SOUND SCOPE PHONE: FOCUS ON A SPECIFIC PART USING FACE AND HAND TRACKING

Masatoshi Hamanaka
RIKEN
Masatoshi.hamanaka@riken.jp

ABSTRACT

This paper describes Sound Scope Phone, an application that enables you to emphasize the part you want to listen to in a song consisting of multiple parts by head direction or hand gestures. The previously proposed interface required special headphones equipped with a digital compass and distance sensor to detect the direction of the head and distance between the head and a hand, respectively. Sound Scope Phone integrates face tracking information on the basis of images from the front camera of a commercially available smartphone with information from the built-in acceleration/gyro sensor to detect the head direction. The built application is published on the Apple App store under the name SoundScopePhone.

1. INTRODUCTION

Sound Scope Phone is an application that enables users to focus on the sound of a specific part of a song more clearly while listening to it (Fig. 1). Turning their head to the left and right enables users to follow the sound of specific instruments on their respective sides. Moving their palms close to their ears in a "listening pose" enables users to focus on only specific parts.

We previously proposed Sound Scope Headphones that achieved the aforementioned functions using a digital compass that detects the orientation of the face and distance sensors that measure the distance between a hand and an ear [1, 2]. Fig. 2 shows the exhibition at ACM SIGGRAPH2009 Emerging Technologies. Since face-to-face exhibitions are temporarily difficult due to social conditions, we considered implementing similar functions in an app so that many people can experience it using a smartphone.

2. RELATED WORKS

Headphones equipped with sensors that detects the direction and position of the head have been proposed, but their purposes were to enhance the sense of presence by fixing the virtual sound source position [3–5]. Therefore, they have not been used for the purpose of emphasizing a specific part as in this study. For example, it was difficult to selectively listen to only the instruments that you want to hear near those that you did not. A spatial audio system enables to change the volume of an instrument by moving the position of the listener's avatar and each part [6, 7]. However, it was difficult for beginners to properly adjust the mixing of each part because it required a complicated operation to place the part where the solo started in the center and the part that returned to the accompaniment far away.

Figure 1. Sound Scope Phone

3. SOUND SCOPE PHONE

When listening to a song consisting of multiple parts, a new way of listening to music is achieved on a smartphone by searching for the part you want to listen to and listening while emphasizing it.

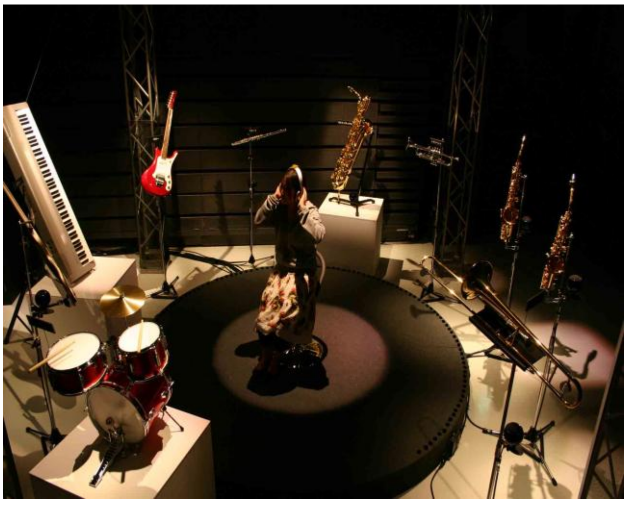

Figure 2. Exhibit of Sound Scope Headphone

Attribution: M. Hamanaka, “Sound scope phone: focus on a specific part using face and hand tracking”, in *Extended Abstracts for the Late-Breaking Demo Session of the 22nd Int. Society for Music Information Retrieval Conf.*, Online, 2021.

3.1 How Parts are Scoped

The part that the user wants to temporarily scope must be differentiated from the others. We adjust each part's volume and localization so that it can be distinguished from the others. That is, we increase that part's volume and turn its localization to the center, while decreasing the volume of other parts and turning their localization left or right. In this way, a user who is a musical novice can easily scope a particular part.

3.2 How Motion is Detected

With Sound Scope Headphones, the direction of the head is detected by mounting a digital compass on the arc of the headphones (Fig. 3). However, Sound Scope Phone detects the direction of the head by processing the image ac-quired by the front camera of a smartphone (upper part of Fig. 4).

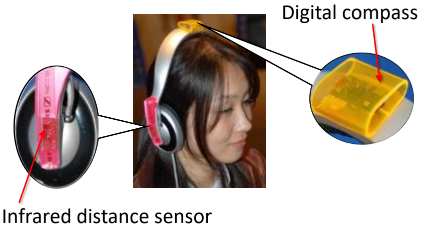

Figure 3. Sensors mounted on the headphones

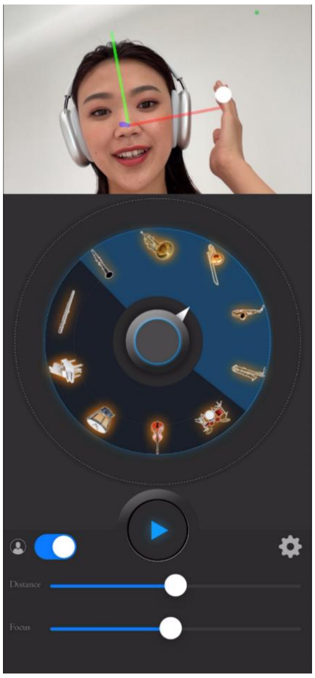

Figure 4. Screen shot of Sound Scope Phone

When detecting the direction of the head using a front camera, if it exceeds 90 degrees, more than half of the front side of the face becomes invisible, making detection difficult. Therefore, with Sound Scope Phone, by integrating the angle information acquired by the acceleration/gyro sensor mounted on the smartphone and the head angle information from the front camera, it is possible to detect the head direction over 360 degrees. In other words, when holding a smartphone in one's hand and looking back, it is difficult to look back with only the neck, and the upper body also moves in tandem. At this time, the hand holding the smartphone also moves in tandem to some extent, making it easier to capture the front of the head in the front camera. This makes it possible to detect the direction of the head even if the head is rotated 90 degrees or more.

With Sound Scope Headphones, a distance sensor is mounted on the outside of the right speaker to detect that a hand is approaching the ear. However, Sound Scope Phone detects that the hand is approaching the ear by performing image processing on the image from the front camera.

3.3 How to Emphasize

With previous sound scope headphones, the sound of the part you want to hear is emphasized in the front by associating the detected movement with the operation of the audio mixer. On the other hand, with the sound scope phone, by arranging *n* parts in the 3D space defined by the 3D sound library OpenAL [8], the sound of the parts arranged in the direction of the head can be heard in front (center of Fig. 4). At this time, let θ_n be the angle created from each of the *n* parts from the listener's avatar.

By changing the amplification rate $h_n{}^{\theta}$ ($0 \leq h_n{}^{\theta} \leq 1$) in accordance with the orientation of the head, the sound of the part at an angle close to the front can be heard to be more emphasized. In the following, the angle between the head direction and the position where the part is placed is θ ($- \pi \leq \theta < -\pi$), and the distance between the hand and ear is δ ($0 \leq \delta \leq 1$).

$$h_n{}^{\theta} = \begin{cases} 0 & \widetilde{h_n{}^{\theta}} < 0 \\ \widetilde{h_n{}^{\theta}} & 0 \geq \widetilde{h_n{}^{\theta}} \end{cases} \quad (1),$$

where

$$\widetilde{h_n{}^{\theta}} = \begin{cases} 0 & \delta = 0 \\ 1 - (\alpha \cdot |\acute{\theta}_n|)/(\pi \cdot \delta) & \delta > 0 \end{cases} (2).$$

α ($0 \leq \alpha \leq 1$) is an adjustable parameter that sets the change in amplification rate when $\delta < 1$. When $\alpha = 0$, the amplification rate of each part does not change even if the hand is closer to the ear, but when $\alpha > 0$, the amplification rate decreases as the hand approaches the ear. At this time, since the decrease in the amplification rate is larger in the direction in which the listener is not facing, the sound in front can be heard relatively loudly.

4. CONCLUSION

We described Sound Scope Phone, an application that enables you to emphasize the part you want to listen to in a song consisting of multiple parts by using head direction or hand gesture. You can access the download link of the Sound Scope Phone and the introductory video to understand how to use it from the following URL.

https://gttm.jp/hamanaka/en/soundscopephone/

If you use the app 30 times, the guide for the page requesting the questionnaire will be displayed. We would appreciate your cooperation.

We plan to build and release an iPad version that also displays the performance video of the performer.

5. REFERENCES

[1] M. Hamanaka and S. Lee, “Sound Scope Headphones,” *ACM Siggraph2009 Talks* TK-201 /*Emerging Technologies* ET-201, 2009.

[2] M. Hamanaka and S. Lee, “Music Scope Headphones: Natural User Interface for Selection of Music”, *in Proc. of the 7th Int. Conf. on Music Information Retrieval*, Victoria, Canada, 2006, pp.302–307.

[3] O. Warusfel and G. Eckel, “LISTEN - Augmenting Everyday Environments Through Interactive Soundscapes,” *in Proc. of IEEE Workshop on VR for public consumption*, IEEE Virtual Reality, Chicago, 2004, pp. 268–275.

[4] J. Wu, C. Duh, and M. Ouhyoung, “Head Motion and Latency Compensation on Localization of 3D Sound in Virtual Reality”, *in Proc. of the ACM symposium on Virtual reality software and technology*, ACM Virtual Reality Software and Technology, Lausanne, Switzerland, 1997, pp. 15–20.

[5] C. Goudeseune and H. Kaczmarski, “Composing Outdoor Augmented-reality Sound Environments”, *in Proc. of the Int. Computer Music Conf.*, International Computer Music Association, Havana, Cuba. 2001, pp. 83–86.

[6] F. Pachet and O. Delerue, “A Mixed 2D/3D Interface for Music Spatialization”, *in Lecture Notes in Computer Science (no. 1434) First Int. Conf. on Virtual Worlds*, Paris, France, 1998. pp. 298–307.

[7] F. Pachet and O. Delerue, “On-The-Fly Multi-Track Mixing”, *in Proceedings of AES 109th Convention*, Audio Engineering Society, Los Angeles, 2000.

[8] “OpenAL,” accessed 15 August 2021 https://www.openal.org/.