Sound Scope Headphones: Controlling an Audio Mixer through Natural Movement

Masatoshi Hamanaka [*], Seunghee Lee[†]

[*]PRESTO, Japan Science and Technology Agency
m.hamanaka@aist.go.jp
[†]University of Tsukuba Graduate School of Comprehensive Human Science
lee@kansei.tsukuba.ac.jp

Abstract

This paper describes the Sound Scope Headphones which enable the user to control an audio mixer through natural movements of the head and hands. This allows musical novices to interact closely with the music they listen to. Commercial audio mixers are too complicated for musical novices to properly control the multi-channel volumes and panpots. Our headphone device controls an audio mixer using three sensors mounted on the headphones to detect movements of the user when listening to music.

1 Introduction

We have been developing a jam session system, called the Guitarist Simulator, that allows a human guitarist to play interactively with virtual guitarists, each imitating the musical personality of a human guitarist (Hamanaka et al. 2003a), (Hamanaka et al. 2003b). Although the Guitarist Simulator was initially intended as a system for musical experiments, we are now extending the system so that it can be used by musical novices. When musical novices use the system, though, they find it difficult to separate each player's performance if they want to clearly hear a particular part. Thus, it is hard for musical novices to recognize the differences between the virtual guitarist personalities even if the system can imitate the musical personalities of actual human guitarists.

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

A music spatialization system (Pachet, and Delerue 1998), (Pachet, and Delerue 2000) allows a user to control the localization of each part in real time through a graphical interface. However, these systems suffer from the same problem as a commercial audio mixer because it is difficult for a musical novice to appropriately change each part location through a graphical interface the moment a solo part begins.

We designed the Sound Scope Headphones so that they would let users control an audio mixer through natural movements, and thus enable a musical novice to separately listen to each player's performance. The main advantage of the headphones is that they detect natural movement, such as head movement or placing a hand behind an ear, and uses the detected movements to control an audio mixer while the user listens to music. Three sensors are mounted on the headphones: a digital compass, a tilt sensor, and a distance sensor.

Previously reported headphones with sensors to detect the direction the user is facing or the location of the head can escalate the musical presence and create a realistic impression, but do not control the volumes and panpots of each part according to the user's wishes (Warusfel, and Eckel 2004), (Wu et al. 1997), (Goudeseune, and Kaczmarski 2001), (Sato 2004). With these headphones, it is difficult to clearly hear a particular part from among many other parts, including some that the user would prefer not to hear.

In contrast, our headphones let a user listening to music scope a particular part that he or she wants to hear. For example, when listening to jazz, one might want to clearly hear the guitar or reduce the sound of the sax. By moving your head left or right, you can hear from a frontal position. By looking up or down, you can better hear the parts allocated to a more distant or a closer position. By simply putting your hand behind your ear, you can adjust the distance sensor on the headphones and focus on a particular part you want to hear.

2 Sound Scope Headphones

Here we explain what enables our headphones to scope a part which the user wants to hear.

2.1 How parts are scoped

To scope a particular part, the part which the user wants to temporarily scope must be differentiated from the other parts. We adjust each part's volume and panpot so that it can be distinguished from the other parts. That is, we increase that part's volume and turn its panpot to the center, while

decreasing the volume of other players and turning their panpots left or right. In this way, a user who is a musical novice can easily scope a particular part.

To control each part's volume and panpot, we have to prepare a sound source where the track of each part is recorded separately. For this, we used the RWC music database (RWC-MDB-J-2001 No.38), which contains raw data before mix-down (Goto et al. 2002).

2.2 How motion is detected

To control an audio mixer through natural movements, the movements must be detected while the user is listening to music. On top of the headphone arc we mounted a digital compass to detect the direction the user was facing and a tilt sensor to detect the face's angle of elevation. On the outside of the right speaker we mounted a distance sensor to detect the distance from hand to ear (Figure 1). Originally, we used a plastic lever with a bend sensor as the distance sensor, but later changed this to an infrared distance sensor. We evaluated which is more practical in the experiment discussed below.

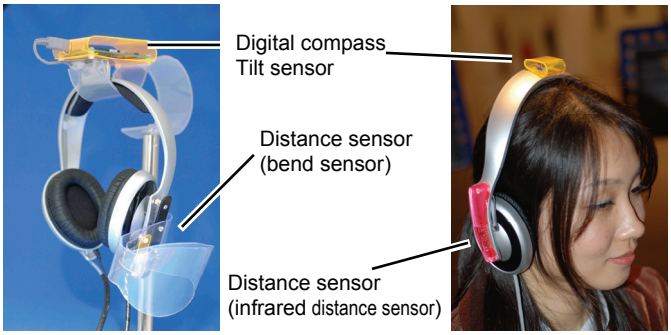


Figure 1. Three sensors mounted on the headphone.

2.3 How the mixer and motion is linked

The usability of our headphones depends on the quality of the links between the mixer manipulations and the natural movement while the user listens to music. We use three links.

Link from the facing direction. When a user moves his head leftwards (rightwards), the part normally heard from the left (right) side can be heard from a frontal position as the digital compass detects the change in the direction the user faces. This allows a user, through natural movement, to scope the part which he wants to hear most clearly and hear it from a frontal position.

Link from the face's angle of elevation. When there are several parts in the frontal position, the user might not be able to hear the desired part clearly after turning his head left or right to hear it from a frontal position. In such a case, the user can change the mix by moving his head up or down and the tilt sensor will detect the change in the face's angle of elevation. By looking up or down, respectively, the user increases the volume of each part located farther away or more closely. Here, we can change each part's position as shown in the graphical user interface in Figure 2. The circle at the center indicates the position of the avatar and its head direction, and the circle numbers around the avatar indicate the positions of the parts.

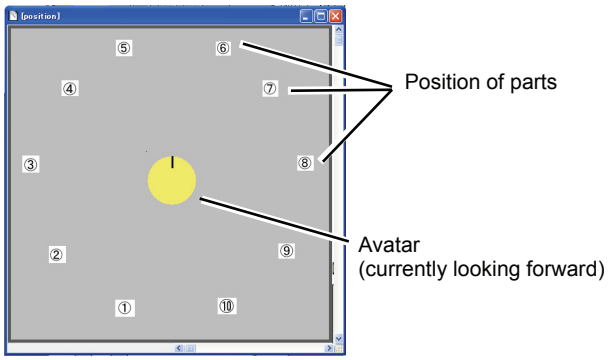


Figure 2. GUI for locating the position of each part.

Link from the distance between hand and ear. The distance sensor detects the motion of putting a hand behind one's ear while listening to the sound from a frontal position. The distance between hand and ear determines the area indicating whether each part is audible. For example, when a user places her hand close to her ear, she can hear only the parts from a frontal position. When the user removes her hand, she can hear all the parts except those behind her. When the user puts her hand in a middle position, she can hear the parts located in the front half position. Therefore, the user can focus on a part she wants to listen to by adjusting the distance between her hand and ear.

3 Implementation

In this section we describe the processing flow of the system. In the following explanation, we use θ ($-\pi \leq \theta < \pi$) as the facing direction detected by the digital compass, φ ($-\pi \leq \varphi < \pi$) as the face's angle of elevation detected by the tilt sensor, and δ ($0 \leq \delta \leq 1$) as the distance between the hand and the ear detected by the distance sensor (Figure 3). We use radians as the angle units and set the starting direction and angle of elevation as zero. We normalized δ from 0 to 1, and the distance 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.

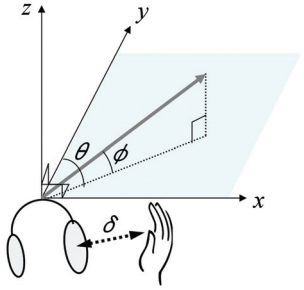


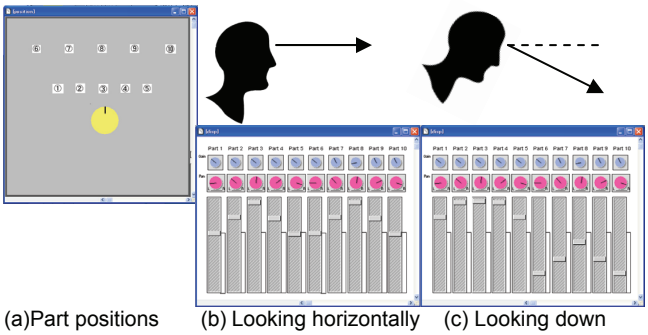
Figure 3. Direction θ, angle of elevation φ, and distance δ.

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

$$h_n^{\phi} = 1 + l_n \sin\phi - \frac{1}{m}\sum_m l_m \sin\phi \qquad (1)$$

When we allocated each part position as shown in Figure 4(a), the mixing console was as shown in Figure 4(b) when φ was zero. When φ was a negative value, the mixing console was as shown in Figure 4(c), indicating that the volume of parts located near to the avatar was increased.



(a)Part positions (b) Looking horizontally (c) Looking down

Figure 4. Angle of elevation φ and the mixing console.

Step 2. $h_n^{\delta}(0 \le h_n^{\delta} \le 1)$ was calculated as the amplification rate for each part n which changes depending on the distance between 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.

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

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 5). In this way, we can eliminate parts the user does not want to hear.

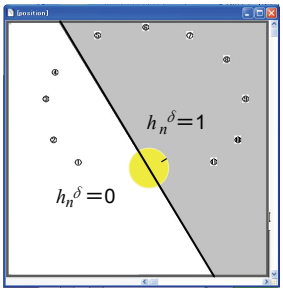
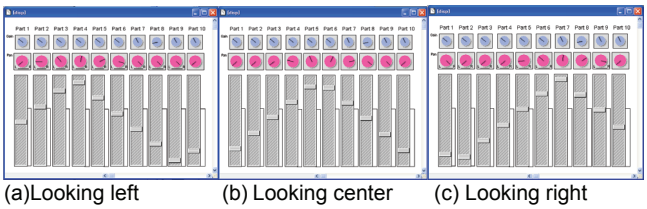


Figure 5. Distance from hand to ear δ and h_n^{δ}.

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

$$h_n^{\theta} = 1 - \frac{\alpha \cdot \left|\theta_n'\right|}{\pi \cdot \delta} \qquad (3)$$

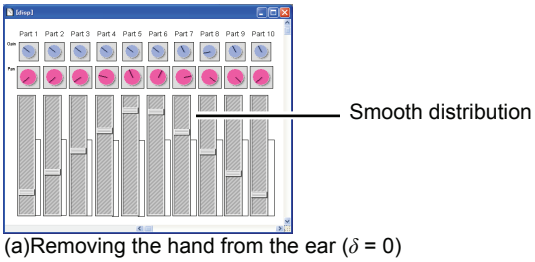
When we allocated each part position as shown in Figure 2, the mixing console was as shown in Figure 6(a) when the user was looking left, as shown in Figure 6(b) when the user was looking straight ahead, and as shown in Figure 6(c) when the user was looking right.



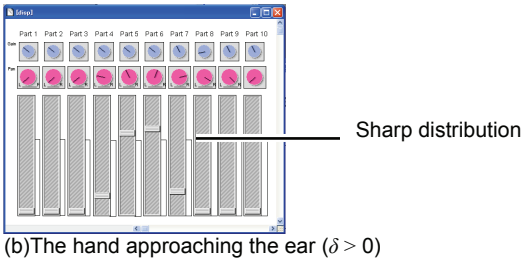
(a)Looking left (b) Looking center (c) Looking right

Figure 6. Direction θ and the mixing console.

We used an adjustable parameter, α $(0 \le \alpha < 1)$, to decrease the amplification rate when the user puts his hand to his ear and $\delta < 1$. When we allocate each part position as shown in Figure 2, the mixing console was as shown in Figure 7(a) when the user moved his hand away from his ear, and as shown in Figure 7(b) when the user moved his hand towards his ear.



(a)Removing the hand from the ear (δ = 0)



(b)The hand approaching the ear (δ > 0)

Figure 7. Decreasing the 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 the ratio is 0:1 and $p_n = 0.5$ indicates the ratio is 1:1. We use an adjustable parameter, β, to change the left/right ratio when the user puts his hand to his ear and $\delta < 1$. When $\beta > 0$ and $\delta < 1$, the panpots of the parts move to

the back except for the part at the frontal position and the user can hear the 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:

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

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 Experimental results

We evaluated the usability of the two types of distance sensor: the bend sensor and the infrared distance sensor. Both headphone sets used the same digital compass and tilt sensor. We asked two musical novices to find a particular instrument, which we specified randomly, while listening to a song using the headphones. The song was played by ten instruments and each musical novice already knew the sound of each instrument. The subjects were allowed to use both types of headphones several times before the experiment to familiarize themselves with the operation. We measured the time the musical novice needed to find the instrument after we specified an instrument.

Table 1 shows the average results from 100 trials. The musical novices changed the headphones after every 10 trials. Both subjects could find an instrument more quickly when using the infrared distance sensor. While the bend sensor was no less accurate than the infrared sensor, it was attached to a plastic lever which made it difficult to precisely control.

Table 1. Comparison of two kinds of distance sensor.

	bend sensor	infrared distance sensor
musical novice A	1.28 sec.	1.12 sec.
musical novice B	1.04 sec.	0.84 sec.

5 Conclusion

The Sound Scope Headphones enable the wearer to control an audio mixer through natural movements. Three sensors are mounted on the headphones: a digital compass, a tilt sensor, and a distance sensor. We tested two kinds of distance sensor and found that an infrared distance sensor was better than a bend sensor for detecting the distance from hand to ear. We are now developing applications for the headphones. For example, Figure 8 shows music stands where the light brightness is controlled depending on each part's sound-level. This allows the user to understand each part's sound-level visually as well as aurally. This should help musical novices who do not know the sound of each instrument learn the relationship between instrument and sound. We plan to use this headphone in a jam session system we are developing so that the user can select a favorite virtual player from among many choices.

Figure 8. Lighting depending on each part's sound-level.

References

Hamanaka, M., Goto, M., Asoh, H., and Otsu, N. 2003. A Learning-Based Jam Session System that Imitates a Player's Personality Model. In *Proceedings of the International Joint Conference on Artificial Intelligence* (IJCAI2003), pp. 51-58. Acapulco.

Hamanaka, M., Goto, M., Asoh, H., and Otsu, N. 2003. A Learning-Based quantization: Unsupervised Estimation of the Model Parameters. In *Proceedings of the International Computer Music Conference,* pp. 369–372. Singapore: International Computer Music Association.

Pachet, F., and Delerue, O. 1998. A Mixed 2D/3D Interface for Music Spatialization. In *Lecture Notes in Computer Science* (no. 1434) First International Conference on Virtual Worlds, pp. 298–307. Paris.

Pachet, F., and Delerue, O. 2000. On-The-Fly Multi-Track Mixing. In *Proceedings of AES 109th Convention,* Los Angeles: Audio Engineering Society.

Warusfel, O., and Eckel, G. 2004. LISTEN - Augmenting Eeveryday Environments Through Interactive Soundscapes. In *Proceedings of IEEE Workshop on VR for public consumption*, pp. 268–275. Chicago: IEEE Virtual Reality.

Wu, J., Duh, C., Ouhyoung, M., and Wu, J. 1997. Head Motion and Latency Compensation on Localization of 3D Sound in Virtual Reality. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pp. 15-20. Lausanne, Switzerland: ACM Virtual Reality Software and Technology.

Goudeseune, C., and Kaczmarski, H. 2001. Composing Outdoor Augmented-reality Sound Environments. In *Proceedings of the International Computer Music Conference,* pp. 83–86. Havana, Cuba: International Computer Music Association.

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

Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. 2002. RWC Music Database: Popular, Classical, and Jazz Music Databases. In *Proceedings of the International Conference on Music Information Retrieval*. pp. 287-288. Paris.