CONCERT SCOPE HEADPHONES

Masatoshi Hamanaka
University of Tsukuba
hamanaka@iit.tsukuba.ac.jp

Seunghee Lee
University of Tsukuba
lee@kansei.tsukuba.ac.jp

ABSTRACT

We designed *concert scope headphones* that are equipped with a projector, an inclination sensor on the top of the headphones, and a distance sensor on the outside right headphone. We previously developed sound scope headphones that enable users to change the sound mixing depending on their head direction. However, the system could not handle images. In contrast, our headphones let the user listening to and watching a music scene scope on a particular part that he or she wants to hear and see. For example, when listening to jazz, one might want to more clearly hear and see the guitar or sax being played. The user can hear the guitar or sax sound from a frontal position by moving their head to the left or right. The user can adjust the distance sensor on the headphones and focus on a particular part they want to hear and see by simply putting their hand behind their ear.

1. INTRODUCTION

Our goal is to create an audio-visual interface that enables for the separation of the listening and watching of each performance if the user wants to clearly hear and see a particular performer. A musical expert at a concert, such as a conductor, can distinguish between the sounds coming from each performer, even if there are many performers playing the same instrument part. However, it is hard for musical novices to distinguish each performer's sound. Therefore, we have developed concert scope headphones that let a user listening and watching a music performance to scope on a particular part of the performance that he or she wants to hear and see.

We had to define two requirements for the interface. First, the user can control it by simply performing the natural actions related to listening. These actions include those made by people in a concert hall audience (e.g., they turn their heads in the direction of the viewing and/or listening target). Thus, with this interface, a user can better enjoy videos of concerts by selecting particular areas and/or performers on the stage by turning his or her head in their direction and cupping a hand to their ear. Second, the constructed device incorporating this interface needs to be small enough for home use. Since projectors have been getting smaller and smaller, we were able to develop a headphone device equipped with a compact projector, an inclination sensor, and a distance sensor (Figure 1). This device detects the user's head direction, detects the distance between the user's cupped hand and ear, and outputs the corresponding image and sound. Moreover, it is small enough to be used in the home as well as many other environments.

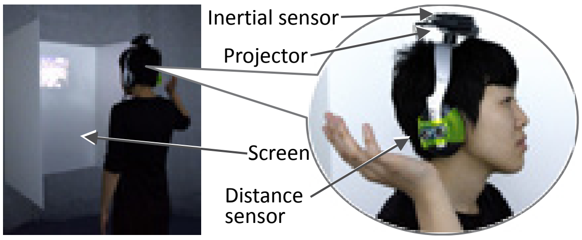

Figure 1. Concert scope headphones.

Figure 2 shows the system flow of our *concert viewing headphones*. First, the user's head direction is detected using inertial sensors. Next, the portion of the wide-angle image that captures the whole stage corresponding to the head direction is extracted, and this portion is projected on the screen. At the same time, the recorded sounds are mixed to emphasize the sounds of the performers within the extracted portion (i.e., the projected image). If the user cups a hand to his or her ear to hear better, the projected image is enlarged to a degree corresponding to the distance between the user's hand and the distance sensor attached to one of the headphones, enabling the user to better focus on a particular performer.

The *concert scope headphones* have three particular features.

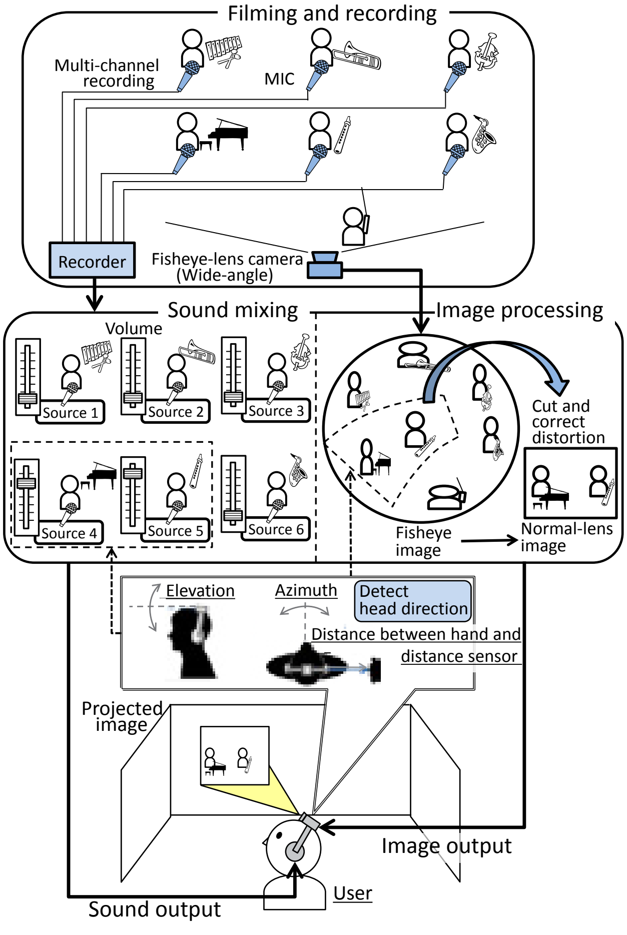

Figure 2. System flow of audiovisual interface.

Use of Imaginary Microphones. Ideally, we capture the sound coming from each performer through a microphone attached to the performer's music stand because the sounds would be mixed to emphasize those within the selected scope. However, this is difficult in terms of time and cost if there are many performers, such as for an orchestra. Therefore, there is a problem with the mixing sounds if performers without microphones are in the selected scope. We solve this issue by creating imaginary microphones for those performers without a real microphone, making it possible to mix sounds as if each performer had a real microphone. The sounds for two performers with real microphones who are near a performer without one are mixed, creating an imaginary microphone for that performer.

This mixed sound is used as the sound recorded in proportion to the distances between the performer without a real microphone and the two performers with real ones.

Use of Image Captured with Fisheye-Lens. A wide-angle image of the whole stage is captured using a camera with a circular fisheye-lens. The user selects the target scope from this image, enabling the user to selectively appreciate a particular portion of the image. The lens captures a 180º image that is characterized by low distortion in the center and large distortion around the edges. The distortion in the extracted area is corrected when the image corresponding to the gaze orientation is extracted. A non-distorted image is then projected.

Projection of Image on Front Wall and Two Side Walls. If the 180º image of the entire stage was projected on only the front wall, it would be difficult to clearly see the performers at the ends of the stage. For this reason, the *concert scope headphones* are premised on being used indoors and on images being projected on not only the wall in front of the user but also the two walls to the sides. Thus, the user can evenly view the entire stage because the performers at the ends of the stage appear on the corresponding side walls when the user's head turns in the given directions. However, when an image is projected on a side wall, it is projected at a tilt, so it is distorted and forms a trapezoid (keystone distortion). This distortion is compensated for by distorting the image counter to the keystone distortion before it is projected.

This paper is organized as follows. In Section 2, we mention the related works, and in Sections 3 and 4 we explain the system for image processing and sound mixing. In Section 5, we describe the implementation and present our experimental results and conclusions in Sections 6 and 7.

2. RELATED WORKS

This is the first report of a device based on a pair of headphones with a projector that enables images and sound sources to be selectively viewed and listened to. However, there has been some related research.

We can prepare a multi-channel recording and adjust the volumes and panoramic potentiometers of an audio mixer to listen to each player's performance separately. However, a commercial audio mixer is too complicated for a musical novice to operate. For example, it is difficult for a novice to increase the guitarist's volume and turn the guitarist's panoramic potentiometers to the center the moment a guitarist begins a solo, and at the same time lower the other players' volumes and properly adjust their panoramic potentiometers.

A music spatialization system [1, 2] allows a user to control the localization of each part in real time using a graphical interface. However, these systems encounter the same problem as for a commercial audio mixer because it is difficult for a musical novice to appropriately change the location of each part using a graphical interface the moment a solo part begins.

Previously reported headphones with sensors for detecting the direction the user is facing or the location of the head can escalate the musical presence and create a realistic impression, but they do not control the volumes and panoramic potentiometers of each part in accordance with the user's wishes [3-6].

We previously developed an interface called sound scope headphones that enabled a user to appreciate sounds by selecting particular sound sources by mixing the sounds on the basis of the user's head direction [7]. However, the interface did not handle images. Furthermore, the interface did not use real sounds recorded at a concert but rather the music in the RWC Music Database [8]. In this study, we use real sounds recorded at a concert.

TWISTER [9] presents stereoscopic images to a user. The user stands in a cylindrical booth, which displays live 360º full-color panoramic and stereoscopic images. Ensphered Vision [10] projects images onto a full-surround spherical screen that surrounds the user, enabling him or her to experience virtual reality. These two systems are not suitable for home use because they must be large enough to cover either one's whole body or their head.

With multiview or wide-angle image systems [11, 12], the user can arbitrarily select the portion of an image to be viewed from a panoramic image captured with omnidirectional cameras or with a camera equipped with a fisheye lens. Such systems are controlled by manipulating a remote control or a mouse. Therefore, the user cannot appreciate the image and sound by simply performing the natural actions related to listening.

We have been constructing headphones with sensors and a projector in which the user wearing this device turns his or her head and the projected image is adjusted corresponding to this movement [13]. Thus, the user can recognize in which direction he or she is looking in the image. However, the sound output of the headphones is monaural, and then, the localization between the image and sound of each performer becomes unnatural. Therefore, we improved the headphones that we call *concert scope headphones* by naturally correlating between the sound mixing and the position of each instrument in the projected image.

3. IMAGE PROCESSING SYSTEM

The *concert scope headphones* are comprised of two systems: image processing and sound mixing. The image processing system comprises image extraction, keystone distortion correction, and image output.

3.1. Image Extraction

For a camera using a circular fisheye-lens, the landscape is reflected in an imaginary hemisphere and projected onto the imaging area at right angles. Thus, a fisheye image is produced and output with low distortion in the center and large distortion around the edges. The distortion is corrected by centering the extracted image onto a point detected by a three-axis attitude sensor. In particular, we correct the distortion by reversing the process on the basis of the principle for capturing images using a fisheye-lens. Figure 3 shows a distorted image and a corrected (non-distorted) image.

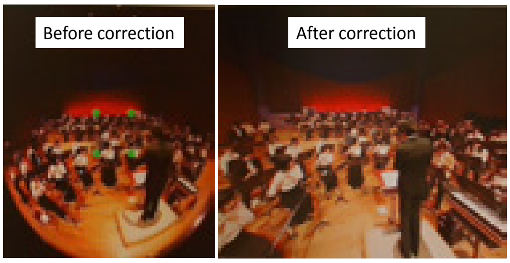

Figure 3. Distorted and corrected images.

3.2. Keystone Distortion Correction

The image projected on a wall is at a tilt, so it is either magnified or demagnified and forms a trapezoid. This keystone distortion must be corrected. The image is thus processed in accordance with the elevation and direction of the user's head so that it is not distorted when it is projected. This processing distorts the image counter to the keystone distortion. The correction is calculated by using the measurements taken from the direction and inclination sensors. The magnification or reduction percentages are determined by comparing the projection distance with the image projected on the front wall (Figure 4).

3.3. Rotation Correction

When a user moves his or her head to the right, left, up, or down, it is easy to naturally tilt their head to one side. If a user tilts his or her head to one side, the projector also rotates and then the frame of projection is rotated. This rotation correction remains horizontal in the projected image by inverse rotating the image before projection (Figure 5).

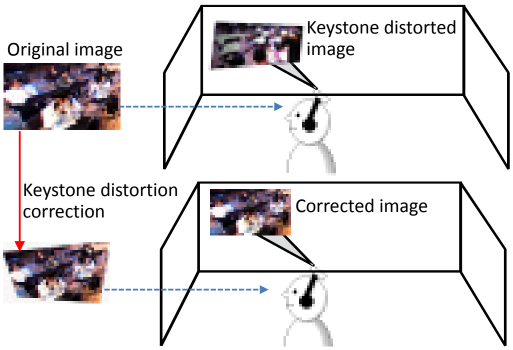

Figure 4. Keystone distortion correction.

3.4. Image Output

Since it is visually unnatural to project 180º images on a flat surface, the *concert scope headphones* are premised, as mentioned above, on being used indoors and on images being projected on not only the front wall but also two side walls. When an image is projected on a side wall, the correction processing appropriately switches its shape for the sidewall projection.

4. SOUND MIXING SYSTEM

The sound mixing system comprises imaginary microphone creation, distance calculation, and mixing.

4.1. Imaginary Microphone Creation

We record sound sources (R_n) using microphones corresponding to each of the instrumental parts of the given music, that is, a microphone is attached to the music stand of one performer for each group of performers playing the same instrument. For those performers without a microphone, the sound of an imaginary microphone (I_n) is created for each one by mixing the sounds recorded by the closest real microphones on the basis of the distance between the performer and the position of the real ones. It is possible that the sound of an imaginary microphone to include instrument sounds that are distant from the imaginary microphone because, in general, microphones pick up sounds from all directions. However, the sound pressure is decreased in inverse proportion to the square of the distance between a microphone and the position of a sound source. Moreover, in typical concerts, because the size of the stage is sufficiently large, the performers are spaced out on the stage with enough distance between each instrumental part of the music. Therefore, in this study, most of the sounds picked up by the microphones are from near them and make up one instrumental part. For this reason, we used a method that creates imaginary microphones on the basis of the distance between the performer and the position of the real microphones.

$$I_m = \frac{1}{n}\sum_n \frac{R_n}{dist(m,n)^2}\,, \tag{1}$$

where $dist(m, n)$ is the distance between the microphone positions of m and n in a concert hall.

4.2. Calculating amplification rate for distance from center

The amplification rate for each performer's sound is calculated on the basis of the distance between the performer and the center of the image extracted by the image processing system. The distance is calculated from the currently projected image and the position of each performer in the coordinate system of the image

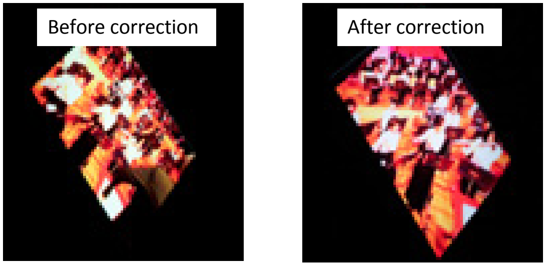

Figure 5. Rotated and corrected images.

captured using a circular fisheye-lens (Figure 6).

$$d_n = \frac{\text{Distance between center and sound source } n}{\text{Half length of a diagonal line of projected image}} \quad (2)$$

The sounds from the performers who are within the projected image are emphasized so that they approach the center of the projected image. The prepared functions are used to determine the mixing rates needed to adjust the sound volumes to increase the volumes with a decrease in the calculated distance.

We prepared five amplification rates h_n, as shown in Figure 7. In this figure, line 1 is constant with the distance, and the volume is the same for all the microphones. Line 2 is a negative gradient. If this function is used, each performer's projected volume has an inverse relationship to the performer's distance from the center of the image. Line 3 is a normal distribution. If this function is used, the user can hear the performers located at the center of the image and around the center. The volumes for the performers at the edge of the image are almost zero. Line 4 is the inverse of the distance from the center of the image. If this function is used, each performer's projected volume has an inverse relationship to the performer's distance from the center of the projected image. Line 5 is the inverse square of the distance from the center of the projected image. If this function is used, each performer's projected volume has an inverse relationship to the performer's distance from the center of the image. We describe our evaluation of these functions in Section 6.

If a user controls the projected image to view the entire stage, it is possible they do not want to emphasize on the sound from a particular performer but to listen to all the performers' sounds. For this reason, the user can select from two sound mixing modes. The first one is the "constant mode," in which the volume is the same for the sounds from all the performers who are within the projected image (line 1 in Figure 7). The other one is the "emphasis mode," in which, as mentioned above, the sounds of the performers who are within the projected image are emphasized so that those performers are centered in the projected image (e.g., line 5 in Figure 7).

4.3. Calculating amplification rate for horizontal position

We control the panoramic potentiometers of each sound source from the real and imaginary microphones to more naturally determine each performer's position in the projected image and localize the sound. This means that a performer on the right side of the projected image can be heard from the right and a performer on the left side can be heard from the left. The amplification rates for the horizontal positions h_n^{right} and h_n^{left} are

$$h_n^{\text{right}} = \frac{d_n^{left}}{d_n^{right} + d_n^{left}}, \text{ and} \quad (3)$$

$$h_n^{\text{left}} = \frac{d_n^{right}}{d_n^{right} + d_n^{left}}, \quad (4)$$

where d_n^{left} is the distance between the position of a performer and the left end of the projected image and d_n^{right} is the distance between the position of a performer and the right end of the projected image (Figure 6).

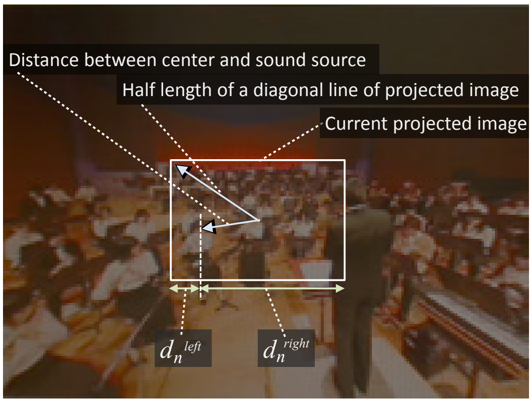
Distance between center and sound source
Half length of a diagonal line of projected image
Current projected image
d_n^{left}
d_n^{right}

Figure 6. Distance calculation.

4.4. Mixing

Each performer's sound is multiplied by the corresponding amplification rates, the sounds are added together, and they are then output.

Right-side output:

$$S_{right} = \sum_n R_n \cdot h_n \cdot h_n^{right} + \sum_m I_m \cdot h_m \cdot h_m^{right}, \quad (5)$$

Left-side output:

$$S_{left} = \sum_n S_n \cdot h_n \cdot h_n^{left} + \sum_m I_m \cdot h_m \cdot h_m^{left}, \quad (6)$$

where n is the numbers of real microphones and m is the numbers of imaginary microphones.

The *concert scope headphones* change the image and sound mixing in accordance with the orientation of the user's head. When the user cups one of their ears to operate the zoom function, the performers that are at the edges of the image are moved out of or moved away from the center of the zoomed image. As a result, the sounds of the performers at the center of the line of sight are emphasized.

5. IMPLEMENTATION

We used an attitude heading reference system (*3DM-GX3-25*™, MicroStrain Inc.) as both a direction and an inclination sensor. It outputs the Euler angles, rotation matrix, delta angle, delta velocity, acceleration angular rate, and magnetic field to a USB device small enough to be mounted on a pair of our headphones.

A proximity sensor (*Ping)))*™, Parallax Inc.) is used as the distance sensor. It uses ultrasonic reflectance to detect the distance within a 0-6 cm range, making it well suited for measuring the distance between the hand and ear. It is mounted on the right headphone. The ultrasonic proximity sensor cannot detect the distance when the

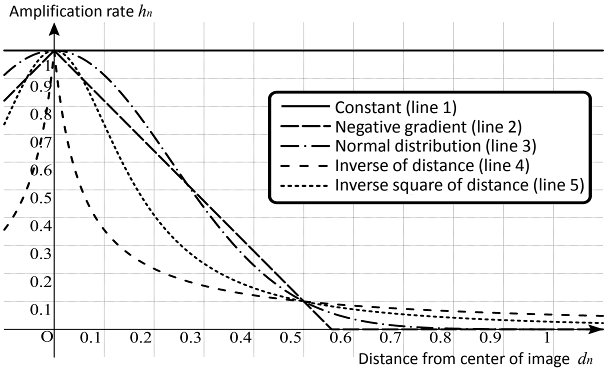

Figure 7. Mixing rate functions.

user completely covers the sensor with their hand, because it cannot receive the ultrasonic reflectance. We thus made a protector for the ultrasonic proximity sensor that was formed by using acrylic resin. So, the user cannot directly touch the sensor because of the protector, and then, it can provide accurate measurements even at a close range.

Data is transmitted from the proximity sensor using an Arduino, which is an open-source electronics prototyping platform. Since the proximity sensor outputs data using serial communication, connecting it to a PC is problematic. We used a USB-serial conversion substrate (*FT232RX*) to enable the proximity sensor-to-USB connection.
A mini USB projector was mounted on the headphones. We selected a projector with less magnetic force because the *3DM-GX3* system is susceptible to it.

The small rectangle or square space for the front and side walls is sufficient enough for image projection when using the *concert scope headphones*. We also made a special screen for exhibitions that consists of a half-through screen stuck to a bent acrylic plate (Figure 8).The screen enables the audience to see the view of the user controlling the *concert scope headphones* from off the screen.

Calibration is needed before using the *concert scope headphones*, because the mapping of the keystone distortion correction will change depending on the screen's direction now being projected among the front and side screens. The calibration has three steps. First, measure the distance from the headphones to the front screen using the distance sensor mounted on the right side headphone. Then, measure the angle of the left and right ends of the front screen using the inertial sensor mounted on the top of headphones.

6. EVALUATION

We evaluated our *concert scope headphones* by first creating an image and sound source. We filmed and recorded a University of Tsukuba Symphonic Band concert at Nova Hall, a concert hall in the city of Tsukuba, Japan. The microphones and camera were configured as shown in Figure 9. The microphones were placed on the music stands. We used 37 lavalier microphones to record the sounds because there were 37 instrumental parts in the performance. We used 10 sets of 4-channel audio recorders (*H4n*$^{\text{TM}}$, Zoom Inc.) to independently record 37 channels, because they work on batteries and are portable. An image of the entire stage with all the performers visible was captured using a 2.5m-high camera (*EOS Kiss X3*$^{\text{TM}}$, Canon inc.) with a circular fisheye-lens (*Raynox*$^{\text{TM}}$ DCR-CF187PRO, Yoshida Inc.). We experimented with the amplification rate for the distance from center h_n described in Section 4.2 to identify the natural correspondence between the changes in the image and the sound mixing.

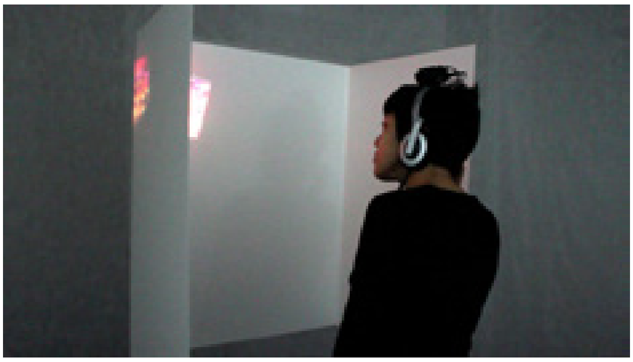

Figure 8. Semi-permeable concavity screen.

6.1. Evaluation of Functions for Mixing

We located a point at which one could hear a 440-Hz sine wave at random positions within the image. The ten evaluation participants attempted to locate this point using only their ears so as to identify the natural correspondence between the head movement and changes in the sound mixing. We recorded the time it took to do this. Each time a participant located the point, the sine wave randomly shifted to another point. This was repeated five times for each function. We defined the function with which the participants could find the sine wave the fastest as the best one. The average times for each function are listed in Table 1. The time was the shortest for the "inverse square of distance" function using the zoom function. All ten participants were adults in their 20s. In addition, three of them had been playing musical instruments on a daily basis, another three had been playing musical instruments for a period of time, and the other four had hardly played an instrument. We determined from the results of each function that the participants' musical experience did not account for any differences.

Function	Time without zoom system [s]	Time with zoom system [s]
Constant	112.5	56.2
Negative gradient	29.2	24.0
Normal distribution	32.6	17.8
Inverse of distance	30.9	20.8
Inverse square of distance	23.8	16.9

Table 1. Average time to locate 440-Hz sine wave

6.2. Evaluation by using eye-mark camera

We demonstrated that the targeted scope naturally corresponded to the changes in the sound mixing by using an eye-mark recorder. In particular, the participants repeated the action of arbitrarily looking at a particular performer and transferring their gaze to another performer. In the meantime, we examined the relation between their viewpoint and the mixing rate of each performer's sound. As a result, we determined that the sound volume of a performer who the participant was looking at was the highest in the mixing. When the participant was transferring his or her gaze, the mixing was switched in real time so that the sound volume of the performer who was closest to the participant's

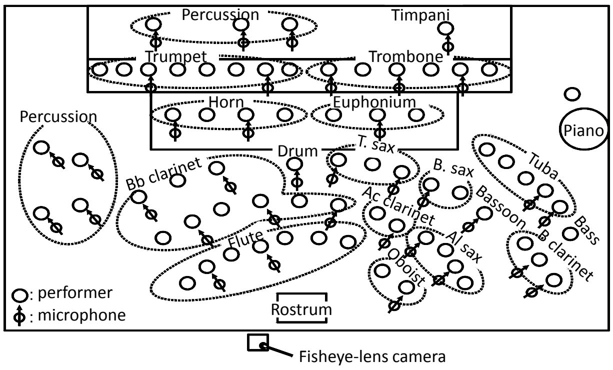

Figure 9. Microphone and camera positions.

viewpoint was the highest. However, we determined that the shifting of the image when the participant quickly turned his or her head sometimes lagged behind that of the sound. If we reduced the file size of the image and/or sound, this lag would be improved, but we think that the image and sound quality should not be degraded in the audiovisual interface. Therefore, we will improve the program of this interface so that the shifting of the image and sound is more natural.

6.3. Questionnaire

A questionnaire containing four questions was completed by the participants after the testing.

Q1: Is the relation between the shifting of the image and that of the sound natural?

Q2: Is the instrumental sound emphasized so that it approaches the center of the projected image?

Q3: When you quickly turn your head, is the relation between the shifting of the image and that of the sound natural?

Q4: What is your feeling about the characteristic or difference among each of the five functions?

Questions 1 to 3 were rated on a 5-point scale with the following possible responses: completely disagree (1), somewhat disagree (2), uncertain (3), somewhat agree (4), and completely agree (5). Table 2 lists the average scores of each function rated by the ten participants for each question. The score was the highest for the "Negative gradient" and "Inverse square of distance" functions. For question No. 4, in the "Inverse square of distance function," most participants had positive opinions, e.g., "I could hear the difference among the sounds very clearly," and "I could notice many different instrumental sounds." Therefore, we determined the effectiveness of using the inverse square of distance function.

Function	Q1	Q 2	Q3:	Average
Constant	2.6	2.7	3.1	2.8
Negative gradient	3.9	4.1	3.9	3.9
Normal distribution	3.4	3.6	3.6	3.5
Inverse of distance	3.6	3.3	3.1	3.3
Inverse square of distance	3.7	3.7	4.0	3.8

Table 2. Average score of each questionnaire

7. CONCLUSION

Our *concert scope headphones*, equipped with a projector, an inclination sensor, and a distance sensor for zoom control, enables a user to selectively view and listen to specific performers in a video-taped group performance. It has both image and sound processing functions. The image processing extracts the portion of the image selected by the user and projects it free of distortion on the front and side walls. The sound processing creates imaginary microphones for those performers without one so that the user can hear the sound from any performer. Testing using the images and sounds captured using a fisheye-lens camera and 37 lavalier microphones showed that the sound localization was the fastest when an inverse square function was used for the sound mixing. Moreover, the zoom function enabled the participants to indicate the desired sound performance.

We will research and discuss the creation of the imaginary microphones in the sound mixing system and design a method to create a more realistic sound in the near future. For example, we will record the instrumental sounds within a more narrow area on the stage by using directional microphones and by adjusting the positions of the microphones. In addition, we will estimate the acoustic transfer function of the positions of the performers without a real microphone.

In order to improve the *concert scope headphones*, we plan to conduct further experiments on both the image and sound functions to determine whether the zooming and image or sound changes are natural. Furthermore, we will conduct experiments with participants of all ages.

8. REFERENCES

[1] Pachet, F., and Delerue, O. "A mixed 2d/3d interface for music spatialization", *In Lecture Notes in Computer Science* (no. 1434), pp. 298–307, 1998.

[2] Pachet, F., and Delerue, O. "On-the-fly multi-track mixing". *Proceedings of AES 109th Convention*, 10 pages, 2000.

[3] Warusfel, O., and Eckel, G. "Listen – augmenting everyday environments through interactive soundscapes", *Proceedings of IEEE Workshop on VR for public consumption*, pp. 268–275, 2004.

[4] Wu, J., Duh, C., Ouhyoung, M., and Wu, J. "Head motion and latency compensation on localization of 3d sound in virtual reality", *Proceedings of the ACM symposium on Virtual reality software and technology*, pp. 15-20, 1997.

[5] Goudeseune, C., and Kaczmarski, H. "Composing outdoor augmented-reality sound environments", *Proceedings of the International Computer Music Conference*, pp. 83–86, 2001.

[6] Sato, K. "Development of Digital Cordless Headphone", *Pioneer R&D* 14(2), 66-73, 2004.

[7] Hamanaka, M., and Lee, S. H. "Sound scope headphones: controlling an audio mixer through natural movement", *Proceedings of the International Computer Music Conference*, pp. 155-158, 2006.

[8] Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. "RWC music database: music genre database and musical instrument sound database", Proceedings of the 4th International Conference on Music Information Retrieval, pp.229–230, 2003.

[9] Tachi, S. "TWISTER: immersive ominidirectional autostereoscopic 3D booth for mutual telexistence", *Proceedings of the Asiagraph*, pp. 1–6, 2007.

[10] Iwata, H. "Full-surround image display technologies", *International Journal of Computer Vision*, 58(3), pp. 227–235, 2004.

[11] Google Maps with Street View, http://maps. google.com/intl/en/help/maps/streetview/

[12] immersive media, http://www.immersivemedia.com/demos/

[13] Miyashita, S., Hamanaka, M., and Lee, S. H. "Concert viewing headphones" ,Proceedings of Internaional Conference on Advances in Computer Entertainment Technology (ACE2010), pp. 108-109, 2010.