Music Scope Headphones: Natural User Interface for Selection of Music

Masatoshi Hamanaka
Presto, Japan Science and Technology Agency
A.I.S.T. Mbox 604 1-1-1 Umezono,
Tsukuba, Ibaraki, 305-8568 Japan
m.hamanaka@aist.go.jp

Seunghee Lee
University of Tsukuba
Tennoudai 1-1-1,
Tsukuba, Ibaraki, 305-8574, Japan
lee@kansei.tsukuba.ac.jp

Abstract

This paper describes a novel audio only interface for selecting music which enables us to select songs without having to click a mouse. Using previous music players with normal headphones, we can hear only one song at a time and we thus have to play pieces individually to select the one we want to hear from numerous new music files, which involves a large number of mouse operations. The main advantage of our headphones is that they detect natural movements, such as the head or hand moving when users are listening to music and they can focus on a particular musical source that they want to hear. By moving their head left or right, listeners can hear the source from a frontal position as the digital compass detects the change in the direction they are facing. By looking up or down, the tilt sensor will detect the change in the face's angle of elevation; they can better hear the source that is allocated to a more distant or closer position. By putting their hand behind their ear, listeners can adjust the focus sensor on the headphones to focus on a particular musical source that they want to hear.

Keywords: Headphones, music interface, digital compass, tilt sensor, infrared distance sensor.

1. Introduction

Although we have recently been able to download a huge number of songs through Internet music delivery services, users are only listening to a small number because opportunities to find unfamiliar musical pieces in the collection are limited. Our goal was to construct a system that would enable people to easily select musical sources that they had an affinity for from many unknown ones.

Previous music retrieval methods that use queries such as similarity-based [1-3] searching, text-based searching [4], or collaborative filtering based searching [5, 6] are useful for narrowing the number of musical pieces, but after the list of rankings has been provided we have to listen to songs one by one because no consideration has been given to finding songs one has an affinity for from the list.

Musicream [7], on the other hand, make it possible to interact with many music collections by applying operations and providing functions for the order of play. Papipuun [8] and SmartMusicKIOSK [9] provide a music summary and allow quick listening in a manner similar to a stylus skipping on a scratched record. All these systems [7-9] enable us to save time by previewing songs from a list of rankings acquired from the results of music retrieval. However, these systems also force us to listen to songs one by one and involve many mouse operations.

In contrast, our system, called Music Scope Headphones, make it possible to select a musical source from the many available without the need for mouse clicks or other visual manipulations by detecting natural movements when users are listening to music and focusing on the particular musical source that they want to hear. The Music Scope Headphones provide a novel music selection interface that enables the following three functions to be applied.

1. *Scoping function*: enable us to scope many musical sources allocated in 2-dimensional space by moving our heads left or right or by looking up or down. The function enables us to landscape songs and save time in previewing them.
2. *Focusing function*: highlights a particular musical source that users want to hear by them placing their hand behind an ear. This function enables us to narrow the area in which sources are audible in 2-dimensional space as if controlling the directivity of a microphone.
3. *Switching function*: seamlessly changes musical sources in 2-dimensional space through users' gestures such as them nodding or shaking their heads, turning them around, or leaning them to one side. For example, when users are turning their head around, the next 10 musical sources in the order on the list acquired from a music retrieval system will be allocated in 2-dimensional space.

We mounted three sensors to the headphones, i.e., a digital compass, a tilt sensor, and a focus sensor, which detect natural movements, such as that of the head or the placement of a hand behind an ear, and this allowed us to use these three functions without the need for a display or a computer mouse. Users are freed from mouse operations and can select music much more actively.

Previously reported headphones with sensors to detect the direction users were facing or the location of the head could improve the sense of musical presence and create a realistic impression, but could not highlight parts according to their wishes [10-12]. It was difficult to clearly hear a particular musical source from many other sources with these headphones, including some that users may have preferred not to hear. There are music spatialization systems [13, 14] that allow users to control the localization of each part in real time through a graphical interface. However, it is difficult to control each musical source's location through this interface.

This paper is organized as follows. Section 2 explains the functions of the headphones. Section 3 describes system processing, and Section 4 discusses the implementation. Sections 6 and 7 present the experimental results and the conclusion.

2. Music Scope Headphones

We constructed the Music Scope Headphones that enabled us to save time in previewing songs based on the following three policies.

Reduced mouse operations When selecting music with a computer, we generally have to play songs individually with many mouse operations and these interrupts break the process of listening just as if a telephone were ringing. To solve this, we propose operations without mouse clicks achieved by detecting natural movements when listening to music and using these to control the computer.

Easy preview of many songs We wanted to increase the number of opportunities for encountering unfamiliar musical pieces in collections. However, the number of songs that can be previewed is limited within a fixed amount of time. This is because the larger the number of songs to be previewed, the shorter the time to listen to each song. To solve this, we propose a novel way of selecting music by playing many musical sources at the same time.

No computer display We wanted the system to be used anywhere and at any time such as at work or when riding or walking without the need to see a computer display. We attempted to construct a system to investigate whether it were possible to select and manipulate songs without a display.

The Music Scope Headphones let users control an audio mixer through natural movements, and thus enable them to select a musical source that they want to listen to from numerous sound sources. We will now explain the problems and solutions we encountered with the Music Scope Headphones based on these policies.

2.1 How musical sources are scoped

A particular musical source that the user temporarily wants to hear must be differentiated from other musical sources to scope it. We automatically adjusted each musical source's volume and panpot so that it could be distinguished from other musical sources. That is, we increased that source's volume and turned its panpot to the center, while decreasing the volume of the other sources and turning their panpots left or right. In this way, a user can easily scope a particular musical source.

2.2 How motion is detected

Natural movements must be detected while the user is listening to music to control the audio mixer through them. To enable this, we mounted the digital compass and the tilt sensor on top of the headband to detect the direction the user was facing and to detect the face's angle of elevation. We also mounted the focus sensor on the outside of the right speaker to detect the distance from the hand to the ear (Figure 1). We prepared three focus sensor prototypes and evaluated how practical they were in an experiment.

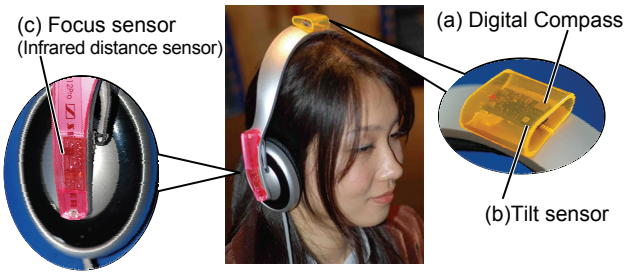

Figure 1. Three sensors mounted to headphone.

2.3 How function and motion are linked

How usable the Music Scope Headphones are depends on the quality of the links between the functions and the users' natural movements while they are listening to music. Let us imagine the following scenario.

- We receive several session recordings from a childhood friend.
- It sounds like the friend is playing a saxophone on one of these recordings.

In such a case, we would ordinarily search for songs with a saxophone part, and we then might want to hear the saxophone playing more clearly. We used the three links that follow to achieve this.

Link for scoping function When users move their head left (right), the musical source normally heard from the left (right) side can be heard from the frontal position as the digital compass detects the change in the direction they are facing. This allows users, through natural movements, to scope the musical source they want to hear most clearly and hear it from the front. When there are several musical sources at the front, users might not be able to hear the desired source clearly even after turning their head left or right to hear it from the front. In such a case, they can change the mix by moving their head up or down; the tilt sensor will detect the change in the face's angle of elevation. By looking up or down, users can increase the volume of sources so that instruments appear farther away or nearer. Here, we changed each source's position in 2-dimensional space, as can be seen in the graphical user interface in Figure 2. The circle at the center indicates the

position of the user's avatar and his/her head direction, and the circled numbers around the avatar indicate the positions of the sources. We also had several preset allocations for musical sources and these were easy to change by putting one's head to one side (Figure 3).

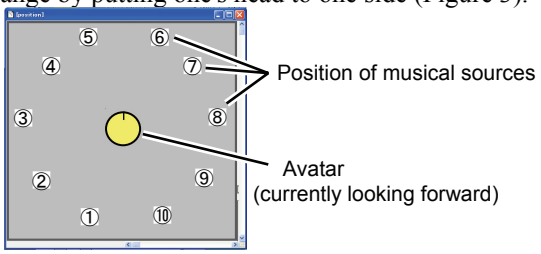

Figure 2. GUI for locating positions of parts.

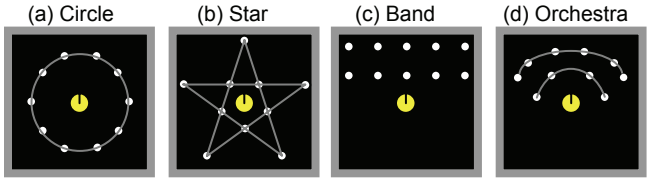

Figure 3. Presets for allocation.

Link for focusing function The focus sensor is used to detect the motion of users putting their hand behind an ear while they are listening to sound coming from the frontal position. The distance between the hand and ear determines the area in which sources are audible. For example, when users place their hand close to their ear, they can only hear the sources from the frontal position. When they removes their hand, they can hear all the sources except those behind them. When they put their hand in the middle position, they can hear the sources located in the front half position. By adjusting the distance between their hand and ear in this way, they can control the focus level and highlight the source of interest.

Link for switching function The system has two modes, a *song selecting mode* and a *part scoping mode*. We can scope and preview 10 songs in the song selecting mode from those listed in order by a music retrieval system allocated in 2-dimensional space. When converging on several songs using the focusing function in the song selecting mode, users can leave focused songs and delete unfocused songs by nodding their head (Figure 4(a)). If they want more convergence, they only need to adjust the focus level and nod their head again. Conversely, users can defocus by shaking their head and return to the previous scenario (Figure 4 (b)). When they only select one song and have a sound source where the tracks for each part have been recorded separately, the system changes to the part scoping mode. The Music Scope Headphones provide novel entertainment with this mode through which users can "scope" onto the part they want to hear more clearly. They can return to the song selection mode by shaking their head. Users can change the preset allocation by putting their head to one side (Figure 4 (c)) during the song selecting mode or the part scoping mode. By turning their head around in the song selection mode, the songs allocated in 2-dimensional space change to the next 10 songs from the list (Figure 4 (d)).

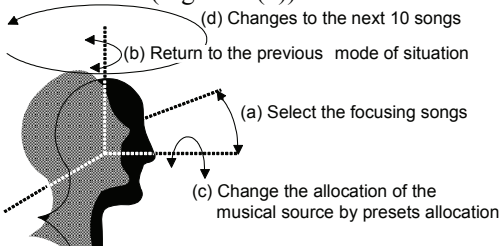

Figure 4. Link for switching function.

3. Processing

This section describes the processing flow for the system. We mainly describe sound processing and have omitted explanations for detecting gestures, nodding, shaking, putting the head to one side, and turning it around because of word limitations. In the following, we use θ ($-\pi \le \theta < \pi$) as the facing direction detected by the digital compass, ϕ ($-\pi \le \phi < \pi$) as the face's angle of elevation detected by the tilt sensor, and δ ($0 \le \delta \le 1$) as the distance between the hand and the ear detected by the focus sensor (Figure 5). We use radians as angle units and set the starting direction and angle of elevation to zero. We normalized δ from 0 to 1, and the focus sensor could detect a distance from 0 to 3 cm. When the distance was 0 cm, δ was output as 0, and when the distance was 3 cm, δ was output as 1. When the distance was between 0 and 3 cm, δ ranged from 0 to 1.

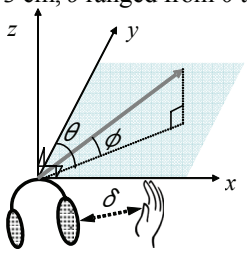

Figure 5. Three sensors mounted to headphone.

Pretreatment We prepared sound source S_n by recording a separate track for each part and allocating a position on the graphical user interface to each part (Figure 2). Here, l_n ($0 \le l_n \le 1$) indicates the distance from the avatar to each part and θ_n indicates the direction of each part. We normalized l_n so that the most distant part would have a value of 1.

Step 1 $h_n{}^{\phi}(0 \le h_n{}^{\phi} \le 1)$ was calculated as the amplification rate for each part, n , which changes depending on the angle of elevation, ϕ . We used the following formula so that when users looked up (down), the volumes of parts located far from (near to) their position would increase.

$$h_n^{\phi} = \begin{cases} 0 & \tilde{h}_n^{\phi} < 0 \\ h_n^{\phi} & 0 \le \tilde{h}_n^{\phi} < 1 \\ 1 & 1 < \tilde{h}_n^{\phi} \end{cases}, \quad (1)$$

where

$$\tilde{h}_n^{\phi} = 1 + l_n \sin\phi - \frac{1}{m}\sum_m l_m \sin\phi$$

m: number of parts .

When we allocated the positions for all parts as in Figure 6 (a), the mixing console was as in Figure 6 (b) when ϕ was zero. When ϕ was negative, the mixing console was as in Figure 6 (c), indicating that the volume of parts located near to (far from) the avatar was increased (decreased).

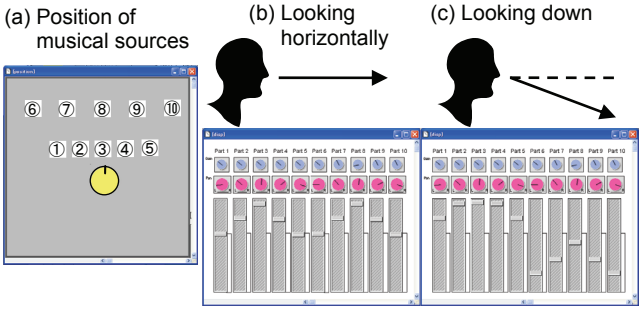

Figure 6. Angle of elevation ϕ and mixing console.

Step 2 h_n^{δ} was calculated as the amplification rate for all parts n, which changes according to the distance between the hand and ear δ. Here, $|a|$ indicates the absolute value of a, and θ_n' ($-\pi \le \theta_n' < \pi$) indicates the angle between θ_n and θ.

$$h_n^{\delta} = \begin{cases} 1 & \pi \cdot \delta \ge \left|\theta_n'\right| \\ 0 & \pi \cdot \delta < \left|\theta_n'\right| \end{cases} \quad (2)$$

For example, $h_n^{\delta} = 0$ corresponds to the parts located behind the user and $h_n^{\delta} = 1$ corresponds to the parts in front of the user when $\theta = \pi/3$ and $\delta = 0.5$ (Figure 7). In this way, we can eliminate parts the user does not want to hear.

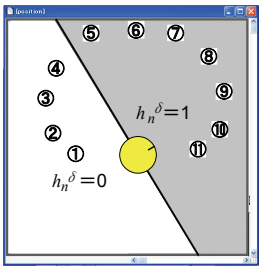

Figure 7. Distance from hand to ear δ and h_n^{δ}.

Step 3 h_n^{θ} ($0 \le h_n^{\theta} \le 1$) is calculated as the amplification rate for all parts n, which changes according to the direction. The h_n^{θ} output has a large value when the part is located in front of the user and becomes smaller when the part is located in another direction.

$$h_n^{\theta} = \begin{cases} 0 & \tilde{h}_n^{\theta} < 0 \\ \tilde{h}_n^{\theta} & 0 \le \tilde{h}_n^{\theta}, \end{cases} \quad (3)$$

where

$$h_n^{\theta} = \begin{cases} 0 & \delta = 0 \\ 1 - \dfrac{\alpha \cdot \left|\theta_n'\right|}{\pi \cdot \delta} & \delta > 0. \end{cases}$$

When we allocated the positions of all parts as in Figure 2, the mixing console was as in Figure 8(a) when users were looking left, as in Figure 8(b) when they were looking straight ahead, and as in Figure 8(c) when they were looking right.

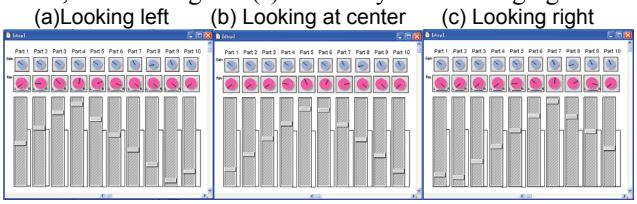

Figure 8. Direction θ and mixing console.

We used an adjustable parameter, α ($0 \le \alpha < 1$), to decrease the amplification rate when users placed their hand on their ear and $\delta < 1$. When we allocated positions for all parts as in Figure 2, the mixing console was as in Figure 9 (a) when users moved their hand away from their ear, and as in Figure 9 (b) when they moved their hand toward their ear.

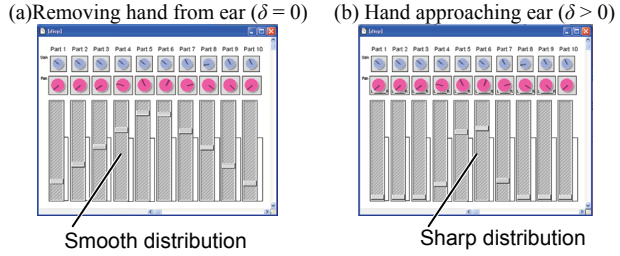

Figure 9. Decreasing amplification rate while $\alpha > 0$.

Step 4 p_n ($0 \le p_n < 1$)is calculated as the left/right volume ratio depending on direction θ. Here, $p_n = 0$ indicates that the ratio is 0:1 and $p_n = 0.5$ indicates that it is 1:1. We used an adjustable parameter, β, to change the left/right ratio when the users put their hand to their ear and $\delta < 1$. When $\beta > 0$ and $\delta < 1$, the panpots of the parts move to the back except for the part in the frontal position, and users can hear music as if focusing on the front part.

$$p_n = \frac{1}{2} + \frac{\beta \cdot \theta_n'}{\pi \cdot \delta} \quad (4)$$

Step 5 The amplification rates acquired in Steps 1 to 4 are multiplied and then the sound is output by summing up the sounds of all parts.

Right-side output:

$$S_{Right} = \sum_n S_n \cdot h_n^{\phi} \cdot h_n^{\delta} \cdot h_n^{\theta} \cdot p_n \quad \text{and} \quad (5)$$

Left-side output:

$$S_{Left} = \sum_n S_n \cdot h_n^{\phi} \cdot h_n^{\delta} \cdot h_n^{\theta} \cdot (1 - p_n) \,. \quad (6)$$

4. Implementation

We presented the processing flow for the software in the previous section. It worked on the Max/MSP [16]. Here, we describe the implementation of the hardware. We implemented the headphones with the two policies that follow so that everyone can easily use them.

- Lightweight yet strong.
- Easily connected to computer.

Headphones We selected headphones (Zenhizer: HD212Pro) that had adjusters inside the headband, because we could mount the focus sensor outside the right headband and this would therefore work stably even if the speakers were moved (Figure 1).

Sensors We mounted the attitude detection module (Aichi Micro Intelligent: AMI302-ATD) to the top of the headband, which consisted of the digital compass (MI sensor) and tilt sensor (Figure 1 (a), and (b)). Generally, the larger the angles of elevation, the larger the margin for error in the digital compass, because it detect the direction of the magnetic line of force. The main advantage of using the module was that the tilt sensor could correct the output of the digital compass. The detection resolution for the module was 2 degrees.

We also mounted the focus sensor to the right of the headband. We compared three focus-sensor prototypes in the experiments with musical novices, which are described below, and we selected the infrared distance sensor (Sharp: GP2S40J) (Figure 1 (c)). The infrared distance sensor consists of illuminant and acceptance of infrared and measures the distance between the sensor to objects by accepting the reflecting infrared. We prepared a circuit for mounting the infrared distance sensor and we mounted a semi-variable resistor so that the sensor could detect from 0 to 3 cm.

Protectors We made protectors for the sensor out of acrylic resin (Figure 1). We tested several colors for the resin and selected a light pink because this was affected least by sunlight from the windows and it widened the detection range of the sensor.

Circuit We integrated the information from the sensors by using a microcomputer (Renesas: R8C/15) mounted inside the headphone speaker housing and it output a serial signal. We could therefore reduce the number of cores in the cable from the headphone to the computer. The microcomputer sent a signal with output information from the sensors every 120 ms.

USB conversion We used a USB converter (Silicon Labs.: CP2102), which converted the serial signal to USB. It was mounted in the middle of the cable from the headphones to enable easy connection to the computer. We mounted LEDs on all sensors to indicate whether they were connected to the computer. If there was a connection they blinked quickly and if there was no connection they blinked slowly and we had to re-connect the USB cable.

Power supply All the sensors and the microcomputer worked on the bus current of the USB, which simplified the connection of the headphones. All we needed were the headphones and the computer.

5. Experimental Results

We designed the Music Scope headphones to enable not only a particular song to be selected from an ordered list acquired from music retrievals but also to highlight a particular instrument in the selected song that a user may want to hear more clearly. The system allows both audio files and MIDI files. In the experiments, we used RWC music database, which contains raw audio data before mix-down [15].

5.1 Evaluation of usability of focus sensors

Here, we discuss our evaluation of how usable the three focus-sensor prototypes were. They were (a) a variable resistor, (b) a bend sensor on a plastic lever, and (c) an infrared distance sensor (Figure 10). All headphones sets used the same digital compass and tilt sensor. We asked three musical novices to find a particular instrument, which we specified randomly, while listening to a song using the headphones. We used the song RWC-MDB-J-2001 No. 38 [15], which was played by 10 instruments located around the avatar as in Figure 2. The subjects already knew the sound of each instrument and were allowed to use all headphones several times before the experiment to familiarize themselves with their operation. The adjustable parameters α and β described in Section 3 were tuned by the subjects as they wanted. The following describes one trial of the experiment.

(1) Before the song was started we specified an instrument to subjects.
(2) We started the song at a midpoint randomly selected for the specified instrument.
(3) We measured the time the subjects needed to find the instrument.

The location of all instruments were randomly changed at every trial. The musical novices changed their headphones after every 10 trials.

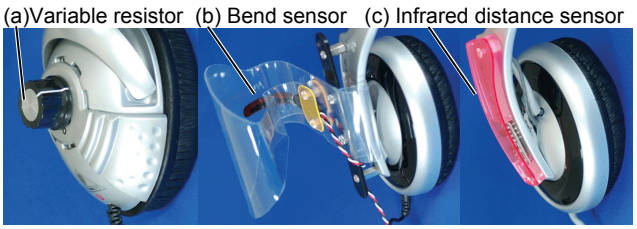

Figure 10. Three types of focus sensors .

Table 1 lists the average results from 100 trials. While the bend sensor was no less accurate than the variable resistor or the infrared sensor, it was attached to a plastic lever, which made it difficult to precisely control. Subjects A and C could find an instrument more quickly when using the infrared focus sensor. Subject B, on the other hand, could find an instrument more quickly when using the variable resistor. We selected the infrared distance sensor because the average time for the three subjects was the shortest.

Table 1. Comparison of three kinds of focus sensors.

	Variable resistance	Bend sensor	Infrared distance sensor
Subject A	1.84 sec.	1.28 sec.	1.12 sec.
Subject B	0.72 sec.	1.04 sec.	0.84 sec.
Subject C	1.02 sec.	2.01 sec.	0.74 sec.
Average	1.19 sec.	1.44 sec.	0.90 sec.

5.2 Evaluations of usability for selecting songs

We evaluated whether users could select a song by using the Music Scope Headphones. We asked three musical novices to find a song with a soprano saxophone from the RWC-MDB-J-2001 database [15]. It had fifty jazz songs and only one song had a soprano saxophone part. We measured the time the subject needed to find the soprano saxophone. The subjects had not heard the songs on the database before the experiment except for RWC-MDB-J-2001 No. 38, which we had used in the experiment in Section 5.1. After measuring the time using the Music Scope Headphones, we measured the time for same trial using Windows Mediaplayer, which is a standard music player pre-installed in Windows XP. The time for each subject to find the musical instrument was only measured one for the Music Scope Headphones and Windows Mediaplayer.

Table 2 lists the results obtained with the Musical Scope Headphones and Windows Mediaplayer. All the subjects could find the song more quickly when using our Music Scope Headphones, but Windows Mediaplayer was handicapped because the subjects may have memorized the songs in the first trial with the Music Scope Headphones. As a result, our experiment revealed that the Music Scope Headphones were superior for previewing songs from an ordered list.

Table 2. Comparison of our system and standard music player.

	Music Scope Headphones	Windows Media Player
Subject A	224 sec.	845 sec.
Subject B	423 sec.	1145 sec.
Subject C	642 sec.	751 sec.
Average	429 sec.	914 sec.

6. Conclusion

The Music Scope Headphones enabled wearers to control an audio mixer through natural movements that enabled them not only to select a song from an ordered list acquired from music retrievals but also to highlight a particular instrument in the selected song that they wanted to hear more clearly. Three sensors were mounted to the headphones: a digital compass, a tilt sensor, and a focus sensor for detecting natural movements. This freed users from mouse operations so they could select music much more actively. We tested how usable three kinds of focus sensors were and found that an infrared distance sensor was better than either a variable resistor or a bend sensor from the average time it took three subjects to locate an instrument. We also tested how efficiently the headphones were in selecting songs and the results revealed that they performed better than the standard Windows Mediaplayer by being able to select a particular song from fifty others.

We are now developing other applications for the headphones. Figure 11 shows where the light's brightness has been controlled according to the sound level at the music stands. This allows the user to experience all sound levels visually as well as aurally. This should help musical novices who do not know what individual instruments sound like to learn the relationship between these and the entire piece. The video is available at http://staff.aist.go.jp/m.hamanaka/video/.

We plan to use these headphones with music retrieval based on voice recognition to construct a system in which a display and a mouse are unnecessary.

Figure 11. Lighting depending on sound levels at music stands.

References

[1] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. IEEE Trans. on Speech and Audio Proc., 10(5): 293− 302, 2002.

[2] F. Vignoli and S. Pauws. A music retrieval system based on user-driven similarity and its evaluation. In Proc. of ISMIR 2005, pp. 272–279, 2005.

[3] E. Pampalk. A MATLAB toolbox to compute music similarity from audio. In Proc. of ISMIR2004, pp. 254–257, 2004.

[4] T. Soding and A. F. Smeaton. Evaluating a music information retrieval system - TREC style. In Proc. of ISMIR2002, pp. 71–78, 2002.

[5] W. W. Cohen and W. Fan, "Web-collaborative filtering: Recommending music by crawling the Web." WWW9/Computer Networks, 33 (1-6): 685− 698, 2000.

[6] A. Uitdenbogerd and R. van Schyndel. A review of factors affecting music recommender success. In Proc. ISMIR2002, pp. 204–208, 2002.

[7] M. Goto and T. Goto. Musicream: New Music Playback Interface for Streaming, Sticking, Sorting, and Recalling Musical Pieces, In Proc. of ISMIR 2005, pp. 404–411, 2005.

[8] K. Hirata and S. Matsuda. Interactive Music Summarization Based on GTTM. In Proc. of ISMIR 2002, pp. 86–93, 2002.

[9] M. Goto: SmartMusicKIOSK: Music Listening Station with Chorus-search Function, In Proc. of UIST 2003, pp. 31–40, 2003.

[10] Warusfel, O. and Eckel, G. LISTEN - Augmenting Everyday Environments through Interactive Soundscapes. In Proc. IEEE VR2004, pp. 268–275, 2004.

[11] Wu, J., Duh, C., Ouhyoung, M., and Wu, J. 1997. Head Motion and Latency Compensation on Localization of 3D Sound in Virtual Reality. In Proc. ACM VRCIA1997, pp. 15–20, 1997.

[12] Goudeseune, C., and Kaczmarski, H.. Composing Outdoor Augmented-reality Sound Environments. In Proc. of ICMC2001, pp. 83–86, 2001

[13] Pachet, F. and Delerue, O. A Mixed 2D/3D Interface for Music Spatialization. In Proc. of ICVW1998, pp. 298–307, 1998.

[14] Pachet, F. and Delerue, O. On-the-Fly Multi-track Mixing.. In Proc. of AES2000, 2000.

[15] Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. RWC Music Database: Popular, Classical, and Jazz Music Databases. In Proc. of ISMIR2002, pp. 287–288, 2002.

[16] cycling74. http://www.cycling74.com/products/maxmsp/, 2006.