International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Enhanced Realistic Audio Sound Generation based on Virtual Speaker Layout

Kwangki Kim

Abstract*: In this paper, we proposed a constant power panning (CPP) based realistic binaural sound generation with head related transfer function (HRTF) coefficients in the virtual twelve speakers layout arranged at intervals of 30 degrees for the VR service. In the proposed method, the original multi-channel audio signals are mapped to the virtual playback system using the CPP according to the users' head movement, and the realistic stereo binaural sound is formed by convolution of the mapped multi-channel audio signals and the HRTF coefficients for the virtual twelve speakers layout. Since the angle difference between the arbitrary adjacent two speakers is fixed as 30 degrees, the azimuthal resolution of the CPP based realistic sound generation is also 30 degrees and we can create the more accurate realistic sound reflecting the users' head movement. The experimental results show that the proposed method has a similar performance to the HRTF based realistic binaural sound generation even though it only needs about 0.79 Mbytes to be 1/30 of data amount of the HRTF coefficients compared with the HRTF based method.

***Keywords*: binaural rendering, constant power panning, HRTF, multi-channel audio, realistic audio.**

I. INTRODUCTION

To supply users with the immersive audio sound composed of 5.1 or more channel audio signals, the users should have the 5.1 or more multi-channel playback system. If the users do not have the multi-channel playback system and he wants to enjoy the realistic sound generated by the multi-channel audio signals via stereo headphones, a binaural rendering should be applied. The binaural rendering is a traditional technology to generate realistic audio sounds by convolution of head-related transfer function (HRTF) coefficients and the input audio signals [1-4]. Meanwhile, virtual reality (VR) is an interactive computer-generated experience that occurs primarily in a simulation environment by integrating auditory and visual feedback. But, since an audio in the VR service is provided based on the stereo headphone environment, a user cannot enjoy a realistic sound by 5.1 or more channel audio signals. Therefore, in order to deliver realistic sound using the stereo headphone to the VR service, the binaural rendering should be applied.

The HRTF based binaural rendering is very efficient in providing realistic sounds to the users' with stereo headphone in the VR service, but there is a limitation that stereo sound can have only fixed sound scenes. That is, since the binaural rendering is performed using the HRTF coefficients for a fixed multi-channel reproduction system, an accurate realistic audio signal cannot be generated in a changed sound scene according to the movement of the users' in the VR service. Therefore, the HRTF coefficients of the multi-channel speaker layout must be changed in accordance with the users' head movement, and the HRTF coefficients for all directions, 360 degrees, must be stored in the memory in order to produce the accurate realistic sound in the VR service.

The data amount of the HRTF coefficients of all directions of 360 degrees is about 23.6 Mbytes, so the HRTF based realistic sound generation may not be implemented in the embedded environment. Therefore, we proposed the constant power panning (CPP) based realistic sound generation method [5, 6]. In the CPP based method, the HRTF coefficients are not changed and, instead, the multi-channel playback environment is regenerated to reflect the users' movement. Then, the original multi-channel signals are mapped to the newly formed multi-channel speaker layout, and the realistic binaural stereo sound is produced by convolution of the newly generated multi-channel signals and the fixed HRTF coefficients. Even if the CPP based method generates rather successfully the realistic sound with only a small amount of HRTF coefficients, it does not fully reflect the users' head movement to the realistic sound due to the low azimuth resolution. Finally, we proposed the CPP based realistic sound generation method in the virtual speaker layout arranged at intervals of 30 degrees. Because the angle between any adjacent speakers is fixed at 30 degrees, the azimuthal resolution of the CPP-based realistic sound generation is also 30 degrees and can produce the more accurate and realistic sound reflecting the users' head movement. This paper consists of as follows. In chapter 2, we describe the HRTF based realistic sound generation method for VR. In chapter 3, we proposed the CPP based realistic sound generation method. In chapter 4 and 5, we give the experimental results and the conclusion.

II. HRTF BASED REALISTIC AUDIO SOUND GENERATION FOR VR SERVICE

Fig. 1 shows a HRTF based binaural rendering for the realistic sound generation. Since the HRTF coefficients represent the signal path from each speaker to the human ear in 5.1 or more multi-channel audio playback system,

Manuscript published on 30 September 2019
* Correspondence Author
Kwangki Kim*, Department of IT convergence, Korea Nazarene University, Cheonan, South Korea. Email: k2kim@kornu.ac.kr

we can create realistic audio sound for a stereo headphone environment by convolving multichannel audio signals with measured HRTF coefficients. If there are 5.1 channel playback system, we need 10 HRTF coefficients, which are the path for moving the audio signal from 5.1 speaker layout to the human's left and right ears. Using 10 HRTF coefficients, we can obtain the stereo binaural sound with the 5.1 channel audio effect as shown in Fig.2 and (1).

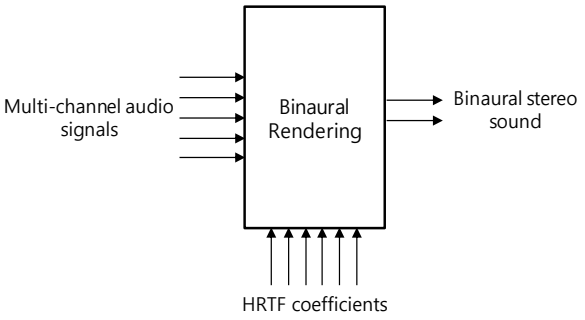

Fig. 1. HRTF based binaural rendering.

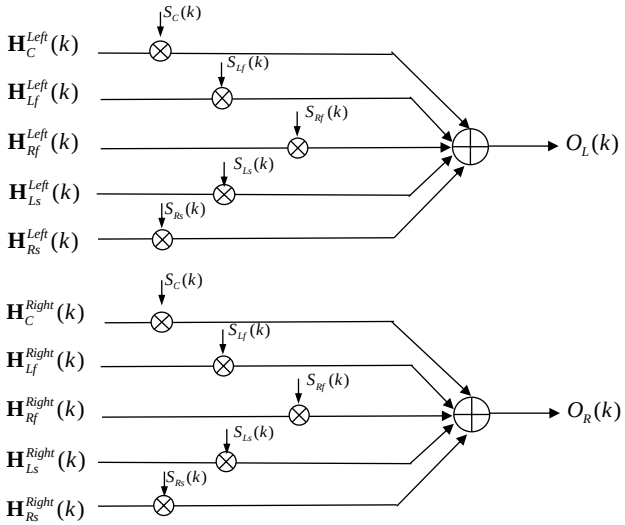

Fig. 2. Overall structure of the binaural rendering for 5.1 channel speaker layout.

$$\begin{bmatrix} O_L(k) \\ O_R(k) \end{bmatrix} = \begin{bmatrix} H_C^{Left}(k) & H_C^{Right}(k) \\ H_{Lf}^{Left}(k) & H_{Lf}^{Right}(k) \\ H_{Rf}^{Left}(k) & H_{Rf}^{Right}(k) \\ H_{Ls}^{Left}(k) & H_{Ls}^{Right}(k) \\ H_{Rs}^{Left}(k) & H_{Rs}^{Right}(k) \end{bmatrix}^T \times \begin{bmatrix} S_c(k) \\ S_{Lf}(k) \\ S_{Rf}(k) \\ S_{Ls}(k) \\ S_{Rs}(k) \end{bmatrix}, \text{ for } 0 \le k \le M-1 \quad (1)$$

Here, $O_L(k)$ and $O_R(k)$ are the left and right binaural sound. C, Lf, Rf, Ls, and Rs are the center, left front, right front, left surround, and right surround in 5.1 channel speaker layout. $H_X^{Left}(k)$ and $H_X^{Right}(k)$ are the HRTF coefficients in frequency domain from the arbitrary channel X to human's left and right ear while $S_X(k)$ is a signal of the arbitrary channel X in frequency domain. K is the frequency index and M is the FFT size. For the arbitrary N channels speaker layout, Fig. 2 and (1) can be generalized as in Fig. 3 and (2).

$$\begin{bmatrix} O_L(k) \\ O_R(k) \end{bmatrix} = \begin{bmatrix} H_1^{Left}(k) & H_1^{Right}(k) \\ H_2^{Left}(k) & H_2^{Right}(k) \\ \vdots & \vdots \\ H_{N-1}^{Left}(k) & H_{N-1}^{Right}(k) \\ H_N^{Left}(k) & H_N^{Right}(k) \end{bmatrix}^T \times \begin{bmatrix} S_1(k) \\ S_2(k) \\ \vdots \\ S_{N-1}(k) \\ S_N(k) \end{bmatrix}, \text{ for } 0 \le k \le M-1 \quad (2)$$

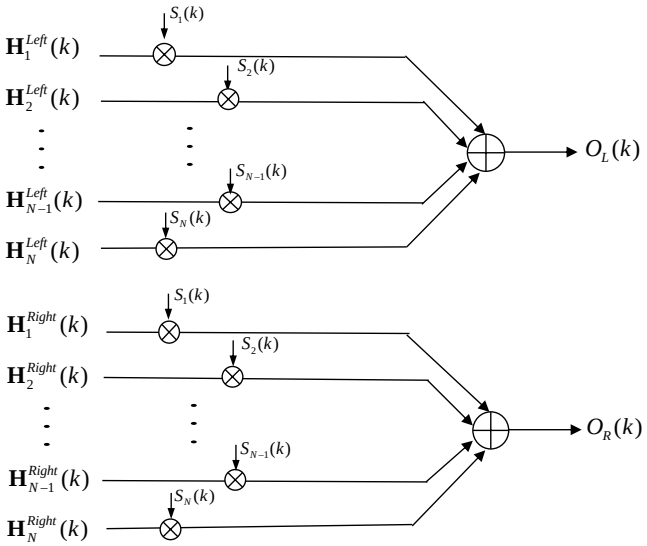

Fig. 3. Overall structure of the binaural rendering for N multi-channel speaker layout.

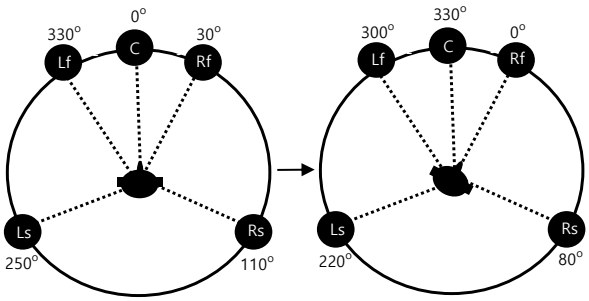

Fig. 4. Example of the change of 5.1 channel speaker layout according to the users' head azimuth change by 30 degrees.

Meanwhile, because the audio signal in the VR service is delivered to the users' through the stereo headphone, the users cannot enjoy the realistic audio sound formed by the multi-channel audio signals. Therefore, the binaural rendering described above should be also applied to the VR service to provide the realistic sound to the users with the stereo headphone. Although the HRTF based realistic sound generation is very efficient for providing the users' with the realistic sound in the VR service with stereo headphone, there is a limitation that the binaural sound only has a fixed sound scene. Namely, since the binaural rendering is performed by using the HRTF coefficients for the fixed multi-channel playback system, it cannot produce the accurate realistic sound with the changed sound scene according to the users' head change in the VR service. For the explanation, if the users' changes his/her head azimuth by 30 degrees, the original multi-channel speaker layout should be also changed.

Retrieval Number: C4518098319/19©BEIESP
DOI:10.35940/ijrte.C4518.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

And the binaural stereo sound is recalculated using the new HRTF coefficients for the changed multi-channel speaker layout. Fig. 4 shows an example of the change of 5.1 multi-channel speaker layout according to the users' head movement by 30 degrees, and the realistic audio sound to reflect the users' head movement is recalculated as in (3). As a same manner, any other multi-channel speaker layout can be changed according to the users' head movement, and the accurate realistic sound is calculated based on the changed multi-channel speaker layout as in (4).

$$\begin{bmatrix} O_L(k) \\ O_R(k) \end{bmatrix} = \begin{bmatrix} H^{Left}_{330^\circ}(k) & H^{Right}_{330^\circ}(k) \\ H^{Left}_{300^\circ}(k) & H^{Right}_{300^\circ}(k) \\ H^{Left}_{0^\circ}(k) & H^{Right}_{0^\circ}(k) \\ H^{Left}_{80^\circ}(k) & H^{Right}_{80^\circ}(k) \\ H^{Left}_{220^\circ}(k) & H^{Right}_{220^\circ}(k) \end{bmatrix}^T \times \begin{bmatrix} S_c(k) \\ S_{Lf}(k) \\ S_{Rf}(k) \\ S_{Ls}(k) \\ S_{Rs}(k) \end{bmatrix}, \text{ for } 0 \le k \le M-1 \quad (3)$$

$$\begin{bmatrix} O_L(k) \\ O_R(k) \end{bmatrix} = \begin{bmatrix} H^{Left}_{1+\theta}(k) & H^{Right}_{1+\theta}(k) \\ H^{Left}_{2+\theta}(k) & H^{Right}_{2+\theta}(k) \\ \vdots & \vdots \\ H^{Left}_{N-1+\theta}(k) & H^{Right}_{N-1+\theta}(k) \\ H^{Left}_{N+\theta}(k) & H^{Right}_{N+\theta}(k) \end{bmatrix}^T \times \begin{bmatrix} S_1(k) \\ S_2(k) \\ \vdots \\ S_{N-1}(k) \\ S_N(k) \end{bmatrix}, \text{ for } 0 \le k \le M-1 \quad (4)$$

Here, θ is the degree of the users' head azimuth change. and $H^{Left}_{X+\theta}$ and $H^{Right}_{X+\theta}$ are the newly replaced left and right HRTF coefficients of the arbitrary channel X according to the users' head azimuth change. Finally, the overall structure of binaural rendering in Fig. 1 is updated as Fig. 5.

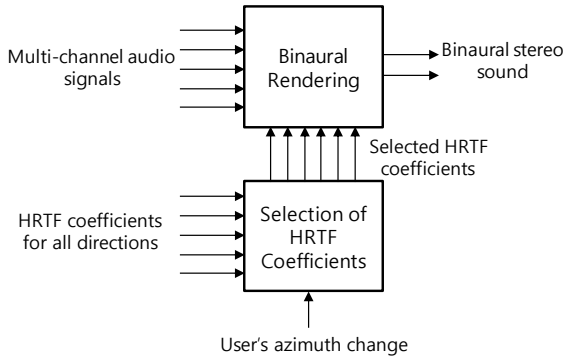

Fig. 5. HRTF based binaural rendering according to user's head movement

As described above, since the HRTF coefficients of the multi-channel speaker layout are changed according to the users' head azimuth variation, the HRTF coefficients for all directions of 360 degrees should be stored in the memory to generate the accurate realistic sound in the VR service.

III. CPP BASED REALISTIC AUDIO SOUND GENERATION FOR VR SERVICE

The data amount of the HRTF coefficients of 5.1 multi-channel speaker layout is about 0.33 Mbytes while that of all directions of 360 degrees is about 23.6 Mbytes. So, the binaural rendering to generate the realistic sound using HRTF coefficients for the VR service may not be implemented in the embedded environment with low memory storage. Therefore, we proposed the CPP based realistic sound generation method. In the proposed method, the HRTF coefficients of the 5.1 or more channel speaker layout are never changed. Instead, the multi-channel speaker layout is newly formed according to the users' head movement and the multi-channel signals are mapped to the newly formed multi-channel speaker layout. So, the data amount of the HRTF coefficients is the same as the conventional HRTF based binaural rendering, and only additional operations are needed for the audio signal mapping. Fig. 6 shows an example of the newly formed multi-channel speaker layout and the audio signal mapping when the users' head azimuth change is 180 degree. According to the CPP based realistic sound generation, Fig. 1 and (4) are updated as Fig. 7 and (5).

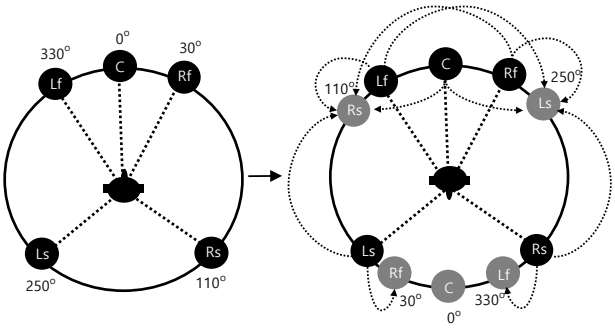

Fig. 6. Example of signal mapping of 5.1 channel audio for user's head movement as 180 degrees.

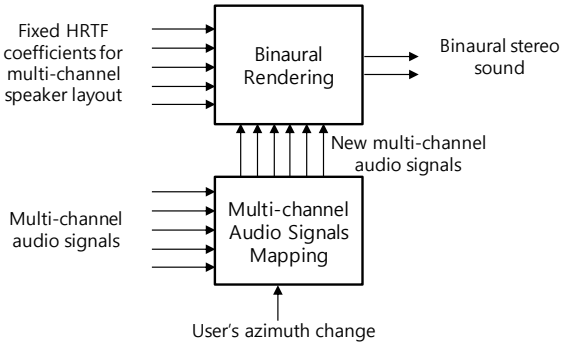

Fig. 7. CPP based binaural rendering according to user's head movement

$$\begin{bmatrix} O_L(k) \\ O_R(k) \end{bmatrix} = \begin{bmatrix} H^{Left}_1(k) & H^{Right}_1(k) \\ H^{Left}_2(k) & H^{Right}_2(k) \\ \vdots & \vdots \\ H^{Left}_{N-1}(k) & H^{Right}_{N-1}(k) \\ H^{Left}_N(k) & H^{Right}_N(k) \end{bmatrix}^T \times \begin{bmatrix} S'_1(k) \\ S'_2(k) \\ \vdots \\ S'_{N-1}(k) \\ S'_N(k) \end{bmatrix}, \text{ for } 0 \le k \le M-1 \quad (5)$$

Here, $S'_X(k)$ is a newly generated signal of channel X when the original signals are mapped to the newly formed multi-channel speaker layout. For the explanation of the CPP method for multi-channel audio signals mapping, let's assume that there are two channels speaker layout- C1 and C2, and a channel C3 lays in between C1 and C3 as the result of the users' head change (θ_n) as shown in Fig. 8. Then, a signal of C3 is mapped to C1 and C2. Given the channel gains of C1($S_1(k)$), C2($S_2(k)$), and C3($S_3(k)$), the $S_3(k)$ is projected to C1 and C2 using (6) and (7).

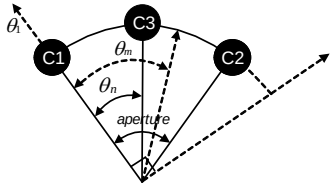

Fig. 8. Example of arbitrary channel mapping using CPP method

$$\theta_m = \frac{(\theta_n - \theta_1)}{(aperture - \theta_1)} \times \frac{\pi}{2} \tag{6}$$

$$\left.\begin{aligned} S_1'(k) &= S_1(k) + S_3(k) \times \cos(\theta_m) \\ S_2'(k) &= S_2(k) + S_3(k) \times \sin(\theta_m) \end{aligned}\right\}, \text{ for } 0 \le k \le M\text{-}1 \tag{7}$$

Here, θ_m is the normalization of the azimuth angle (θ_n) of C3 located between C1 and C2 at 90 degrees, and it is less than 90 degrees. θ_1 is the azimuth angle of C1 and *aperture* is the angle between C1 and C2. $S_1'(k)$ and $S_2'(k)$ are the newly generated signals of C1 and C2 as the result of channel mapping using the CPP [7, 8]. In this way, the entire original audio signals are mapped to the newly formed speaker layout, and HRTF is applied to the newly generated multi-channel audio signals to generate the realistic audio sound according to the users' head azimuth change.

Meanwhile, the more speakers in the playback environment, the more immersive sound can be produced, and the same is true for stereo binaural sound based on the same playback environment. In other words, the binaural sound based on the 10.1 channel reproduction environment is more immersive than the binaural sound based on the 5.1 channel reproduction environment. Also, although the CPP based realistic sound generation rather successfully generates the realistic audio sound only with the HRTF coefficients for the multi-channel speaker layout, it can more efficiently generate a realistic sound that successfully reflects user 's change of azimuth angle in a playback environment with a larger number of channels. Since the users' head movement may not be accurately reflected due to the low azimuthal resolution of the CPP based realistic sound generation in the playback environment with a small number of channels, we proposed the CPP based realistic sound generation method in twelve virtual playback environments arranged at intervals of 30 degrees as shown in Fig. 9. The whole multi-channel audio signals are mapped to the new virtual playback system using the CPP according to the users' head movement, and the newly generated multi-channel audio signals are convolved with the HRTF coefficients for the virtual twelve speaker layouts to create the realistic stereo binaural sound. Since the angle between the arbitrary adjacent speakers is fixed as 30 degrees, the azimuthal resolution of the CPP based realistic sound generation is also 30 degrees and we can create the more accurate realistic sound reflecting the users' head movement. Consequently, the data amount of the HRTF coefficients is slightly increased to about 0.79 Mbytes and (5) is updated as (8).

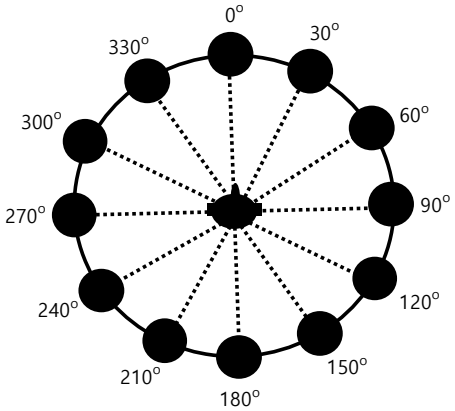

Fig. 9. Proposed virtual twelve speaker layouts

$$\begin{bmatrix} O_L(k) \\ O_R(k) \end{bmatrix} = \begin{bmatrix} H_{0^\circ}^{Left}(k) & H_{0^\circ}^{Right}(k) \\ H_{30^\circ}^{Left}(k) & H_{30^\circ}^{Right}(k) \\ H_{60^\circ}^{Left}(k) & H_{60^\circ}^{Right}(k) \\ \vdots & \vdots \\ H_{270^\circ}^{Left}(k) & H_{270^\circ}^{Right}(k) \\ H_{300^\circ}^{Left}(k) & H_{300^\circ}^{Right}(k) \\ H_{330^\circ}^{Left}(k) & H_{330^\circ}^{Right}(k) \end{bmatrix}^T \times \begin{bmatrix} S_{0^\circ}'(k) \\ S_{30^\circ}'(k) \\ S_{60^\circ}'(k) \\ \vdots \\ S_{270^\circ}'(k) \\ S_{300^\circ}'(k) \\ S_{330^\circ}'(k) \end{bmatrix}, \text{ for } 0 \le k \le M\text{-}1 \tag{8}$$

IV. EXPERIMENTAL RESULTS

To check the validity of the proposed CPP based realistic sound generation in the virtual twelve speakers layout, the subjective listening test was performed. For the test, we used three 5.1 multi-channel audio contents listed in Table-I. Five subjects were participated in the test. They determined the azimuthal change of the realistic audio sound by three systems listed in Table-II compared to the original multi-channel audio signals when the users' 's azimuth change is 45 degrees and 90 degrees.

Table-I: Test materials

Material	Description
ARL_applause	Ambience
Chostakovitch	Music (back: direct)
Fountain_music	Pathological

Table-II: System under Test.

Classification	Description
HRTF	Binaural sound generation with HRTF coefficients for all directions
CPP	Binaural sound generated with HRTF coefficients for fixed multi-channel speaker layout and CPP method
ECPP	Binaural sound generated with HRTF coefficients for virtual twelve multi-channel speaker layout and CPP method

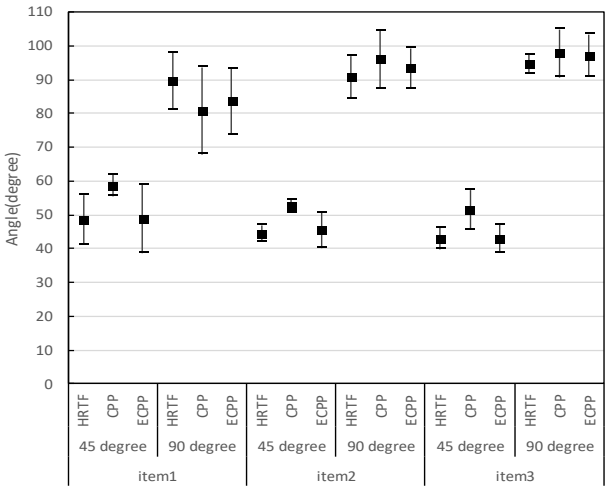

Fig. 10. Subjective listening test results.

Fig. 10 shows the subjective listening test results. For all test items and degrees, 'HRTF' shows the best realistic sound successfully reflecting the users' 's change of azimuth angle while 'CPP' shows the worst realistic sound rather unsuccessfully reflecting the users' head movement. Although 'ECPP' shows a somewhat lower performance than 'HRTF', 'ECPP' has almost similar performance to 'HRTF'. Since 'ECPP' requires about 1/30 of data amount compared with 'HRTF' and shows almost similar performance, it can be confirmed that 'ECPP' is a very effective technology.

V. CONCLUSION

In this paper, we proposed the CPP based realistic binaural sound generation with the HRTF coefficients in the virtual twelve speakers layout for the VR service. In the proposed method, the original multi-channel audio signals are mapped to the virtual playback system using the CPP according to the users' head movement, and the realistic stereo binaural sound is produced by the convolution of the the newly generated multi-channel audio signals and the HRTF coefficients for the virtual twelve speakers layout. The experimental results show that the proposed method has a similar performance to the HRTF based realistic binaural sound generation even though it only needs about 1/30 of data amount of the HRTF coefficients compared with the HRTF based method. Since the proposed method considers the users' head movement, it cannot reflect the users' elevation change. The more reliable realistic sound generation considering the users' free movement including elevation change remains as a future work.

ACKNOWLEDGMENT

This work was funded by the research fund of Korea Nazarene University in 2019. This research also was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2017R1D1A3B03034951) and the MSIT(NRF-2018R1A4A1025559).

REFERENCES

1. B. Gardner and K. Martin, "HRTF Measurements of a KEMAR Dummy Head Microphone," MIT Media Lab Perceptual Computing -technical Report #280, May 1994
2. Breebaart J., Herre J., Jin C., Kjörling K., Koppens J., Plogsties J., & Villemoes L., "Multi-channel goes mobile: MPEG Surround binaural rendering," In Proceedings of the Audio Engineering Society Conference: 29th International Conference: Audio for Mobile and Handheld Devices. Audio Engineering Society, 2006.
3. K. Kim and J. Kim, "Binaural decoding for efficient multi-channel audio service in network environment," In Proceedings of the 2014 IEEE 11th Consumer Communications and Networking Conference, pp. 525-526, Jan. 2014.
4. K. Kim, "A study on complexity reduction of binaural decoding in multi-channel audio coding for realistic audio service," Contemporary Engineering Sciences, Vol. 9, 2016, no. 1, pp. 11-19, Jan. 2016.
5. K. Kim, "Sound scene control of multi-channel audio signals for realistic audio service in wired/wireless network," International Journal of Multimedia and Ubiquitous Engineering, vol. 9, no. 2, 2014.
6. E. Zwicker and H. Fastl, Psychoacoustics, Springer-Verlag, Berlin,Heidelberg, 1999.
7. V. Pulki, "Virtual sound source positioning using vector base amplitude panning," Journal of Audio Engineering Society, vol. 45, pp. 456-466, 1997.
8. M. A. Gerzon, "Panpot laws for multispeaker stereo," In Proceedings of the 92nd Convention of the AES, Journal of Audio Engineering Society, Preprint 3309, 1992.

AUTHORS PROFILE

Kwangki Kim received the B.S. degree in electronic engineering from Korea Aviation University, Koyang, South Korea, in 2002, the M. S. degree and the Ph. D. degree in department of Information and Communications Engineering at Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2004 and 2011, respectively. In 2012, he was with the Samsung DMC R&D center. From 2013, he has been an associate professor of the Department of Information Technology Convergence, Korea Nazarene University, Chonan, South Korea. His research interests include spatial audio coding, audio object coding, 3D audio processing and their applications.