Development of shape-based average head-related transfer functions and their applications

Tsukasa Suenaga*, Shoken Kaneko† and Hiraku Okumura‡

Research & Development Division, Yamaha Corporation,
10–1 Nakazawa-cho, Naka-ku, Hamamatsu, 430–8650 Japan

Abstract: Three-dimensional (3D) audio reproduction systems incorporated into a set of headphones are growing in popularity alongside the evolution of virtual reality (VR)/augmented reality (AR) technology. In this paper, some applications of binaural 3D audio systems will be presented. Game audio is one of the application fields for binaural 3D audio systems. We have developed a binaural 3D audio system for game development. When applying the binaural 3D audio system to games, it is necessary to solve the problem of calculation load. Therefore, we developed a hybrid system of virtual loudspeakers and object sound sources. In addition, the problem of timbre change was highlighted. Meanwhile, 360° video is another potential application field for 3D audio. In conventional binaural systems, it is difficult to express ambient sound sources such as the sound of rustling leaves. Our system has made it possible to reproduce ambient sound sources by combining higher-order ambisonics (HoA) and head-related transfer functions (HRTFs).

Keywords: 3D audio, HRTFs, HoA, Headphones

PACS number: 43.60.Lq, 43.66.Jh [doi:10.1250/ast.41.282]

1. INTRODUCTION

Binaural three-dimensional (3D) audio systems are growing in popularity alongside the evolutions of VR/AR technology. Virtual auditory display (VAD) systems have become easier to implement as interactive systems because they can access motion track sensors mounted on a head-mounted display (HMD). In particular, video games and 360° video are the most prominent prospective applications in this field because of their strong interactivity with the player. However, from the viewpoint of content production in these fields, a 3D audio system is not easy to realize because of problems such as poor sound quality and large computational load.

Although 3D audio systems using head-related impulse responses (HRIRs) of dummy heads [1] are widely applied, they have poor quality of sound localization and the timbre effect. Owing to advances in 3D shape scanning, shape processing technology, and computational engineering, 3D human shape morphing and the numerical calculation of HRIRs [2] can be performed more accurately than before. By applying these techniques, we attempted to create original HRIRs based on an average human shape model.

In this paper, we report the following developments:
(1) Create of original shape based average HRIRs.
(2) Development of a 3D audio application for game audio
(3) Development of a 3D audio application for 360° video.

In the two applications described in this paper, the HRTFs and VAD system described in Sect. 3 are used.

However, in the case of application to game audio, it is important that the system has only a small increase in computational loads and small timbral differences compared with the case without 3D audio processing. In contrast, we consider that the combination of binaural audio processing with 3D sound field recording and reproduction technology is important in the application to 360° video.

In this article, we discuss some of the considerations that we made when adapting a VAD system to each application.

2. VAD SYSTEM

A 3D audio reproduction system using a set of headphones is based on VAD technology [3,4]. In this section, we provide an overview of VAD systems.

Human beings have the ability to perceive the directions of sound sources. VAD systems are required to provide the ears of a listener with appropriate sound
information to mimic real-world sound localization. It is known that HRIRs provide perceptual cues for sound localization [5]. Therefore, VAD systems are implemented as convolution systems with sound sources and HRIRs. HRIRs represent impulse responses transmitted from a sound source to the ears of a listener.

A simple VAD system model is shown in Fig. 1. The input of the VAD system consists of a single-channel (monophonic) audio signal and its localization information. According to the relative localization information between the sound source and the listener, the VAD system selects the appropriate HRIR pair from the database. The VAD system convolves the input signal with the HRIR of the left and right ears. Finally, listeners can hear the sound using a set of headphones as if they were at the same position.

3. SHAPE-BASED AVERAGE HRIRs

A VAD system consists of a two-channel simple convolution system. Therefore, an important technology required for localization and good sound quality is a means of obtaining the HRIR coefficients.

It has been reported that the use of individualized HRIRs, i.e., the listener’s own responses, is the best approach for accurate sound localization because HRIRs differ among individuals.

Conversely, from a practical point of view, a system using the HRIRs of each listener is not practical to implement. A system that uses typical HRIRs suitable for many people would still be an important milestone. Hence, we adopted a method of calculating shape-based average HRIRs through computational analysis to obtain coefficients suitable for many people [6].

To calculate shape-based average HRIRs, we employed the following three steps:

1. Collection of 3D human shape data through 3D scanning and life casting.
2. Generation of average shape models from the 3D shape database using 3D shape morphing.
3. Calculation of shape-based average HRIRs using the boundary element method (BEM).

3.1. 3D Scanning of Human Shapes

Optical 3D scanners are suitable for human shape scanning because they do not affect the shapes of the ear, head, and body during measurement. However, the interior parts of the cymba conchae and cavum conchae are difficult to scan using 3D scanners because the scanners have a shadow area caused by the difference in the optical axis between the projector and the camera in the scanner. Conversely, life casting using silicone is an effective method of capturing the shapes of the cymba conchae and cavum conchae. However, the impression material affects the shape of the pinna because of its weight.

By taking advantage of the characteristics of these two methods, we measured human shapes by both 3D scanning and life casting and combined the obtained models into a single human shape model. An example of an ear shape obtained is shown in Fig. 2.

3.2. 3D Shape Processing

We have developed a 3D ear shape modeling method [6] to generate an average shape model for human ears. The scanned 3D shape data consist of point clouds and polygon surfaces. However, the shape data represented as a polygon soup are not registered. Different polygon number is assigned to the same part of the shape data for two different people.

To calculate the arithmetic mean of multiple shapes, a registration procedure is required as the first step in shape processing.

One ear is chosen from the database as the base ear, which defines the origin of the ear shape space. Nonrigid point cloud registration [7] was applied between the base ear and each ear of an individual. The point clouds associated with each ear were registered as values relative to the base ear. In other words, we obtained the warping vectors for each ear from the base ear.

As the second step, the warping vectors from the base ear to the average ear were generated by calculating the average of the obtained warping vectors.

By the same method, an average head model was generated and combined with the average ear model. The average shape model obtained from the ears of 30 subjects and the heads of 11 subjects is shown in Fig. 3.
3.3. Calculating HRTFs

Because of the ease of creating a mesh model and the availability of a high-speed calculation method [8], we chose the BEM for HRIR calculation.

The conditions for the calculation areas were as follows. The mesh size was set to 1.0 mm around the ears and 2.0 mm in other areas. Generally, the mesh size should be set below 1/6 of the wavelength corresponding to the maximum frequency in the calculation [9]. In this case, a mesh size of 2.0 mm is sufficient to perform calculations up to 24 kHz. However, there are some concave parts in human pinnae; hence, the mesh size of the ear area was set to 1.0 mm to maintain the volumetric capacity of the cymba conchae and cavum conchae. The average shape model had a total of 113,824 elements. Using reciprocal theory, a spherical sound source was set at the entrance of the earhole, and the sound pressure of the sound source was 1 Pa. The BEM solver calculated 256 frequencies (93.75 Hz steps from 93.75 to 24 kHz).

To validate the accuracy of the calculated HRIRs, we created a dummy head of an average shape using a 3D printer and measured its HRIRs at RIEC, Tohoku University, Japan. Figure 4 shows the measured and calculated HRTFs for the left ear at azimuth = 0° and elevation = 0°. As the graph shows, the frequency characteristics below 10 kHz are in good agreement. Moreover, a subjective evaluation was conducted to confirm the improvement in sound localization. The result showed that the accuracy of frontal sound localization was close to that using individual HRIRs [6].

4. APPLICATION TO GAME AUDIO

Video games are one of the most important prospective applications of 3D audio technology. In this section, we present our system developed for game audio.

4.1. System Developed for Game Audio

An overview of the system developed for game audio is shown in Fig. 5. At present, many games are being developed using middleware such as Unity [10], Unreal Engine [11], and original in-house systems. Furthermore, in the development of large-scale games, audio middleware is used to develop the audio part independently from the general game development. Game middleware is often called the game engine.

With this background, our 3D audio systems were developed as software plug-ins for the audio middleware “Wwise” [12].

The behavior of game audio systems can be described as follows. Game middleware and audio middleware have common event IDs. When an event occurs in a game, the game middleware sends the event ID and some parameters to the audio middleware. Once the audio middleware receives the event ID, it performs audio processing using the received parameters, inner functions, and plug-ins.

In most cases, the sound event contains localization information on where the sound event occurred. For example, when a gunshot event occurs, the game middleware transmits the gunshot event ID and the position to the audio middleware. In addition, if the game player uses an HMD, the audio middleware can also receive the position and direction of the player and calculate the relative position between the object and the player.

The audio middleware and its plug-ins can use the localization information of the player and the game object...
to process the audio signal. Therefore, 3D audio reproduction systems are suitable for application to video games.

4.2. Computational Cost Reduction

Although the computing power of personal computers and video game consoles is increasing, the proportion of this power that can be used for audio processing is limited. Hence, calculation cost reduction is still one of the primary issues in game audio processing.

In 3D audio processing, computational consumption considerably fluctuates depending on the number of sound objects. However, this situation is undesirable for game sound designers. A 3D audio plug-in should function with low computing power and little fluctuation.

Therefore, a system combining virtual loudspeaker localization and object localization was developed. The virtual loudspeaker localization system has a predetermined position for each loudspeaker and fixed HRIR coefficients from each virtual loudspeaker to the player. When the sound source or the player moves, the system describes the movement using 3D panning. Although it is difficult for the virtual loudspeaker system to express positions closer than the position of the virtual loudspeakers, the computational cost becomes lower than that required of objective localization. Table 1 shows the usage of the central processing unit (CPU) in objective localization and virtual loudspeakers. In this case, the number of virtual loudspeakers is 15.

Using this combination system, sound designers can choose objective localization if they want to set up sound sources close to the player, and other sound sources are processed by virtual loudspeaker localization to reduce the computational cost.

4.3. Timbral Effects

Undesirable timbral effects due to the application of HRIRs are the largest obstacles to implementing 3D audio systems using a set of headphones. Game sound designers do not want to change the original sound that they designed. Instead, the system should reduce timbral effects as much as possible.

The timbral effect is thought to be caused by a disparity between the HRIRs implemented in the system and individual HRIRs, and the characteristics of headphones. Ideally, the system should use individual HRIRs and equalize the characteristics of the headphones. However, it is difficult to obtain individual HRIRs of users and the characteristics of headphones in use. Therefore, HRIRs are often modified to reduce timbral effects in practical applications [13,14].

Since the modification of HRIRs is known to cause the deterioration of sound localization cues, the modification method is selected in accordance with the application.

In game audio systems, the sound coming from in front of the player should maintain the timbre of the original sound quality, as described above. By contrast, sound localization is important for the sound coming from behind the player. Therefore, the HRIRs for the sources in front of the player are modified to be close to a flat response of the frequency characteristic, whereas the HRIRs for the sources behind the player retain their original characteristics.

Figure 6 shows the original and modified HRTFs in front of the player, and Fig. 7 shows the HRTFs behind the player. For ease of viewing, the curve of the processed HRTF is shifted to $5\text{ dB}$.

Through the above modification, the sound designer can create balanced sound with both good sound quality and sound localization.

5. APPLICATION TO 360° VIDEO

There are several kinds of 360° video cameras on the market, and video sharing services are also spreading. Conversely, the audio configuration has still remained stereo in many cases.
When a viewer is watching a 360° video using an HMD, he can rotate his head and see the view behind him. However, with conventional stereo reproduction systems, it is impossible to present sounds corresponding to the movement of the viewer.

Therefore, we have built a 3D audio recording/reproduction system intended for use with a 360° camera for technical verification.

The basic idea behind this system is to combine higher-order ambisonics (HoA) recording, multichannel virtual loudspeaker reproduction, and HRIR convolution.

In principle, owing to the difficulty of defining a sound source position, it is difficult to express an ambient sound field using the HRIR convolution system. However, together with HoA techniques, distributed sound fields such as the sound of rustling leaves can be expressed.

5.1. Recording and HoA Encoding
We developed a 64-channel spherical microphone array system [15] for this application. Using this microphone array with a 360° camera, we can capture the sound field surrounding the camera.

The microphone array system includes 64 omnidirectional microphones, analog-to-digital converters, and a Dante [16] converter. After analog-to-digital conversion, the audio signal is converted to a signal in Dante format. Then the Dante signal is sent to the recording equipment, such as a PC, via Ethernet cables.

The 64-channel audio signal is encoded into a sixth-order ambisonics signal.

5.2. HoA Decoding and Binaural Processing
The Sixth-order ambisonics signal is decoded into a multichannel virtual loudspeaker array signal in the reproduction stage. The number of loudspeakers is determined by the balance between the computing power of the reproduction system and the required precision of the sound direction.

The next step is to combine 3D audio with 360° video using Unity [10].

Unity is one of the most popular game development tools. The 360° video is very similar to game audio in terms of the interactivity between the sound field and the listener. Hence, game development tools are also suitable for producing 360° video.

First, we create a virtual spherical screen to project the 360° image and set the position of the listener at the center of the sphere. Then, we set the virtual loudspeakers on the surface of the spherical screen.

Each virtual loudspeaker is assigned a decoded audio signal, and HRIRs from the virtual loudspeaker to both ears of the listener are calculated using the same 3D audio processing library as that of the plug-in shown in Sect. 4.

Finally, the 360° video content with 3D audio is created as an executable application.

Figure 8 is a screenshot of a 360° video that we have created.

6. CONCLUSIONS
In this paper, shape-based average HRIRs were presented, together with two applications of 3D audio systems using a set of headphones.

The spread of HMDs with position sensors has had a major impact on the practical application of 3D audio technology. Using these sensors mounted on HMDs, a dynamic VAD system that can follow the movement of a listener becomes easy to implement and can be applied to various products.

Although shape-based average HRIRs are valuable from a practical viewpoint, individualized HRIRs are preferable. To apply individualized HRIRs in 3D audio systems, it is necessary to develop an easy way to obtain individualized HRIRs [17]. Equalizing the characteristics of headphones and reproducing the reflection characteristics of original sound fields are also essential in reducing timbral effects.

Conversely, the present audio production environment is optimized for multichannel loudspeaker systems such as stereo and 5.1 channel. Therefore, when sound designers attempt to use objective-based audio and scene-based audio, there will be many difficulties in designing and monitoring the sound position, distance and movement.

As mentioned above, for the further spread of 3D audios, including video games and 360° video, it is essential not only to improve the quality but also to increase the usability of the production process. Therefore, our future work will focus on making continuous improvements in these fields.

ACKNOWLEDGEMENTS
We would like to thank members of the sound production team of CAPCOM Co., Ltd., and members of
the Yamaha R&D spatial audio group for useful suggestions and discussions.

REFERENCES

[1] http://sound.media.mit.edu/resources/KEMAR.html (accessed 30 September 2018).
[2] B. F. G. Katz, “Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation,” J. Acoust. Soc. Am., 110, 2440–2448 (2001).
[3] E. M. Wenzel, “Localization in virtual acoustic displays,” Presence: Teleop. Virt., 1, 80–107 (1992).
[4] B. Xie, Head-Related Transfer Function and Virtual Auditory Display (J. Ross, Florida, 2013).
[5] M. Morimoto and Y. Ando, “On the simulation of sound localization,” J. Acoust. Soc. Jpn. (E), 1, 167–174 (1980).
[6] S. Kaneko, T. Suenaga, M. Fujiwara, K. Kumebara, F. Shirakihara and S. Sekine, “Ear shape modeling for 3D audio and acoustic virtual reality,” Proc. AES 61st Conf. on Audio for Games (2016).
[7] B. Jian and B. C. Vemuri, “Robust point set registration using gaussian mixture models,” IEEE Trans. Pattern Anal. Mach. Intell., 33, 1633–1645 (2011).
[8] R. Coifman, V. Rokhlin and S. Wandzura, “The fast multipole method for the wave equation,” IEEE Antennas Propag. Mag., 35, 7–12 (1993).
[9] R. D. Ciskowski and C. A. Brebbia, Boundary Element Methods in Acoustics (Springer, New York, 1991).
[10] https://unity3d.com/ (accessed 30 September 2018).
[11] https://www.unrealengine.com/ (accessed 30 September 2018).
[12] https://www.audiokinetic.com/ (accessed 30 September 2018).
[13] J. Merimaa, “Modification of HRTF filters to reduce timbral effects in binaural synthesis,” Proc. AES 127th Convention, 7912 (2009).
[14] J. Merimaa, “Modification of HRTF filters to reduce timbral effects in binaural synthesis part 2: Individual HRTFs,” Proc. AES 129th Convention, 8265 (2010).
[15] S. Kaneko, T. Suenaga, H. Akiyama, Y. Miyake, S. Tominaga, F. Shirakihara and H. Okumura, “Development of a 64-channel spherical microphone array and a 122-channel loudspeaker array system for 3D sound field capturing and reproduction technology research,” Proc. AES 144th Convention, 10021 (2018).
[16] https://www.audinate.com/ (accessed 30 September 2018).
[17] S. Kaneko, T. Suenaga and S. Sekine, “DeepEarNet: Individualizing spatial audio with photography, ear shape modeling, and neural networks,” Proc. AES Conf. Audio for Virtual and Augmented Reality (2016).

Tsukasa Suenaga received his B.E. degree from Shibaura Institute of Technology, Japan, in 2003, and M.S. degree from Tohoku University, Japan, in 2005. He is currently working as a research engineer at Yamaha Corporation, Japan.

Shoken Kaneko received his B.E. degree in electronic engineering from the University of Tokyo, Japan, in 2008, and M.S. degree in physics from Radboud University Nijmegen, the Netherlands, in 2010. He is currently working as a research engineer at Yamaha Corporation, Japan.

Hiraku Okumura received his B.E. and M.E. degrees in architecture from Kyoto University in 1999 and 2001, respectively. He joined Yamaha Corporation, Japan, in 2001. He is a manager of the Signal Processing Group, Development Department 1, Research & Development Division, Yamaha Corporation. He was at Rochester Institute of Technology (Rochester, NY, USA) as a visiting professor from 2015 to 2016. He is a member of the Acoustical Society of Japan and Audio Engineering Society. He has been developing algorithms and software for 3D auditory control and adaptive digital filters.