \mathrm{conv}(v_R^{5ch}, SIR_{R-LT}) + \mathrm{conv}(v_{LS}^{5ch}, SIR_{LS-LT}) \\ &\quad + \mathrm{conv}(v_{RS}^{5ch}, SIR_{RS-LT}) \\ v_R^{HP\_v} &= \mathrm{conv}(v_C^{5ch}, SIR_{C-RT}) + \mathrm{conv}(v_L^{5ch}, SIR_{L-RT}) \\ &\quad + \mathrm{conv}(v_R^{5ch}, SIR_{R-RT}) + \mathrm{conv}(v_{LS}^{5ch}, SIR_{LS-RT}) \\ &\quad + \mathrm{conv}(v_{RS}^{5ch}, SIR_{RS-RT}) \end{aligned} \tag{4}$$

Figure 17 shows the rigid sphere's impulse responses for the generation of headphone reproduction signals using 5-channel inverse filtered signals. As shown in the figure, because of the symmetrical characteristics of the rigid sphere, $SIR_{C\text{-}LT}$ and $SIR_{C\text{-}RT}$ are the same and $SIR_{R\text{-}LT}$, $SIR_{R\text{-}RT}$, $SIR_{RS\text{-}LT}$, $SIR_{RS\text{-}RT}$ are equal with $SIR_{L\text{-}LT}$, $SIR_{L\text{-}RT}$, $SIR_{LS\text{-}LT}$, $SIR_{LS\text{-}RT}$, respectively. The length of the rigid sphere's impulse responses is 128 points.

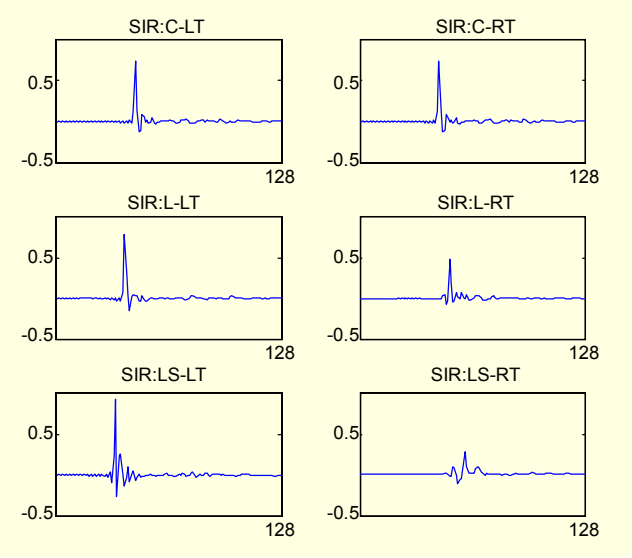


Fig. 17. SIR for the generation of headphone reproduction signals using $v_{C,\,L,\,R,\,LS,\,RS}$ (amplitude vs. no. of samples).

IV. Subjective Experiments

To verify the performance of our system, we carried out subjective experiments for a multi-channel loudspeaker and headphone reproduction environments. The experiment was carried out in the anechoic chamber of Tokyo Denki University with ten students who had normal hearing ability and were inexperienced in these kinds of localization experiments. The test contents were made by simulation using a mono source and a rigid sphere's impulse responses. Three test contents consisted of male and female voices and a classical music clip. Each of the three test contents was a virtual source from 0 degree (front) to 180 degrees (back) with 15 degree intervals, as shown in Fig. 18. The duration of content was 30 seconds.

After the reproduction of each content, a 10 seconds pause was given to the listeners. For a multi-channel localization experiment, the listener was seated in the center of five loudspeakers and a virtual source was played randomly only once per sound position. We hid the loudspeaker positions using a curtain and the listener could see the marks indicating the angle of the right hemisphere at a resolution of 15 degrees as shown in Fig.18. After hearing the virtual sources, the listener wrote down the perceived direction. During the multi-channel localization experiment, the listener was allowed to move his head slightly.

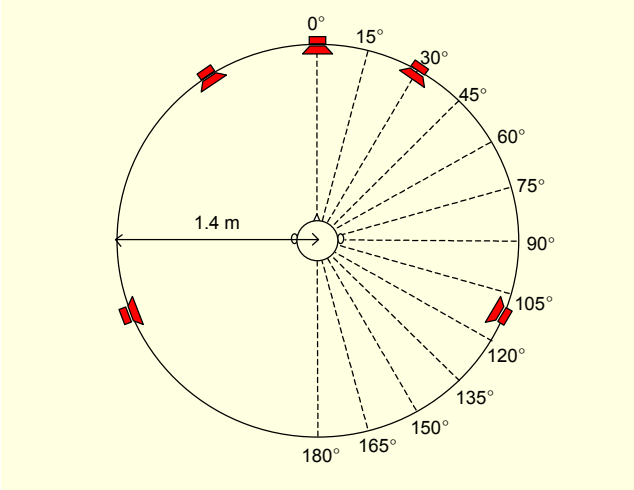


Fig. 18. Layout of localization experiment and direction of virtual sources.

1. Results of 5-Channel Localization Experiments

For the 5-channel localization experiment, we placed five loudspeakers in the ITU 5.1 loudspeaker layout except for the subwoofer for LFE. For 4-channel reproduction, we removed the center channel from the 5-channel reproduction layout. Figure 19 shows the results of the 5-channel localization experiment. We can find in the figure that there is very little front/back confusion.

2. Results of 4-Channel Localization Experiments

Figure 20 shows the results of the 4-channel localization experiment. As we can see in the figure, 4-channel reproduction also reduces the front/back confusion. But a 4-channel reproduction's front image is decreased slightly compared to 5-channel reproduction. In 5-channel reproduction, for the 0 degree reproduction (front), the center channel reproduces high energy and improves center images. But in 4-channel reproduction, the two front loudspeakers reproduce high energy and it makes the front image more ambiguous than in 5-channel reproduction.

Some listeners said that there were small changes of timbre between each source (coloration). The reason for this phenomenon is the inverse filter's frequency characteristics. As we can see in Figs. 9 and 11, there are slight performance variations of crosstalk cancellation over a full frequency range, and this causes the coloration.

3. Results of Headphone Localization Experiments.

We performed a headphone localization experiment using the headphone signals generated by the two methods in Sec. III.3 and KEMAR's HRTF. The test contents were male and female voices and music clips. Each experiment source was generated using a KEMAR and a rigid sphere's impulse responses. The experimental method was the same as the multi-channel localization experiment except that headphones were used instead of loudspeakers. The results are shown in Fig. 21. As we can see in the figure, headphone reproductions have more severe front/back confusion than multi-channel reproduction experiments. Headphone reproduction using 5-channel recorded signals has font/back confusion similar to the experiment with the KEMAR dummy head. But headphone reproduction using 5-channel inverse filtered signals reduced the front/back confusion compared to the experiment with the KEMAR dummy head. The main reason for the increase of front/back confusion is that in headphone reproduction a listener can't use head movements. Some listeners said that in headphone reproduction experiments using 5-channel inverse filtered signals, there were slight timbre changes of the front/back sources, and this helped the decision of direction.

Headphone reproduction using 5-channel recorded signals is very simple and has a low complexity. However, it has front/back confusion like in the dummy head reproduction. Headphone reproduction using 5-channel inverse filtered signals requires a complex calculation and implies coloration of sound images, but it can reduce the front/back confusions compared to the dummy head reproduction.

V. Conclusion and Future Work

In this paper, we suggested a new sound acquisition and reproduction method using multiple microphones on a rigid sphere and post processing for crosstalk cancellation. The purpose of post processing is the reproduction of original signals all around a single listener in an anechoic chamber. For this purpose, we generated various inverse filters, which can eliminate the crosstalk between loudspeakers and microphone points on a rigid sphere. The result of our multi-channel reproduction experiment shows that we can reduce the

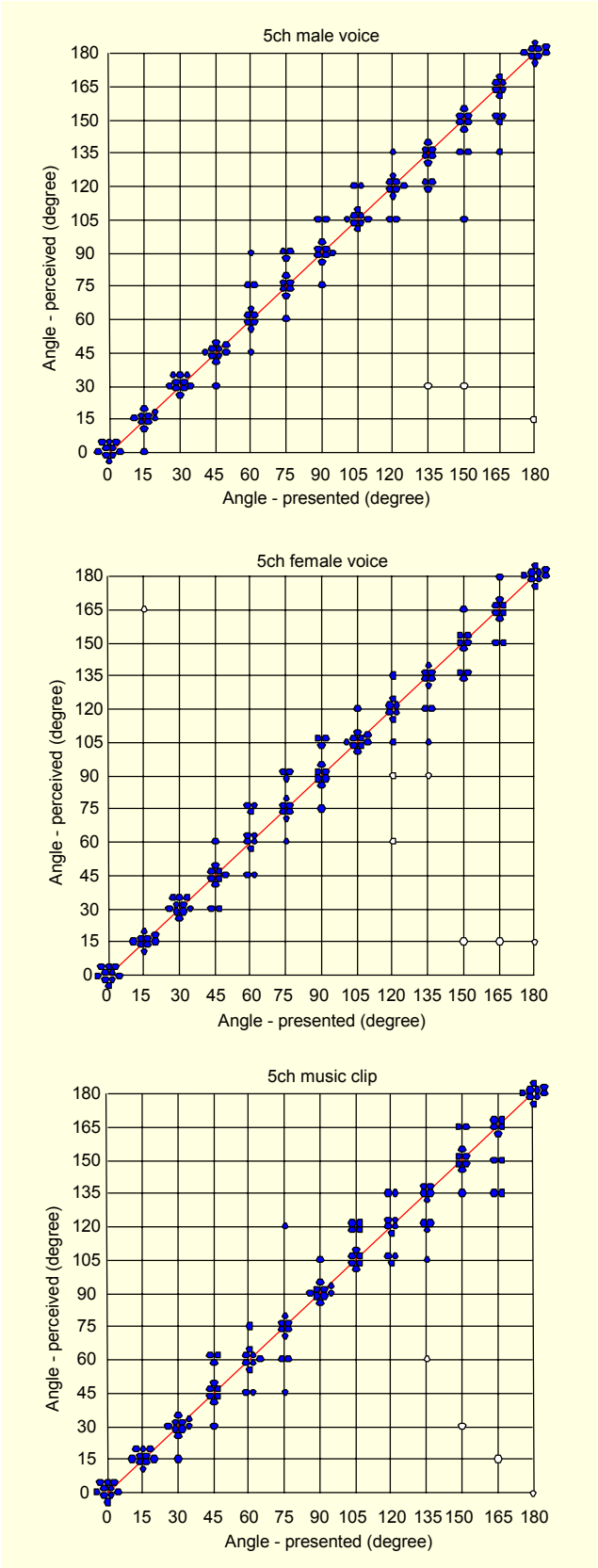


Fig. 19. Results of 5-channel localization experiments (perceived angle vs. presented angle).

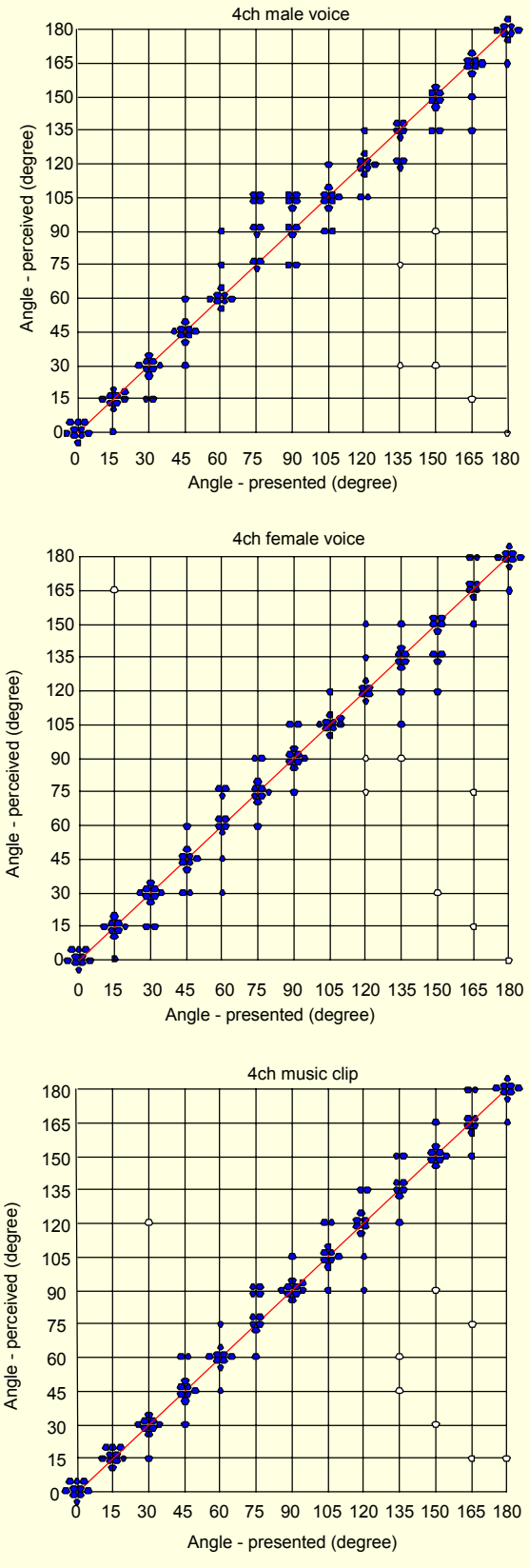


Fig. 20. Results of 4-channel localization experiments (perceived angle vs. presented angle).

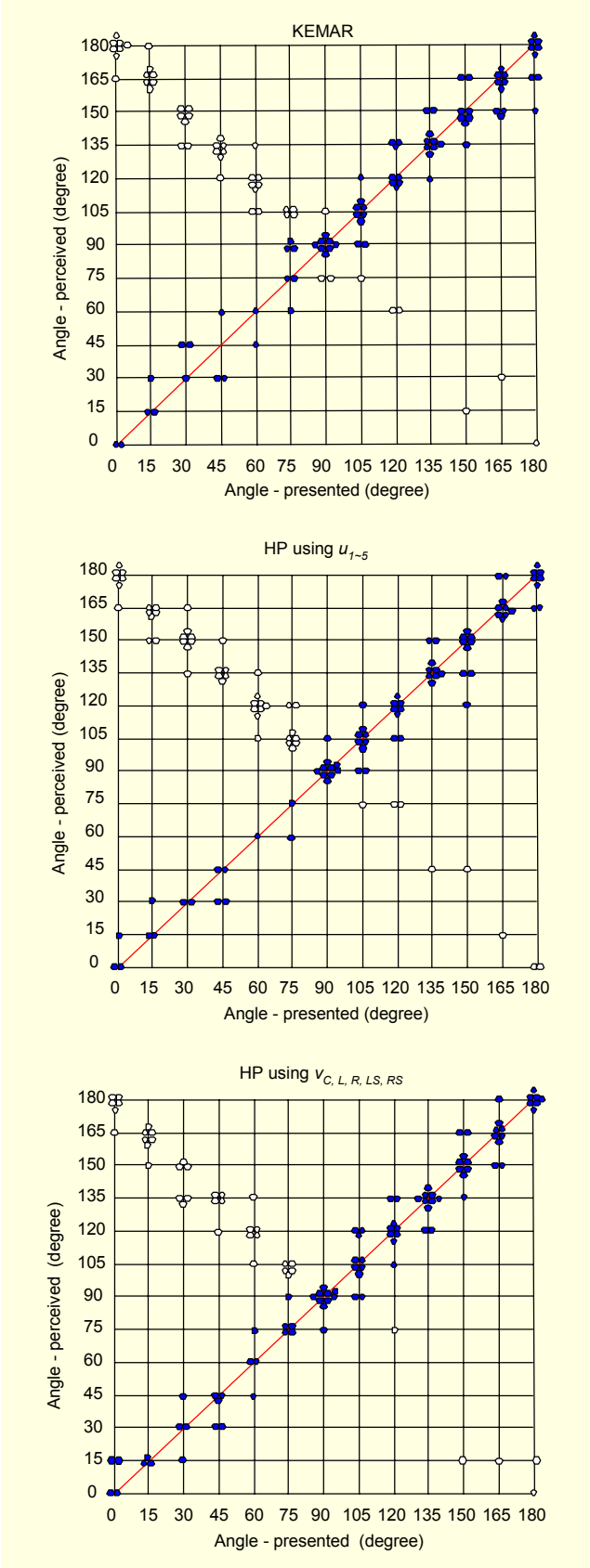


Fig. 21. Result of headphone localization experiments.

front/back confusion on the horizontal plane. The result of headphone reproduction experiment shows that we can slightly reduce the front/back confusions compared to dummy head reproduction.

For the post processing, we used 1024-tap and 512-tap filters, but to implement post processing into DSP hardware systems we need to reduce the filter length for the real-time processing. We also need to resolve the slight timbre changes between front and rear sound which is generated by a variation of the inverse filter's frequency characteristics. Because of inverse filtering using a rigid sphere's impulse responses at a center of reproduction environment, the sweet-spot is restricted to the center of the reproduction environment. For public use, we need to consider an extension of the sweet-spot.

References

[1] Francis Rumsey and Tim McCormick, *Sound and Recording: An Introduction*, Focal Press, 2002.

[2] H. Møller, "Fundamentals of Binaural Technology," *Applied Acoustics,* vol. 36, 1992, pp. 171-218.

[3] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization,* The MIT Press, 1983.

[4] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia,* AP Professional, 1994.

[5] F. Alton Everest, *Master Handbook of Acoustics,* McGraw-Hill, 2001.

[6] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Evaluation of Artificial Heads in Listening Tests," *AES 102nd Convention,* Mar. 1997, Preprint 4404.

[7] Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman, "Localization Using Nonindividualized Head-Related Transfer Functions," *J. Acoust. Soc. Am.,* vol. 94, no. 1, July 1993, pp. 111-123.

[8] Frederic L. Wightman and Doris J. Kistler, "Headphone Simulation of Free-Field Listening. II: Psychophysical Validation," *J. Acoust. Soc. Am.,* vol. 85, no. 2, Feb. 1989, pp. 868-878.

[9] P. A. Nelson, F. Orduna-Bustamante, and Hareo Hamada, "Experiments on a System for the Synthesis of Virtual Acoustic Sources," *J. Audio Eng. Soc.,* vol. 44, no. 11, Nov. 1996, pp. 990-1007.

[10] D. H. Cooper and J. Bauck, "Prospects for Transaural Recording," *J. Audio Eng. Soc.,* vol. 37, no. 1/2, Jan. 1989, pp. 3-19.

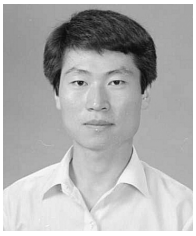
[11] William G. Gardner, *3-D Audio Using Loudspeakers,* Kluwer Academic Publisher, 1998.

[12] J. Bauck and D. H. Cooper, "Generalized Transaural Stereo," *AES 93rd Convention*, Dec. 1992, Preprint 3401.

[13] P. Damaske, "Head-related Two-Channel Stereophony with Loudspeaker Reproduction," *J. Acoust. Soc. Am.*, vol. 50, no. 4, Oct. 1971, pp. 1109-1115.

[14] Daryl Sartain, *3D Audio Programming,* Infinity Publishing.com, 2000.
[15] Robert M. Lambert, "Dynamic Theory of Sound-Source Localization," *J. Acoust. Soc. Am.,* vol. 56, no. 1, July 1974, pp. 165-171.
[16] Jean-Marie Pernaux, Marc Emerit, Jerome Daniel, and Rozenn Nicol, "Perceptual Evaluation of Static Binaural Sound Synthesis," *AES 22nd Int'l Conference*, 2002.
[17] M. Kleiner, "Problems in the Design and Use of Dummy-Heads," *Acoustica,* vol. 41, 1978, pp. 183-193.
[18] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural Technique: Do We Need Individual Recording?" *J. Audio Eng. Soc.,* vol. 44, no. 6, June 1996, pp. 451-469.
[19] Tomlinson Holman, *5.1 Surround Sound Up and Running*, Focal Press, 2000.
[20] http://www.schoeps.de/E/kfm6.html.
[21] http://www.theaudio.com/bs-3d.html.
[22] Yuvi Kahana, Philip A. Nelson, Ole Kirkeby, and Hareo Hamada, "A Multiple Microphone Recording Technique for the Generation of Virtual Acoustic Images," *J. Acoustic. Soc. Am.*, vol. 105, no. 3, Mar. 1999, pp. 1503-1516.
[23] Ole Kirkeby, Philip A. Nelson, Hareo Hamada, and Felipe Orduna-Bustamante, "Fast Deconvolution of Multichannel Systems Using Regularization," *IEEE Trans. Speech and Audio Processing,* vol. 6, no. 2, Mar. 1998, pp. 189-194.
[24] Hareo Hamada, Hironori Tokuno, Yuko Watanabe, and P. A. Nelson, "3D Sound Generation Using Two Loudspeakers – Stereo Dipole System and Its Applications," *AES 15th Int'l Conference*, 1998.
[25] Ole Kirkeby, P. A. Nelson, and Hareo Hamada, "The Stereo Dipole – A Virtual Source Imaging System Using Two Closely Spaced Loudspeakers," *J. Audio Eng. Soc.,* vol. 46, no. 5, May 1998, pp. 387-395.
[26] ITU-RT BS.775-1, "Multichannel Stereophonic Sound System With and Without Accompanying Picture," *Rec., Int'l Telecommunications Union,* 1992~1994.
[27] Bill Gardner and Keith Martin, "HRTF Measurement of a KEMAR Dummy-Head Microphone," *MIT Media Lab Technical Report,* no. 280, May 1994.
[28] http://www.noe.co.jp, "AEIRM Impulse Response Measurement System."
[29] Zhan Huan Zhou, "Sound Localization and Virtual Auditory Space," *Project report of University of Toronto,* 2002.
[30] Mikio Tohyama and Tsunehiko Koike, *Fundamentals of Acoustic Signal Processing*, Academic Press, 1998.
[31] Bernard Widrow and Samuel D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, 1985.
[32] M. R. Schroeder, "Models of Hearing," *Proc. IEEE,* vol. 63, no. 9, Sep. 1975, pp. 1332-1335.
[33] M. Miyoshi and Y. Kaneda, "Inverse Filtering of Room Acoustics," *IEEE Trans. Acoust. Speech and Signal Processing,* vol. 36, no. 2, Feb. 1998, pp. 145-155.
[34] O. Kirkeby and P. A. Nelson, "Reproduction of Plane Wave Sound Fields," *J. Acoust. Soc. Am.,* vol.94, no. 5, Nov. 1993, pp. 2992-3000.
[35] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C,* Second Edition, Cambridge University Press, 1992.
[36] J. H. Wilkinson, *The Algebraic Eigenvalue Problem,* Oxford University Press, 1965.
[37] Hironori Tokuno, Ole Kirkeby, Philip A. Nelson, and Hareo Hamada, "Inverse Filter of Sound Reproduction Systems Using Regularization," *IEICE Trans. Fundamentals*, vol. E80-A, no. 5, May 1997, pp. 809-820.

Taejin Lee received the BS and MS degrees in electronics engineering from Chonbuk National University, Jeonju, Korea, in 1996 and 1998. He worked for Mobens Co., Ltd. Korea, from 1998 to 2000. Since 2000, he has been with the Electronics and Telecommunications Research Institute (ETRI), Korea, as a Senior Member of Research Staff. From 2002 to 2003, he was a Visiting Researcher at Tokyo Denki University, Japan. His research interests include audio signal processing and interactive broadcasting technologies.



Daeyoung Jang received the BS degree in electronic engineering from Pukyong National University, Busan, Korea, in 1991 and the MS degree in computer science from Paichai University, Daejeon, Korea in 2000. He has been engaged at Electronics and Telecommunications Research Institute, Korea, since 1991 and he has been a Senior Research Engineer since 1995. He has researched electro-acoustics for telecommunications and broadcastings, and developed MPEG-1, 2, 4 audio systems for broadcasting and telecommunication systems. Now he is developing an interactive surround audio system for 3D audio visual broadcasting and telecommunications.

Kyeongok Kang received his BS and MS degrees both in physics from Pusan National University, Busan, Korea, in 1985 and 1988, and his PhD degree in electrical engineering from Hankuk Aviation University, Seoul, Korea, in 2004. He has been in Electronics and Telecommunications Research Institute (ETRI) since 1991, and he is now a Principal Member of Engineering Staff and the Leader of 3D Media Research Team. His research interests are in personalized broadcast technologies based on MPEG-7 and TV-Anytime, and audio signal processing including 3D audio.

Jinwoong Kim received his BS and MS degrees both from Seoul National University, Busan, Korea, in 1981 and 1983, and his PhD degree from the Department of Electrical Engineering from Texas A&M University, United States, in 1993. Since 1983, he has been on the Research Staff in Electronics and Telecommunications Research Institute (ETRI), Korea. He has been engaged in the development of the TDX digital switching system, MPEG-2 video encoder, HDTV encoder system, and MPEG-7 technology. His research interests include digital signal processing in the fields of video communications, multimedia systems, and interactive broadcast systems.

Dae-Gwon Jeong received the BS in electrical engineering from Hankuk Aviation University in 1979, and the MS and PhD in electrical engineering from Texas A&M University in 1987 and 1990. He has worked for the ADD and ETRI, Korea, from March 1979 to May 1984 and from October 1990 to August 1991, respectively. Since 1991, he has been with the Department of Avionics Engineering at Hankuk Aviation University, Korea. His research interest includes multimedia compression for broadcasting and mobile multimedia communication systems.

Hareo Hamada has over 25 years of experience in fields of psychoacoustics, digital signal processing and sound research. Since 1983, he has held several teaching positions at Tokyo Denki University where he currently holds a faculty position as a Tenured Professor in the School of Information Environment of Tokyo Denki University. He holds a BS in electrical communication engineering and an MS and PhD in electrical engineering, all from Tokyo Denki University. He held a Visiting Professor position at Institute of Sound and Vibration Research, the University of Southampton, UK, from 1988 to 1990. He has been a chairman of Advanced Acoustic Research Meeting of IEICE (The Institute of Electronics, Information and Communication Engineers). He has also organized many international meetings, especially in the field of audio research and in the active control of noise and vibrations. He founded a high-tech venture with university links, DiMAGIC Co., Ltd., in June 1999.