Paper

3D Audio System Using Multiple Vertical Panning for Large-screen Multiview 3D Video Display

Toshiyuki Kimura †, Hiroshi Ando †

Abstract In this paper, a 3D audio system using multiple vertical panning (MVP), which matches audio to a large-screen multiview 3D video display system, is proposed. The vertical position of sound images is synthesized by the panning between two loudspeakers placed at the upper and lower sides of the screen. The horizontal position of the sound images is controlled by the position of two loudspeakers. Using the proposed system, multiple viewers can simultaneously feel the sound images at the position of 3D objects. A listening test was used to examine whether viewers can perceive the synthesized sound images at the position between two loudspeakers. The results of an audio-visual experiment show that the proposed 3D audio system was effective as compared with a conventional system because viewers could always feel the synthesized sound images at the position of the 3D video object.

Key words: ultra-realistic communication, 3D audio system, vertical panning, multiview 3D video display system

1. Introduction

Ultra-realistic communication techniques have been investigated in the NICT[1]. If these techniques are applied, this will enable more realistic forms of communication (e.g., 3DTV phone and 3D teleconferencing) than those currently offered by conventional video and audio techniques (HD video and 5.1-channel audio).

At the NICT, a glasses-free 3D video technique using a projector array has been proposed and a multiview 3D video display system, in which the size of a screen is 200 inches[2], has been developed. The basic configuration of the developed system is shown in **Fig.1**. Parallax videos are projected to a Fresnel lens by projector units, which are components of the projector array. These parallax videos are only projected in the horizontal direction because of the diffusion characteristics of a diffuser screen placed in front of the Fresnel lens (small diffusion angle in the horizontal direction and wide diffusion angle in the vertical direction). As a result, because this system allows viewers to view parallax videos according to the horizontal position, several viewers can view natural 3D objects simultaneously according to each particular viewing position, without the need for special glasses. However, the developed system presents only visual sensations. In order to achieve a realistic auditory sensation, a 3D audio system must be developed that corresponds with a large-screen multiview 3D video display system .

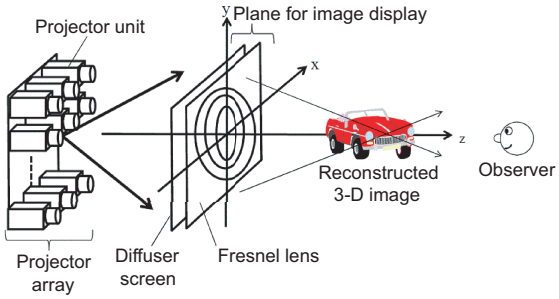

Fig. 1 Basic configuration of a large-screen multiview 3D video display system[3].

According to the principle of the developed large-screen multiview 3D video display system, the 3D audio system needs to satisfy following technical requirements:

(i) Because multiple people view the 3D objects depicted by the 3D video display system, multiple viewers can simultaneously feel the sound images at the position of the 3D objects in any viewing position.

(ii) Because multiple people view the 3D objects without needing special glasses, viewers can listen to the sound without wearing hearing devices.

(iii) Sound playing devices are not placed between the projector array and the viewing position because sound devices placed between the projector array and the viewing position prevent the projec-

Received May 20, 2013; Revised September 6, 2013; Accepted October 29, 2013

† Universal Communication Research Institute, National Institute of Information and Communications Technology (Kyoto, Japan)

tion of the 3D video display system.

(iv) If a 3D audio system is applied as an interactive communication system such as 3D teleconferencing, microphones for recording are not placed between the projector array and the viewing position because microphones placed between the projector array and the viewing position prevent the projection of the 3D video display system.

From requirement (i), it is difficult to apply stereophonic[4] and 5.1 channel[5] systems because the viewer must listen to a sound at one particular point in these systems. From requirement (ii), it is difficult to apply a binaural[6] system because a viewer wears headphones in this system. From requirement (iii), it is difficult to apply 22.2 channel audio[7], higher order ambisonics[8], and wave field synthesis[9] systems because loudspeakers are placed between the projector array and the viewing position in these systems. From requirement (iv), it is difficult to apply transaural[10] and boundary surface control[11] systems because microphones for recording are placed between the projector array and the viewing position in these systems. Thus, conventional 3D audio systems do not satisfy all of the technical requirements described above. In order to solve this problem, a novel 3D audio system must be developed from a different viewpoint.

In this paper, based on a different viewpoint from conventional systems, a novel 3D audio system using the multiple vertical panning (MVP) method is proposed in order to match the developed 3D video display system. In Section 2, the principle of the proposed system is described. Using this system, it is indicated that multiple viewers can simultaneously feel the sound images at the position of 3D objects depicted by the 3D video display system, without wearing hearing devices. Section 3 describes the listening test used to evaluate the auditory performance of the proposed system. Using this test, it is shown that viewers can feel the synthesized sound images at a position between two loudspeakers placed at the top and bottom of the vertically aligned loudspeaker array. Section 4 describes the audio-visual experiment used to evaluate the audio-visual performance of the proposed system. The results of this experiment indicate that the proposed system is effective as compared with conventional audio systems such as stereophonic audio.

2. Diagram of proposed system

The basic configuration of the proposed system is shown in **Fig.2**. First, as shown in the left-hand side of Fig.2, two loudspeakers are placed at the upper and lower sides of the position of the 3D object depicted in the screen by the developed 3D video display system because the basis of the depth of the 3D objects is the position of the Fresnel lens in Fig.1. If a sound is played from two loudspeakers using the panning between two loudspeakers (called "vertical panning"), it is expected that viewers can perceive a sound image between two loudspeakers. If the sound pressure level difference between two loudspeakers is properly adjusted, because sound playing devices are only two loudspeakers placed at the upper and lower sides of the screen (called "vertically panned loudspeakers"), it is expected that multiple viewers can perceive a sound image at the position of the 3D object, regardless of the viewing position. Note that the viewing position in the proposed system means only the horizontal position because the viewing position in the 3D video display system is the horiontal position.

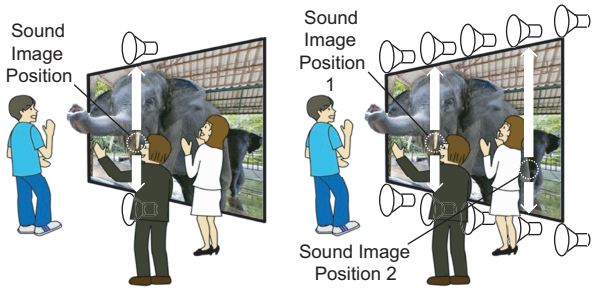

Fig. 2 Basic configuration of the proposed 3D audio system.

Second, as shown in the right-hand side of Fig.2, sound image positions are also expanded in the left-right direction by placing multiple vertically panned loudspeakers at the upper and lower sides of the screen. As a result, multiple viewers can simultaneously feel the sound images at the position of 3D objects depicted by the 3D video display system, regardless of the viewing position. In the proposed system, viewers do not need to wear hearing devices. Because loudspeakers are placed at the upper and lower sides of the screen, there are no sound devices between the projector array and the viewing position in the proposed system. Because this system only has to directly record the speech of participants in the teleconference and does not restrict the position of the recording microphones, it does not need to place recording microphones between the projector array and the viewing position in the proposed system. Thus, the proposed system satisfies all the technical re-

Fig. 3 Image of experimental environment in the listening test (Left: Reverberation time 140 ms, Right: Reverberation time 1030 ms).

quirements described in Section 1.

3. Listening test

The minimum required component of the proposed system consists of summing localization between vertically panned loudspeakers. Because the auditory performance of the proposed system can be denoted by the superposition of minimum components, it is sufficient to evaluate the perceived heights of sound images synthesized by vertically panned loudspeakers.

Because the experimental conditions of the vertical panning performed in the past study[13] do not match the conditions of the developed large-screen multiview 3D video display system, the proper sound pressure level difference for vertically panned loudspeakers described in Section 2 is unknown. In this section, a listening test evaluating the perceived height of the synthesized sound images is performed by two loudspeakers assuming the placement at the upper and lower sides of the screen of the developed large-screen multiview 3D video display system in order to set a vertical panning curve for the proper sound pressure level difference between two loudspeakers.

3.1 Environment and conditions

The listening test was performed in the ATR variable reverberation room[12]. The reverberation time can be changed from 140 ms (total absorption) to 1030 ms (total reflection) in this room as shown in **Fig.3** by rotating the cylinders and ceiling louvers that are components of the walls. A background noise level had an A-weighted level of 14 dB when the reverberation time was 140 ms and an A-weighted level of 22 dB when the reverberation time was 1030 ms.

Twenty-seven loudspeakers were placed in a vertical line, as shown in **Fig.4**. Loudspeakers were manufactured by mounting a loudspeaker unit (Fostex: FE103En) on a loudspeaker enclosure (width: 11 cm, depth: 25 cm, height: 11 cm). The total height of the loudspeaker array was 2.97 m (=11 cm×27). The viewing position was set at a distance of 5.5 m from the loudspeaker array according to the appropriate viewing distance in the developed large-screen multiview 3D video display system[2]. The height of the viewing position was set to 1.485 m, at the ear position of the viewers. The sound pressure level in the viewing position was set to an A-weighted level of approximately 70 dB.

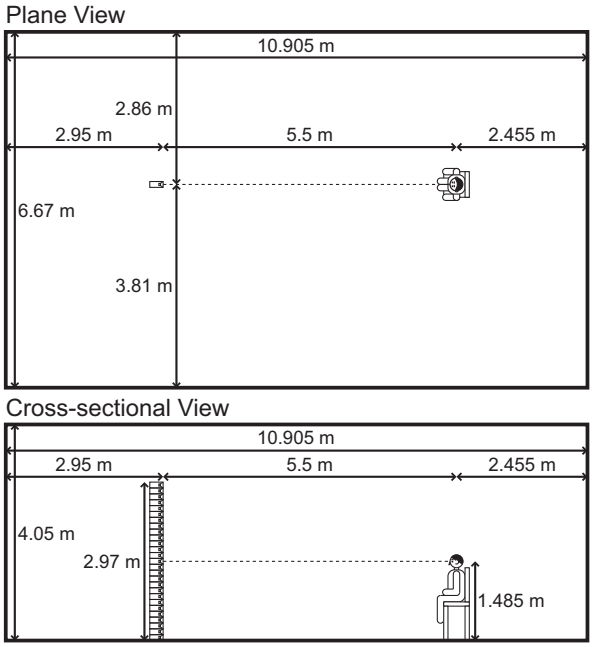

Fig. 4 Position of the viewer and the loudspeaker array in the listening test.

The experimental conditions in the listening test are shown in **Fig.5**. The gray loudspeakers denote the loudspeaker from which a sound is not replayed in each condition. In the panning condition (a), the sound calculated from the sound source signal, $s(n)$, was replayed from two loudspeakers placed at the upper and lower sides in the loudspeaker array according to the following equations:

$$x_U(n) = a_U s(n), \tag{1}$$

$$x_D(n) = a_D s(n), \tag{2}$$

where $x_U(n)$ and $x_D(n)$ denote the sound signals replayed from two loudspeakers of the upper and lower sides, respectively, and a_U and a_D ($a_U^2 + a_D^2 = 1$) denote the gain coefficients in each sound signal. If the level difference, ΔA [dB], is defined as follows:

$$\Delta A = 20\log_{10}\frac{x_U(n)}{x_D(n)} = 20\log_{10}\frac{a_U}{a_D}, \tag{3}$$

a_U and a_D are calculated as follows:

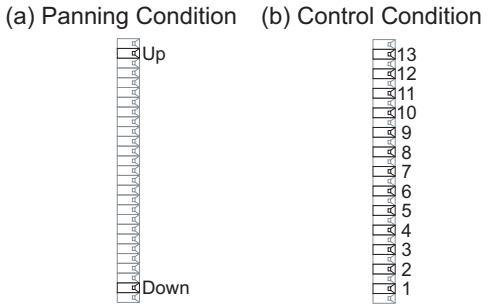

Fig. 5 Experimental conditions used in the listening test (Left: Panning condition, Right: Control condition).

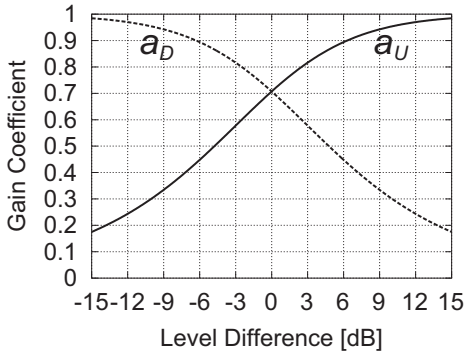

Fig. 6 Gain coefficients of two loudspeakers of the upper and lower sides in the listening test.

$$a_U = \frac{10^{\frac{\Delta A}{20}}}{\sqrt{10^{\frac{\Delta A}{10}}+1}}, \tag{4}$$

$$a_D = \frac{1}{\sqrt{10^{\frac{\Delta A}{10}}+1}}. \tag{5}$$

In this test, the level difference, ΔA, was set from -15 dB to 15 dB at the interval of 1 dB. The curves of a_U and a_D are shown in **Fig.6**. If $\tan\alpha = 10^{\frac{\Delta A}{20}}$ is applied to Eqs. (4) and (5), the curves of a_U and a_D are transformed to the curves of sines and cosines (i.e., $\sin\alpha$ and $\cos\alpha$). In the control condition (b), the sound source signal, $s(n)$, was replayed from one loudspeaker selected from a group of thirteen loudspeakers.

3. 2 Design and procedure

Twelve subjects (age: 21–32, six males and six females), whom the audibility was normal in daily life, participated as viewers in this test. Three types of sounds (white noise, speech, and flute) were used as a sound source. The flowchart of the listening test is shown in **Fig.7**. The test was divided into six sessions for reverberation times and sound sources. The presented orders of sessions were counterbalanced in all viewers. Twelve practice trials and eighty-eight main trials were performed in each session. During the main trials, rest periods were allowed after every set of forty-four trials. The presentation orders of trials were randomized for each viewer. The details of the practice and main trials are shown in **Table 1**.

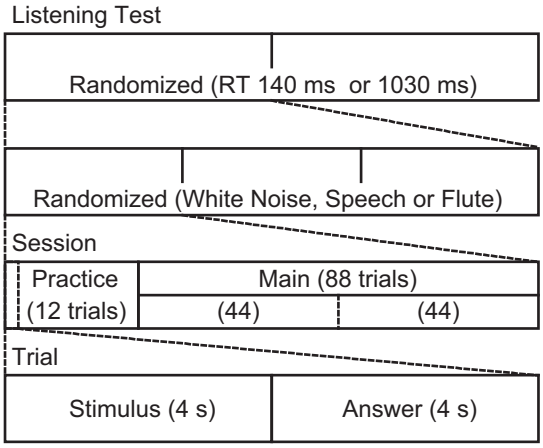

Fig. 7 Flowchart of the listening test.

Table 1 Practice and main trials in the listening test.

	Element	Note
Practice (12)	= 7 conditions	ΔA=0, ±5, ±10 and ±15 in (a) of Fig.5
	+ 5 positions	1, 4, 7, 10 and 13 in (b) of Fig.5
Main (88)	= [31 conditions + 13 positions] × 2 repetitions	ΔA=-15∼15 in (a) of Fig.5 1∼13 in (b) of Fig.5

The viewers were instructed to report the perceived height of the sound images by listing the indexes of the heights in an answer sheet. The relation between the perceived heights and the answer indexes is shown in **Fig.8**. This index ranges from 1 to 27, and the height of the loudspeaker, of which the index is 14, is the same as that of the ears and eyes of viewers. If two loudspeakers are placed near the ceiling and floor of a room for vertical panning, because the sound quality of loudspeakers widely varies owing to the reflected sounds from the ceiling and floor of a room, viewers may feel the two sound images at the position of the two loudspeakers in the panning condition. Thus, in order to verify the phenomenon described above, the viewers could list multiple indexes in an answer sheet if viewers felt multiple sound images in the trials. When the perceived height of viewers was not fitted on the loudspeaker array, the viewers listed the indexes of edges in the loudspeaker array (i.e., 1 and 27). The viewers were allowed to move their heads and upper bodies freely while listening to the sounds.

3. 3 Results and discussion

(1) Response rate of synthesized sound image

Because viewers also entered the number of heard sound images in this test, viewers perceived one synthesized sound image between two loudspeakers if the

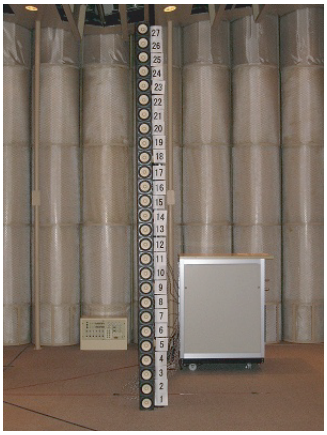

Fig. 8 Relation between perceived heights of sound images and answer indexes in the listening test.

entered number of indexes in the trials was one. The response rates were calculated based on the number of trials in which viewers listed one index in order to evaluate whether viewers could feel a synthesized sound image between two loudspeakers in the panning condition.

The results of the response rates in both conditions are shown in **Figs.9** and **10**. Error bars denote the 95% confidence interval of the response rates. In the panning condition, response rates are always greater than or equal to 0.875. On the other hand, response rates in the control condition are also greater than or equal to 0.875. Thus, it is indicated that viewers can feel synthesized sound images between two loudspeakers in the panning condition because response rates in the panning condition are almost the same as those in the control condition.

(2) Perceived height of synthesized sound image

The perceived heights of the sound images was calculated from the answer indexes of viewers according to the following equation:

$$H_{\mathrm{per}}[\mathrm{m}] = (I_{\mathrm{ans}} - 14)\times 0.11, \tag{6}$$

where I_{ans} and H_{per} denote the answered index of the loudspeakers and the perceived height of the sound image, respectively. Note that the answers where viewers listed multiple indexes eliminated before calculating the perceived heights.

The results of the averages of the perceived heights in the panning condition are shown in **Fig.11**. Error bars denote the 95% confidence interval of the average heights. Labels in the right side of Fig.11 denote the elevation angles calculated from the percieved heights (i.e., $\tan^{-1}\frac{H_{\mathrm{per}}}{5.5}$). In all conditions, the perceived height of the sound images is approximately 0 m (i.e., the middle point between two loudspeakers placed at the upper and lower sides of the screen) when the level difference is approximately 0 dB, and the perceived height of the sound images linearly changes relative to the level dif-

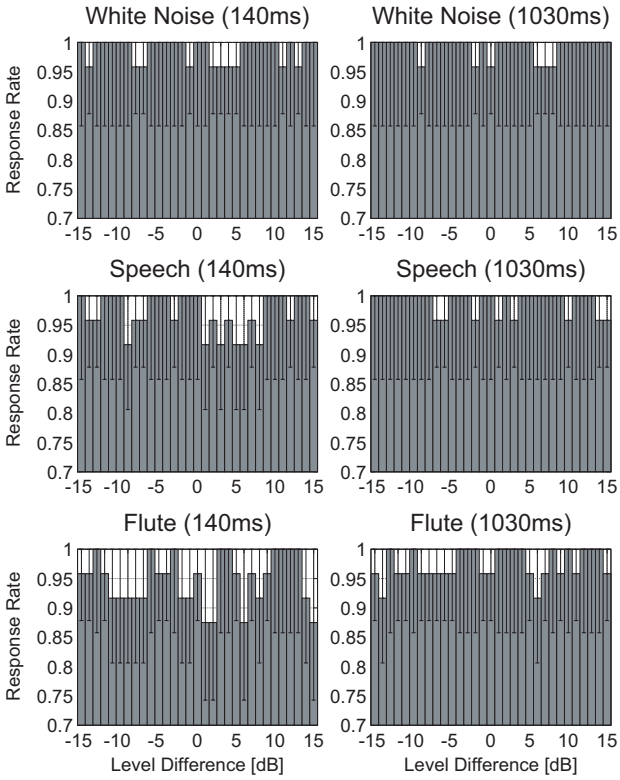

Fig. 9 Response rates of the panning condition in the listening test.

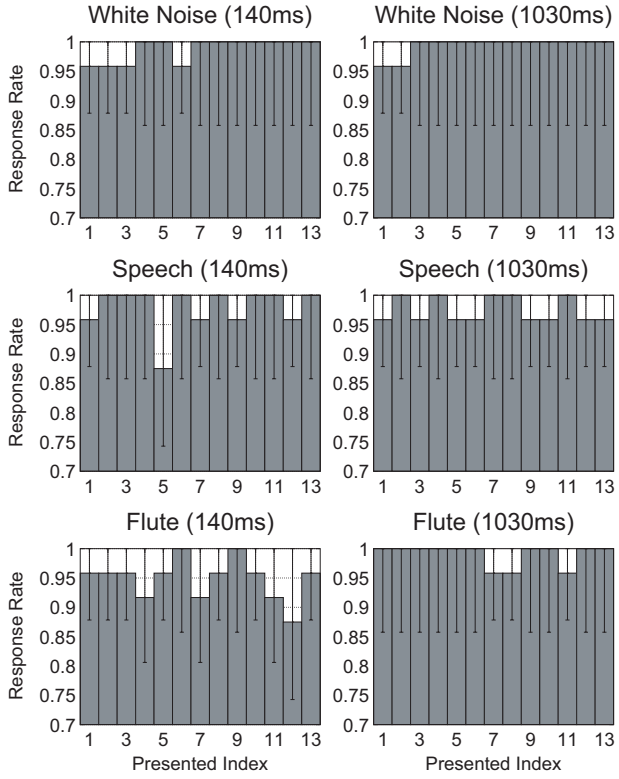

Fig. 10 Response rates of the control condition in the listening test.

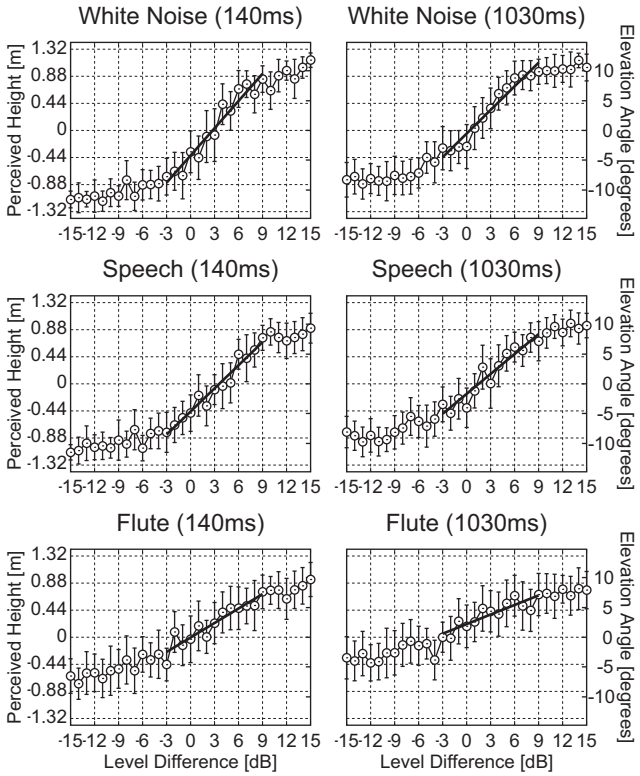

Fig. 11 Results and regression lines of the panning condition in the listening test.

ference when the level difference ranges from -3 dB to 9 dB. As the result of linear regressions in the range from -3 dB to 9 dB, the following regression lines were obtained in each condition:

(White noise, Reverberation time 140 ms)

$$H_{\mathrm{per}} = 0.1475\Delta A - 0.4066, \tag{7}$$

(Speech, Reverberation time 140 ms)

$$H_{\mathrm{per}} = 0.1253\Delta A - 0.4499, \tag{8}$$

(Flute, Reverberation time 140 ms)

$$H_{\mathrm{per}} = 0.0784\Delta A - 0.0045, \tag{9}$$

(White noise, Reverberation time 1030 ms)

$$H_{\mathrm{per}} = 0.1279\Delta A - 0.0510, \tag{10}$$

(Speech, Reverberation time 1030 ms)

$$H_{\mathrm{per}} = 0.1079\Delta A - 0.1635, \tag{11}$$

(Flute, Reverberation time 1030 ms)

$$H_{\mathrm{per}} = 0.0518\Delta A + 0.2130. \tag{12}$$

Regression lines are shown in Fig.11 as bold lines. It is shown that these lines are correctly estimated with the level difference in the range from -3 dB to 9 dB.

According to averaging the regression lines obtained in six conditions, panning curve H_{pan} was calculated as follows:

$$H_{\mathrm{pan}} = \begin{cases} -1.32 \quad (\Delta A < -11.05) \\ 0.1065\Delta A - 0.1437 \\ \qquad (-11.05 \leqq \Delta A \leqq 13.74) \\ 1.32 \quad (\Delta A > 13.74) \end{cases}. \tag{13}$$

The differential limens of the perceived heights of sound images (DL^{+}_{pan} and DL^{-}_{pan}) were also calculated according to the following equations:

$$DL^{+}_{\mathrm{pan}} = \tan(\tan^{-1}(H_{\mathrm{pan}}/5.5) + \phi)\times 5.5, \tag{14}$$

$$DL^{-}_{\mathrm{pan}} = \tan(\tan^{-1}(H_{\mathrm{pan}}/5.5) - \phi)\times 5.5, \tag{15}$$

where ϕ denotes the differential angle of the perceived height of a sound image. The value of ϕ was set to 9° with reference to past studies[14]. The panning curves and differential limens in the panning condition are shown in **Fig.12**. Gray areas denote the area outside of the differential limens. If the average of the perceived heights of the sound images is in the gray areas, viewers can discriminate the difference between the presented height of sound images according to the panning curve and the perceived height of sound sources presented by the 3D video. In five conditions except one (Flute, Reverberation time 1030 ms), because the average of the perceived heights of sound images was not in a gray area, the auditory performance of the panning curve is so high that viewers cannot discriminate the differences among the heights. However, in that condition (Flute, Reverberation time 1030 ms), because the average of the perceived heights of the sound images is in a gray area, viewers may be able to discriminate the differences among the heights.

On the other hand, the results and differential limens in the control condition are shown in **Fig.13**. Error bars denote the 95% confidence interval of the average heights. The differential limens in the control condition (DL^{+}_{ctrl} and DL^{-}_{ctrl}) were calculated according to following equations:

$$DL^{+}_{\mathrm{ctrl}} = \tan(\tan^{-1}(H_{\mathrm{pre}}/5.5) + \phi)\times 5.5, \tag{16}$$

$$DL^{-}_{\mathrm{ctrl}} = \tan(\tan^{-1}(H_{\mathrm{pre}}/5.5) - \phi)\times 5.5, \tag{17}$$

where $H_{\mathrm{pre}}(= (I_{\mathrm{pre}} - 14)\times 0.11)$ denotes the height of the presented sound source, and I_{pre} denotes the index of the presented loudspeakers. In five conditions except one (Flute, Reverberation time 1030 ms), the average of the perceived height of the sound images is not in a gray area. However, in one condition (Flute, Reverberation time 1030 ms), the average of the perceived heights of the sound images is in a gray area. Thus,

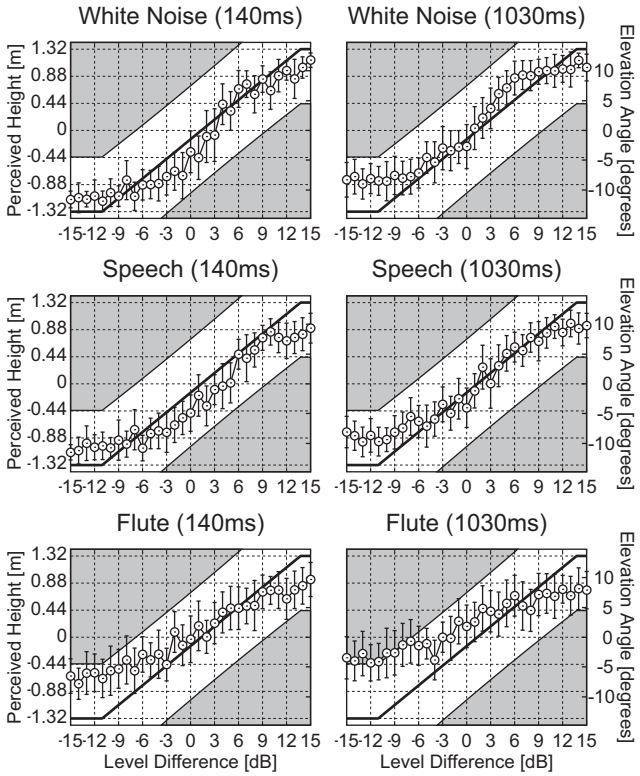

Fig. 12 Panning curves and differential limens of the panning condition in the listening test.

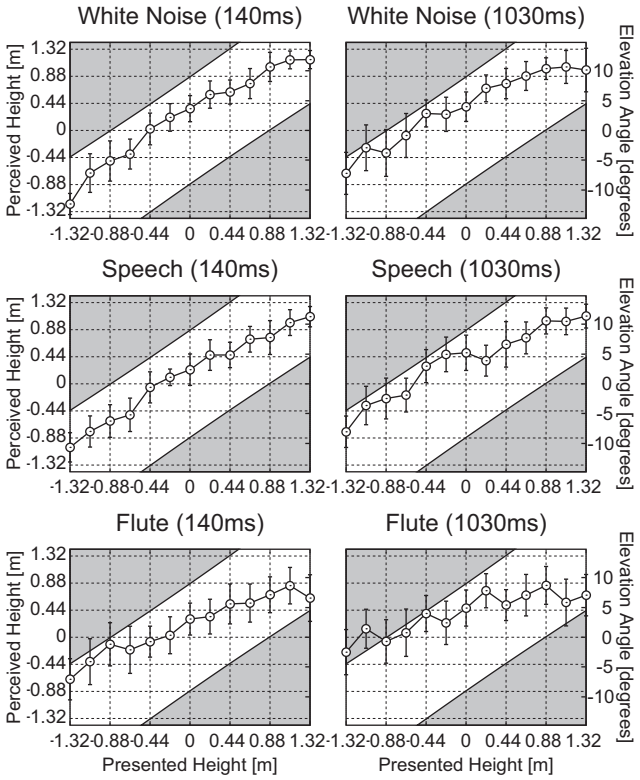

Fig. 13 Results and differential limens of the control condition in the listening test.

when the sound source is a flute, the effect of the reverberation time on the height perception of sound sources should be evaluated as future interesting topics because viewers may not perceive the height of the sound source itself because of the reverberation time.

4. Audio-visual experiment

In this section, the audio-visual experiment evaluating the audio-visual performance of the proposed system is described.

4.1 Environment and conditions

The experiment was performed in a room where a 200-inch rear-projection visual screen was set up. Two projectors for the 2D video of the left and right eyes are set up behind the screen. Because polarization plates are set up in front of the projectors, viewers can see the 3D video by wearing the polarization glass. The reverberation time of the room was 258 ms, and the background noise level had an A-weighted level of 41 dB.

Loudspeakers were manufactured by mounting a loudspeaker unit (Fostex: FE103En) on a loudspeaker enclosure (width: 11 cm, depth: 25 cm, height: 11 cm). In order to place manufactured loudspeakers at the upper and lower sides of the screen densely in the horizontal direction, eighty-two loudspeakers were placed as shown in **Fig.14**. Note that loudspeakers were placed in the forward position at 0.5 m from the screen because loudspeakers could not be placed over and under the screen, which was fixed to the ceiling and floor of the room by wires. The total width of the two loudspeaker arrays was 4.51 m (= 11 cm × 41). Considering the proper viewing distance in the developed large-screen multiview 3D video display system[2], the viewing distance should be set at 5.5 m from the screen. However, because the video operation desk was fixed at the back of the room, the viewing distance was set at 5.2 m from the screen. The proper viewing width of the developed system is 2 m across, centered around the front position of the screen when the viewing distance is 5.5 m from the screen. In this experiment, two viewing positions were set at a front position from the screen and at a lateral position, which was 2 m to the left of the front position. The height of two viewing positions was set to 1.4 m at the ear position of viewers. In addition, in order to compare the proposed system with the conventional system, two loudspeakers were placed at the left and right sides of the screen. These loudspeakers were also placed at a forward position at 0.5 m from the screen, and the height of these loudspeakers was also 1.4 m. The sound pressure level was set to an A-weighted level of approximately 70 dB in the front viewing posi-

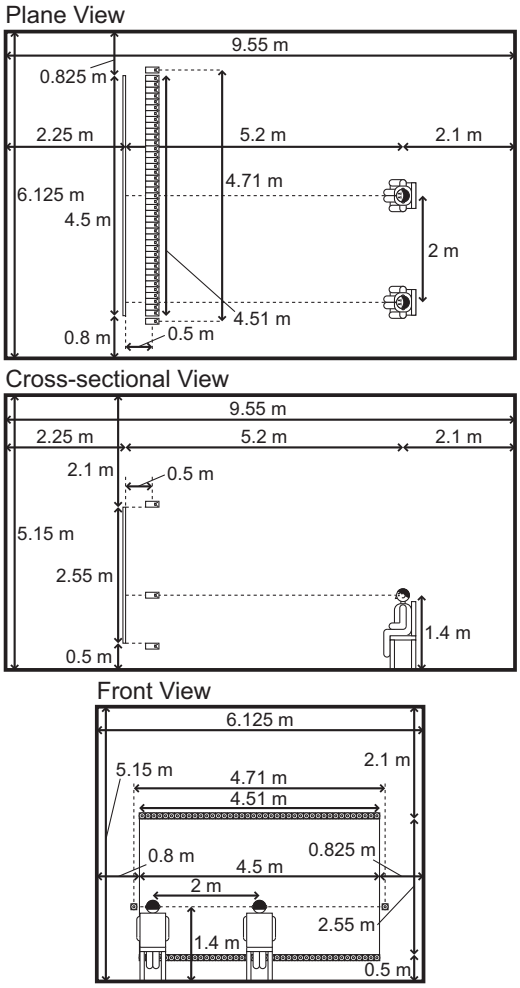

Fig. 14 Position of viewers, screen, and the loudspeaker array in the audio-visual experiment.

tion.

Two types of sounds (white noise and speech) were used as the sound source. The experimental conditions in this experiment are shown in **Fig.15**. The gray loudspeakers denote the loudspeaker from which a sound is not replayed in each condition. In both conditions, grid points and lines are always presented to the screen as a 2D image. In the proposed system condition (a), two loudspeakers placed at the upper and lower sides of the screen were selected according to the horizontal position of the presented sound images, and the sound calculated from the sound source signal, $s(n)$, was replayed from two selected loudspeakers according to the following equations:

$$x_U(n) = a_U s(n), \tag{18}$$

$$x_D(n) = a_D s(n), \tag{19}$$

where $x_U(n)$ and $x_D(n)$ denote the sound signals replayed from two loudspeakers of the upper and lower sides, respectively, and a_U and a_D (the gain coefficients in each sound signal) are calculated from the level difference ,ΔA [dB], as follows:

$$a_U = \frac{10^{\frac{\Delta A}{20}}}{\sqrt{10^{\frac{\Delta A}{10}}+1}}, \tag{20}$$

$$a_D = \frac{1}{\sqrt{10^{\frac{\Delta A}{10}}+1}}. \tag{21}$$

In this experiment, the level difference, ΔA, was set according to the panning curve in Section 3 as follows:

$$\Delta A = \frac{P_V + 0.1437}{0.1065}, \tag{22}$$

where P_V denotes the vertical position of sound images. The height of the sound images is the same as that of the ear position of the viewers if P_V is -0.33. In the stereo condition (b), according to the conventional stereophonic audio, the sound calculated from the sound source signal, $s(n)$, was replayed from two loudspeakers on the left and right as follows:

$$x_L(n) = a_L s(n), \tag{23}$$

$$x_R(n) = a_R s(n), \tag{24}$$

where $x_L(n)$ and $x_R(n)$ denote the sound signals replayed from two loudspeakers on the left and right side, respectively, and a_L and a_R (the gain coefficients in each sound signal) are calculated according to the tangent law[15] as follows:

$$a_L = \frac{1 - \frac{\tan\theta}{\tan\theta_0}}{\sqrt{2\{1+(\frac{\tan\theta}{\tan\theta_0})^2\}}}, \tag{25}$$

$$a_R = \frac{1 + \frac{\tan\theta}{\tan\theta_0}}{\sqrt{2\{1+(\frac{\tan\theta}{\tan\theta_0})^2\}}}, \tag{26}$$

where θ and θ_0 denote the angles of the sound images and the two loudspeakers on the left and right in the front viewing position, respectively. In this experiment, the angles were set as follows:

$$\frac{\tan\theta}{\tan\theta_0} = \frac{P_H}{2.355}, \tag{27}$$

where P_H denotes the horizontal position of a sound image. If P_H is 0, the sound image is placed at the center of the screen. The curves of a_L and a_R are shown in **Fig.16**.

3D videos shown in **Fig.17** were simultaneously presented in addition to the sound. The equipment for playing 3D videos was synchronized with audio-playing devices by Longitudinal Timecode (LTC). The 3D videos of the loudspeaker (left-hand side of Fig.17) and the character (right-hand side of Fig.17) were presented when the sound sources were white noise and speech.

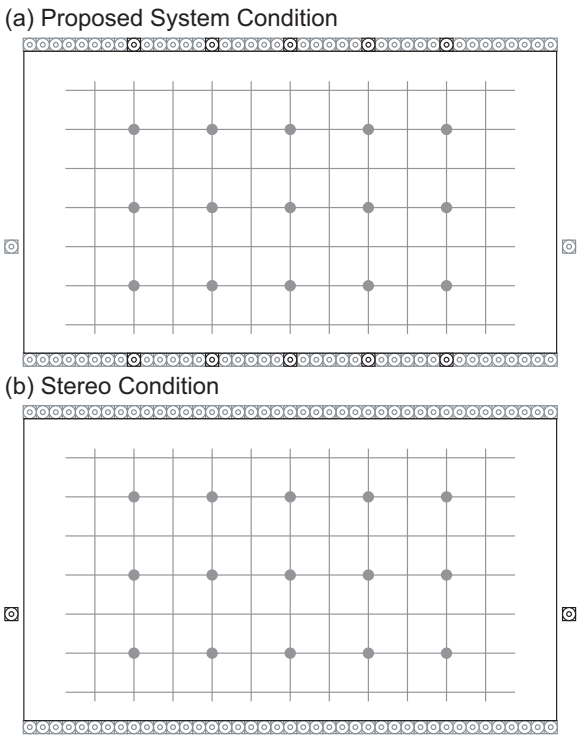

Fig. 15 Sound conditions used in the audio-visual experiment.

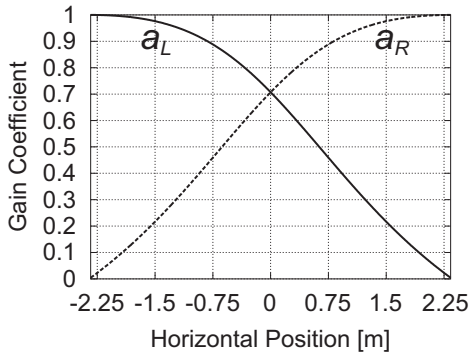

Fig. 16 Gain coefficients of two loudspeakers on the left and right side in the audio-visual experiment.

Table 2 Experimental conditions in the audio-visual experiment.

Index	Sound	3D video
(I)	Stereo	Sound only
(II)	Proposed system	Sound only
(III)	Stereo	Sound & video
(IV)	Proposed system	Sound & video

3D videos were depicted by MAYA software[16]. The proper viewing distance and the parallax of 3D videos were 5.5 m and 0.0625 m, respectively. Thus, four experimental conditions listed in **Table 2** were set in this experiment. Because 3D viewing videos change according to the viewing positions in the developed 3D video display system, the presented 3D videos were changed according to the viewing positions in this experiment.

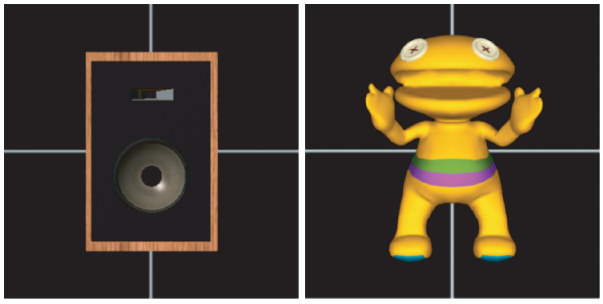

Fig. 17 3D videos used in the audio-visual experiment (Left: white noise, Right: speech).

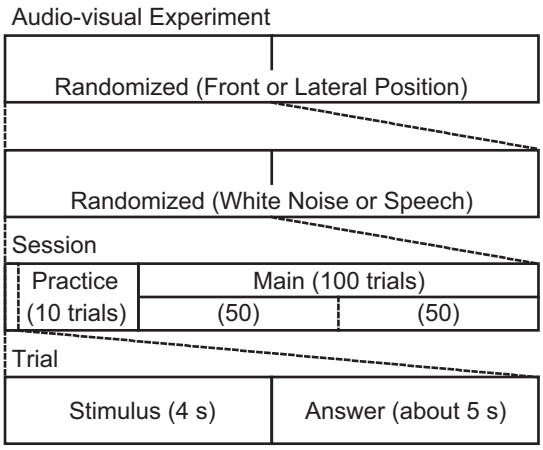

Fig. 18 Flowchart of the audio-visual experiment.

Table 3 Sound image positions in the audio-visual experiment.

Index	P_H [m]	P_V [m]	Index	P_H [m]	P_V [m]
1	-1.32	0.66	9	0.66	0
2	-0.66	0.66	10	1.32	0
3	0	0.66	11	-1.32	-0.66
4	0.66	0.66	12	-0.66	-0.66
5	1.32	0.66	13	0	-0.66
6	-1.32	0	14	0.66	-0.66
7	-0.66	0	15	1.32	-0.66
8	0	0			

4. 2 Design and procedure

Twelve subjects (age: 21–40, six males and six females), of which the audibility was normal in daily life, participated as viewers in this experiment. The flowchart of the audio-visual experiment is shown in **Fig.18**. The experiment was divided into four sessions for viewing positions and sound sources. The presented orders of the sessions were counterbalanced in all viewers. Ten practice trials and one hundred main trials were performed in each session. During the main trials, rest periods were allowed after every set of fifty trials. The presentation orders of the trials were randomized for each viewer. The position of the sound images and the detail of the practice and main trials are shown in **Tables 3** and **4**, respectively.

The viewers were instructed to report the perceived

Table 4 Practice and main trials in the audio-visual experiment.

	Element	Note
Practice (10)	= 2 conditions × 5 positions	(II) & (IV) of Table 2 1, 5, 8, 11 & 15 in Table 3
Main (100)	= [1 condition × 5 positions + 3 conditions × 15 positions] × 2 repetitions	(I) in Table 2 $P_H = -1.32, -0.66, 0, 0.66, 1.32$ (II)-(IV) in Table 2 1-15 in Table 3

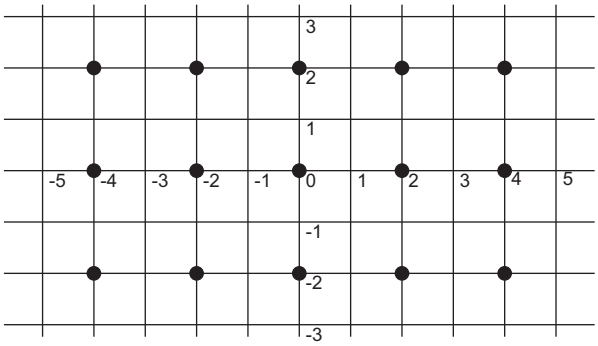

Fig. 19 Relation between perceived positions of sound images and answer grids in the audio-visual experiment.

position of the sound images by listing the indexes of the positions in an answer sheet. Note that the viewers were instructed to gaze at a 3D object when the 3D video and the sound were presented. The relation between the perceived position and the answer grids is shown in **Fig.19**. This grid corresponds to the grid lines and points shown in Fig.15. The position of black circles in Fig.19 corredponds to the position of presented sound images shown in Table 3. The viewers could choose a horizontal index of 11 patterns (from -5 to 5) and a vertical index of 7 patterns (from -3 to 3). If viewers perceived multiple sound images in the trials, the viewers could list multiple indexes in an answer sheet. The viewers were allowed to move their heads and upper bodies freely while listening to the sounds.

4. 3 Results and discussion

(1) Response rate of synthesized sound image

Because viewers also entered the number of heard sound images in this experiment, viewers perceived one synthesized sound image if the entered number of indexes in the trials was one. The response rates were calculated based on the number of trials in which viewers listed one index in order to evaluate whether viewers could perceive a synthesized sound image in the proposed system condition.

The results of the response rates are shown in **Fig.20**. Error bars denote the 95% confidence interval of the response rates. In the proposed system condition (i.e., condition (II) and (IV)), response rates are always 1. Thus, it is indicated that viewers can always feel synthesized sound images in the proposed system.

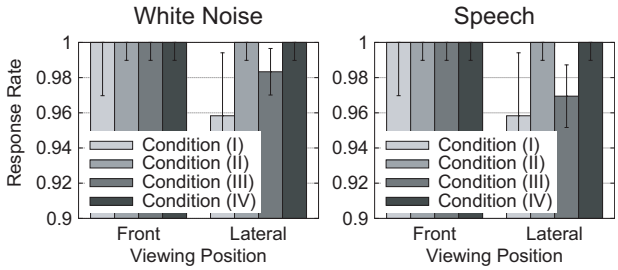

Fig. 20 Response rates in the audio-visual experiment.

(2) Perceived height of synthesized sound image

After eliminating the answers where viewers listed multiple indexes, the averages of the horizontal and vertical indexes were calculated from the answer indexes of viewers. Results of the averages in each experimental condition are shown in **Figs.21-24**. Error bars of horizontal and vertical directions denote the 95% confidence interval of the averages of the horizontal and vertical indexes. Because the gray circles denote the presented positions of the sound images, it is shown that viewers accurately feel sound images at the presented position if the perceived positions of sound images are close to the gray circles.

When the sound is only presented in the stereo condition, although the horizontal localized positions of the sound images are generally accurate in the front viewing position, the vertical localized positions of the sound images are higher than the input position. In the lateral viewing position, the horizontal and vertical positions of the sound images are not accurately localized. This is attributed to the fact that the stereophonic system assumes that viewers listen to a sound in the front viewing position.

On the other hand, in the proposed system, the horizontal localized accuracy of the sound images is improved in the lateral viewing position. The vertical localized accuracy of sound images is improved in the front viewing position. Thus, when only the sound is presented in the proposed system, it is indicated that the localized accuracy of the sound images is improved as compared with the stereophonic system at any viewing position.

When the sound and 3D video are presented in the stereo condition, the localized position of the sound images is generally accurate because of the ventriloquism effect if the viewing position is frontal. However, in the

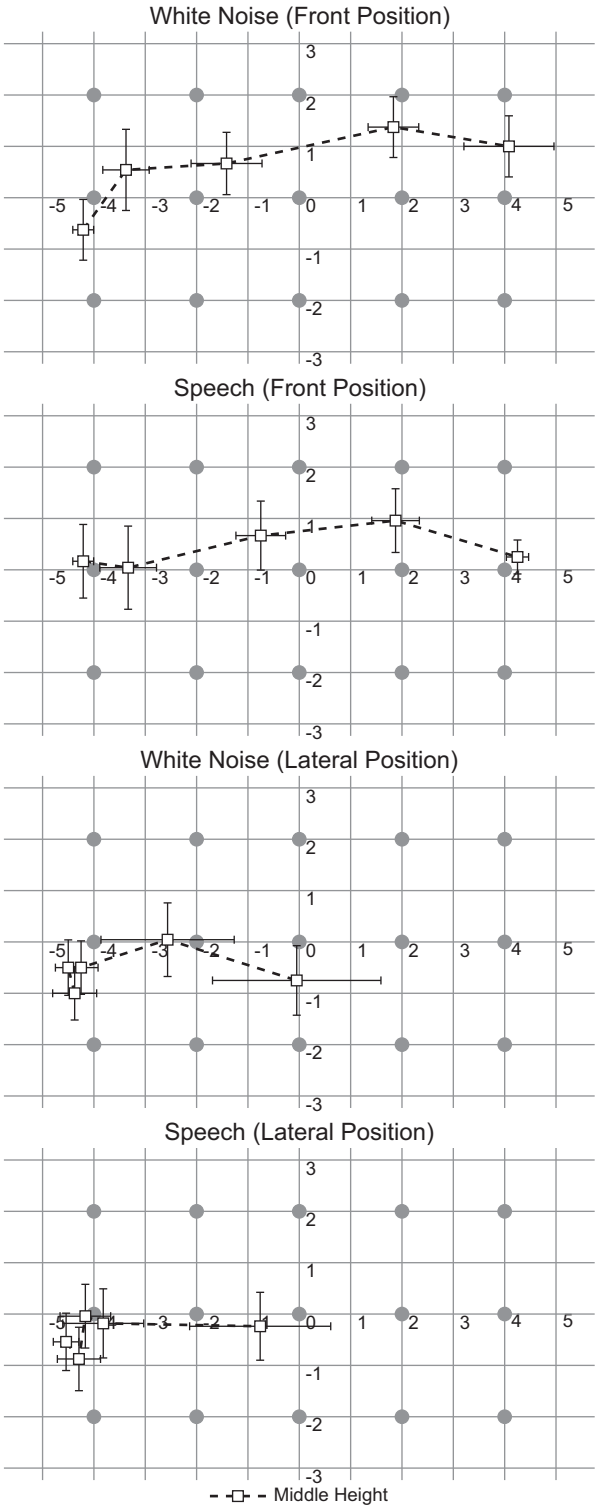

Fig. 21 Results of condition (I) (Stereo sound is only presented) in the audio-visual experiment.

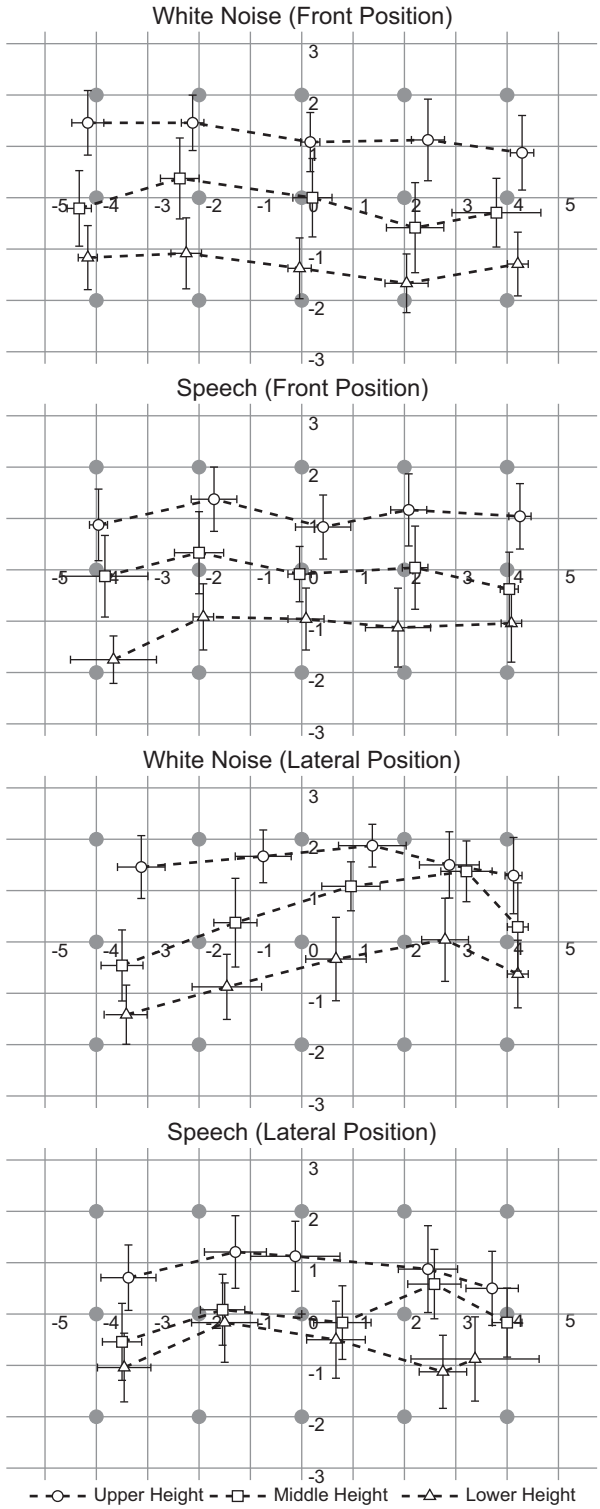

Fig. 22 Results of condition (II) (Sound of proposed system is only presented) in the audio-visual experiment.

lateral viewing position, although seven viewers localized the sound images at the position of the 3D video because of the ventriloquism effect, five viewers did not localize the sound images at the position of the 3D video because the ventriloquism effect does not occur. As a result, the averaged localized position of the sound images is biased to the left side.

On the other hand, the localized position of the sound images is the same as the position of 3D objects in the lateral viewing position in the proposed system. Thus, when the sound and 3D video are presented, it is indicated that the proposed system is effective as compared with a conventional system such as stereophonic audio

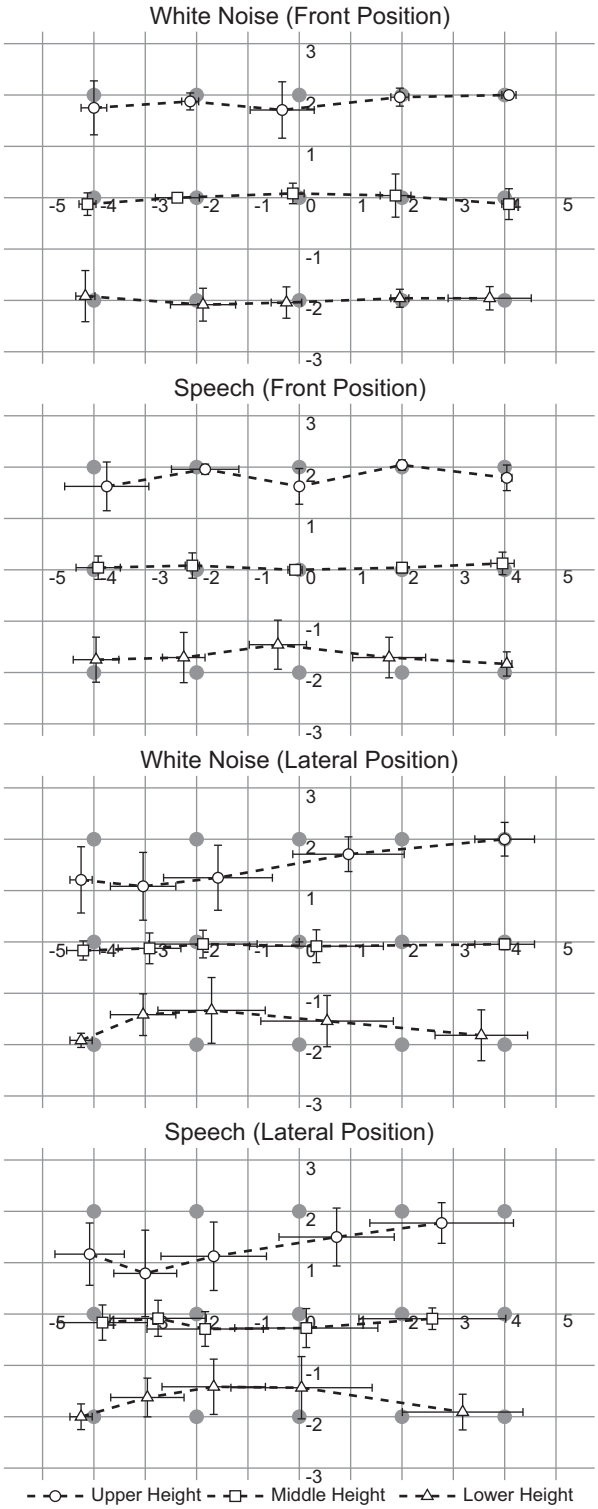

Fig. 23 Results of condition (III) (Stereo sound and 3D video are presented) in the audio-visual experiment.

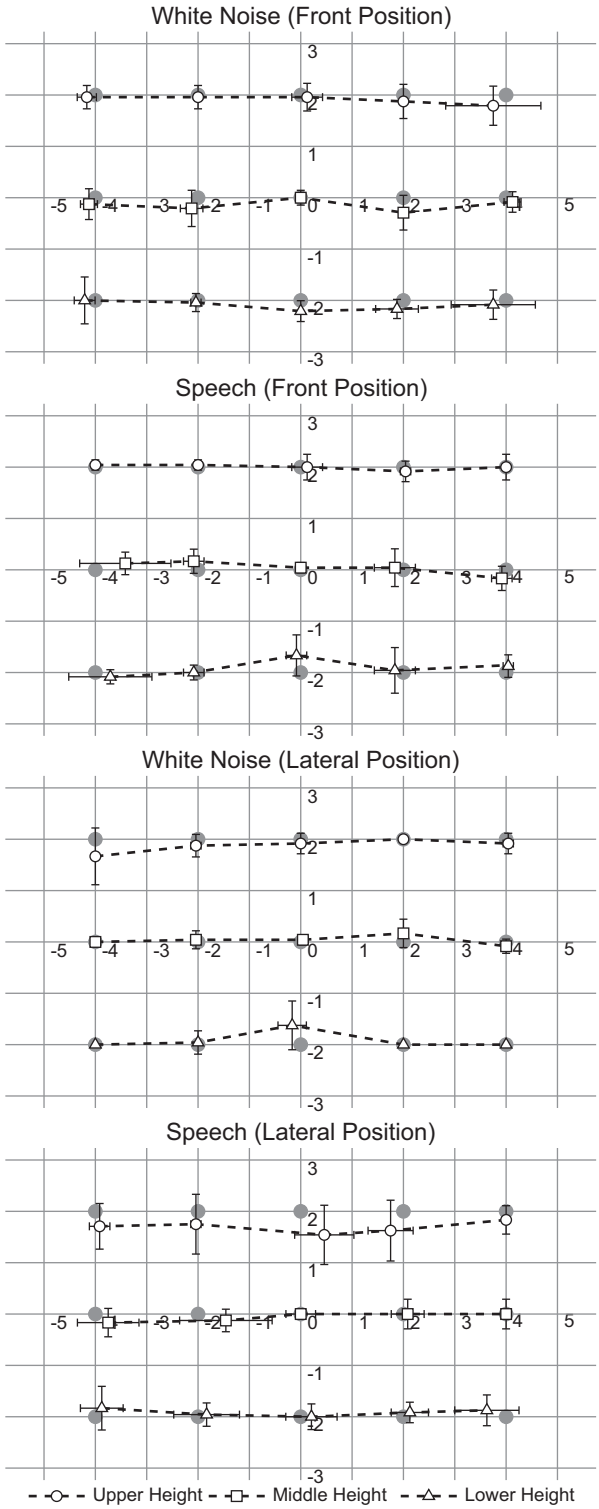

Fig. 24 Results of condition (IV) (Sound of proposed system and 3D video are presented) in the audio-visual experiment.

because viewers can always perceive the sound images at the position of the 3D objects at any viewing position.

5. Conclusion

In this paper, in order to match the developed large-screen multiview 3D video display system, a novel 3D audio system based on multiple vertical panning (MVP) was proposed. In order to evaluate the auditory performance of the proposed system, the listening test was designed by using a loudspeaker array in which twenty-seven loudspeakers were vertically aligned. As a result, it was indicated that the auditory performance of

the proposed system was so high that listeners could not discriminate the differences among the perceived heights of the sound images. In order to evaluate the audio-visual performance of the proposed system, an audio-visual experiment was performed by using a loudspeaker array in which eighty-two loudspeakers were placed at the upper and lower sides of the 200-inch screen. As a result, the proposed 3D audio system was effective as compared with a conventional system such as stereophonic audio because viewers could always perceive the synthesized sound images at the position of the 3D object at any viewing position when the sound of the proposed system was presented with 3D video.

In future work, the feasibility of a practical realization of the proposed system should be studied by reducing the number of loudspeakers and by constructing the method of recording and transmission. The means of expression of the 3D sound distance by changing sound pressure amplitudes should also be developed.

6. Acknowledgment

The authors would like to thank Dr. S. Iwasawa for constructing the environment of the audio-visual experiment. The authors would like to thank Dr. L.-G. Roberto and Dr. M. Makino for depicting 3D videos in the audio-visual experiment. The listening test and the audio-visual experiment in this paper were performed with the approval of the ethical committee of the National Institute of Information and Communications Technology (NICT), Japan.

References

1) K. Enami : "Research on Ultra-realistic Communications", ECTI Trans. Electr. Eng. Electron. Comm., 6, 1, pp.22-25 (Feb. 2008)
2) S. Iwasawa and M. Kawakita : "Quantifying Capabilities of the Prototype 200-inch Automultiscopic Display", Proc. International Conference on 3D Systems and Applications, S1-5, pp.105-109 (June 2012)
3) M. Kawakita, S. Iwasawa, G. Sabri and N. Inoue : "Development of Glasses-free 3D Video System", Proc. International Universal Communication Symposium, pp.323-327 (Oct. 2011)
4) A. D. Blumlein : "Improvements in and Relating to Sound-transmission, Sound-recording and Sound-reproducing Systems", British Patent, 394325 (1931)
5) ITU-R Recommendation BS.775-1: "Multichannel Stereophoneic Sound System with and without Accompanying Picture" (1992-1994)
6) J. Blauert : "Spatial hearing", revised edition, MIT Press, Cambridge, Mass., pp.372-392 (1997)
7) K. Hamasaki, T. Nishiguchi, R. Okumura, Y. Nakayama and A. Ando : "A 22.2 Multichannel Sound System for Ultrahigh-definition TV (UHDTV)", SMPTE Mot. Imag. J., 117, 3, pp.40-49 (April 2008)
8) M. A. Poletti : "Three-dimensional Surround Sound Systems Based on Spherical Harmonics", J. Audio Eng. Soc., 53, 11, pp.1004-1025 (Nov. 2005)
9) A. J. Berkhout, D. de Vries and P. Vogel : "Acoustic Control by Wave Field Synthesis", J. Acoust. Soc. Am., 93, 5, pp.2764-2778 (May 1993)
10) M. R. Schroeder, D. Gottlob and K. F. Siebrasse : "Comparative Study of European Concert Halls: Correlation of Subjective Preference with Geometric and Acoustic Parameters", J. Acoust. Soc. Am., 56, 4, pp.1195-1201 (Oct. 1974)
11) S. Ise : "A Principle of Sound Field Control Based on the Kirchhoff-Helmholtz Integral Equation and the Theory of Inverse Systems", ACUSTICA - Acta Acustica, 85, 1, pp.78-87 (Jan./Feb. 1999)
12) T. Hirahara, C. Muller and Y. Tohkura : "Structure and Acoustic Characteristics of the ATR Variable Reverberation Room", J. Acoust. Soc. Jpn., 48, 5, pp.301-308 (May 1992) (in Japanese)
13) V. Pulkki : "Localization of Amplitude-panned Virtual Sources II: Two- and Three-dimensional Panning", J. Audio Eng. Soc., 49, 9, pp.753-767 (Sep. 2001)
14) J. Blauert : "Spatial hearing", revised edition, MIT Press, Cambridge, Mass., pp.37-50 (1997)
15) B. Bernfeld : "Attempts for Better Understanding of the Directional Stereophonic Listening Mechanism", Proc. 44th Audio Eng. Soc. Convention, C-4, pp.1-24 (Feb. 1973)
16) Website of Maya Software, `http://www.autodesk.co.jp/maya`

Toshiyuki Kimura received B. E., M. A. and Ph. D. degrees from Nagoya University, Japan, in 1998, 2000, and 2005, respectively. He was a research fellow of the Japan Society for the Promotion of Science, a research fellow of Nagoya University, Japan, and a research associate of Tokyo University of Agriculture and Technology, Japan, from 2003 to 2007. He is currently a limited-term researcher of the National Institute of Information and Communications Technology, Japan, since 2007. His research interests include 3D systems, spatial perception, and array signal processing. He is a member of the IEICE, ASJ, VRSJ, HIS, and AES.

Hiroshi Ando received his B. S. in Physics in 1983, his M. A. in Psychology in 1987 from Kyoto University, and his Ph.D. in Computational Neuroscience from Department of Brain and Cognitive Sciences, MIT in 1993. In 1992, he joined a research group at ATR (Advanced Telecommunications Research Institute International) and became Head of Department of Cognitive Dynamics at ATR until 2010. He is currently Director of Multisensory Cognition and Computation Laboratory at Universal Communication Research Institute, NICT. He has also been guest professor at Osaka University. His research has focused on cognitive and brain mechanisms of multisensory perception and multisensory man-machine interfaces.