Complex Plane based Realistic Sound Generation for Free Movement in Virtual Reality

Kwangki Kim
Department of IT Convergence
Korea Nazarene University, Cheon-an, South Korea

***Abstract*—A binaural rendering is a technology that generates a realistic sound for a user with a stereo headphone, so it is essential for the stereo headphone based virtual reality (VR) service. However, the binaural rendering has a problem that it cannot reflect the user's free movement in the VR. Because the VR sound does not match with the visual scene when the user moves freely in the VR space, the performance of the VR may be degraded. To reduce the mismatch problem in the VR, the complex plane based stereo realistic sound generation method was proposed to allow the user's free movement in the VR causing the change of the distance and azimuth between the user and the speaker. For the calculation of the distance and the azimuth between the user and the speaker by the user's position change, the 5.1 multichannel speaker playback system and the user are placed in the complex plane. Then, the distance and the azimuth between the user and the speaker can be simply calculated as the distance and the angle between two points in the complex plane. The 5.1 multichannel audio signals are scaled by the estimated five distances according to the inverse square law, and the scaled multichannel audio signals are mapped to the newly generated virtual 5.1 multichannel speaker layout using the measured five azimuths and the azimuth by the head movement. Finally, we can successfully obtain the stereo realistic sound to reflect the user's position change and the head movement through the binaural rendering using the scaled and mapped 5.1 multichannel audio signals and the HRTF coefficients. Experimental results show that the proposed method can generate the realistic audio sound reflecting the user's position and azimuth change in the VR only with less than about 5 % error rate.**

Keywords*—*Virtual reality; realistic sound; binaural rendering; constant power panning; head related transfer function

I. INTRODUCTION

In general, users should have their own multi-channel audio playback environment to enjoy the realistic sound by multi-channel audio signals. However, most of the users have a stereo headphone environment, so they are unable to enjoy realistic audio by the multi-channel audio signals. Therefore, a head related transfer function (HRTF) [1] based binaural rendering has been proposed to solve this limitation [2-6]. In particular, the binaural rendering is essential to deliver the more realistic audio signal to the users in a system such as a virtual reality (VR) service based on the stereo headphone environment [8-10]. In the binaural rendering, the stereo realistic sound is generated using the multi-channel audio signals and the HRTF coefficients. The stereo realistic sound generation based on the binaural rendering can efficiently supply the realistic sound with the user in the VR service, but there is a critical limitation that the existing stereo realistic sound generation based on the binaural rendering does not reflect the user's position change. Since the stereo realistic sound generation through binaural rendering with a fixed HRTF cannot reflect the user's position change, there is a gap between the visual scene and the sound causing the performance degradation of the VR service. To solve the fixed sound scene problem in the VR, the sound scene control of the stereo realistic sound in the VR was introduced to reflect the user's head azimuth change [11]. In [11], the HRTF coefficients are replaced by the new HRTF coefficients corresponding to the user's azimuth change, and the realistic sound with the controlled sound scene is calculated with the multi-channel audio signals and the replaced HRTF coefficients. Although the realistic sound generation with the substitution of the HRTF coefficients can successfully generate the stereo realistic sound with the controlled sound scene according to the user's head movement, it needs very high data amount of the stored HRTF coefficients for all azimuth directions. The data rate of the HRTF coefficients are 23.6 Mbytes to be 32 times compared with that of the HRTF coefficients of the 5.1 multi-channel speaker layout. Therefore, the sound scene control of the realistic sound with the substitution of the HRTF coefficients is not suitable for the embedded system with low memory storage. Accordingly, the constant power panning (CPP) based sound scene control of the realistic sound was introduced [12-15]. The CPP based sound scene control scheme used only the HRTF coefficients of the 5.1 multi-channel speaker layout, so the data rate of the HRTF coefficients is exactly same as the original binaural rendering. Instead, the CPP based sound scene control method mapped the original multi-channel audio signals onto the new 5.1 multi-channel speaker layout rearranged by the user's head movement. The CPP based method can be applied to the embedded system with the low memory storage because it can generate the realistic sound reflecting the user's head movement without the increase of the HRTF coefficients.

Meanwhile, the VR service allows the user's free movement in the VR space, so the VR service should consider the user's not only head movement but also position change. Namely, the VR service should generate the realistic sound reflecting the user's free movement. However, since the sound scene control based on the HRTF coefficients and the CPP method only focuses on the modification of the stereo realistic sound scene according to the user's azimuth change, its' stereo realistic sound cannot imply the user's distance change.

This work was funded by the research fund of Korea Nazarene University in 2021. Also, this research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A3B03034951).

Therefore, there is still a mismatch between the VR scene and the VR sound when the user freely moves in the VR space and the overall performance of the VR service may be very poor. The realistic sound generation method based on a complex plane for tracking the user's movement is proposed to improve the performance of the VR service by reflecting the user's free movement in the VR sound. The user's free movement (position change) in the VR space causes both changes of the distance and the azimuth between the user and the speaker, while the user's head movement only effects on the azimuth change. Therefore, the proposed method separately handles the user's position change and the head movement and it calculates the distance and the azimuth between the user and the speaker by the user's free movement. Then, the proposed method can generate the realistic sound by scaling the audio signal using the measured distance and by adjusting the sound scene using the final azimuth formed by adding the measured azimuth for the position change and the azimuth change for the head movement. In conclusion, the proposed method can improve the overall performance of the VR service by generating the realistic sound that reflects the user's free movement including the head movement. This paper consists of as follows. In Section 2, the stereo realistic sound generation through the binaural rendering and the sound scene control of the realistic sound is described. In Section 3, the realistic sound generation for the user's free movement in the VR is proposed. In Sections 4 and 5, the experimental result and the conclusion will be given, respectively.

II. Stereo Realistic Sound Generation based on Binaural Rendering for VR

A. Binaural Rendering for VR

The VR system needed the stereo realistic sound generation method for the immersive effect by the multi-channel audio signals since the VR system used the stereo headphone for the delivery of the VR sound. The VR system adopted the conventional binaural rendering for generating the stereo realistic sound [1-7]. The binaural rendering is a technology that generates the stereo realistic audio sound with the multi-channel audio effect for stereo headphone environment using HRTF coefficients to characterize all signal paths from speakers to human ears [1]. As shown in Fig. 1, the binaural rendering is computed with the input multi-channel signal and the HRTF coefficients. To generate the output stereo realistic sound, the input multi-channel audio signals are convolved with the HRTF coefficients as in (1) [2-4].

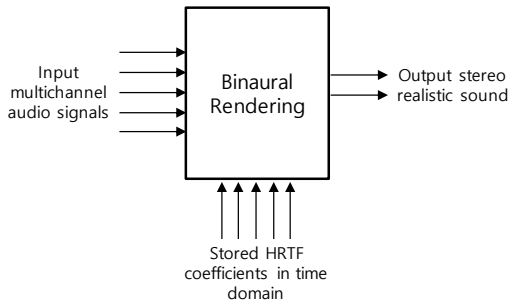

Fig. 1. Stereo Realistic Sound Generation through Binaural Rendering.

$$o_L = \sum_{n=1}^{N}\left(s_n \otimes h_n^L\right),\ o_R = \sum_{n=1}^{N}\left(s_n \otimes h_n^R\right) \tag{1}$$

where h_n^L and h_n^R are the stored HRTF coefficients in time domain from the nth channel to human left and right ear. o_L and o_R are the output left and right realistic signals in time domain and s_n is the nth channel input signal in time domain. N is the channel number of the multi-channel audio signals and $\otimes$ is the linear convolution. Since the linear convolution in time domain between the input signals and the HRTF coefficients has very high computational complexity, the binaural rendering is calculated as the multiplication of the input signals and the HRTF coefficients in the frequency domain as in (2)[5] and Fig. 1 is updated as Fig. 2.

$$O_L = \sum_{n=1}^{N}\left(S_n \cdot H_n^L\right),\ O_R = \sum_{n=1}^{N}\left(S_n \cdot H_n^R\right) \tag{2}$$

where H_n^L and H_n^R are the stored HRTF coefficients in frequency domain from the nth channel to human left and right ear. O_L and O_R are the output left and right realistic signals in frequency domain and S_n is the nth channel input signal in frequency domain.

Meanwhile, (2) can be rewritten in matrix form for 5.1 multi-channel audio signals as in (3) [11].

$$\begin{bmatrix} O_L(k) \\ O_R(k) \end{bmatrix} = \begin{bmatrix} H_C^{Left}(k) & H_C^{Right}(k) \\ H_{Lf}^{Left}(k) & H_{Lf}^{Right}(k) \\ H_{Rf}^{Left}(k) & H_{Rf}^{Right}(k) \\ H_{Ls}^{Left}(k) & H_{Ls}^{Right}(k) \\ H_{Rs}^{Left}(k) & H_{Rs}^{Right}(k) \end{bmatrix}^T \times \begin{bmatrix} S_c(k) \\ S_{Lf}(k) \\ S_{Rf}(k) \\ S_{Ls}(k) \\ S_{Rs}(k) \end{bmatrix}, \text{for } 0 \le k \le M\text{-}1 \tag{3}$$

where $O_L(k)$ and $O_R(k)$ are the output left and right realistic sound. $S_X(k)$ is the arbitrary channel X signal in frequency domain and k is the frequency index. C, Lf, Ls, Rf and Rs are the center, left front, left surround, right front, and right surround of the 5.1 multi-channel audio signals. $H_X^{Left}(k)$ and $H_X^{Right}(k)$ are the stored HRTF coefficients in frequency domain from the arbitrary channel X to human's left and right ear. M is the number of the FFT size.

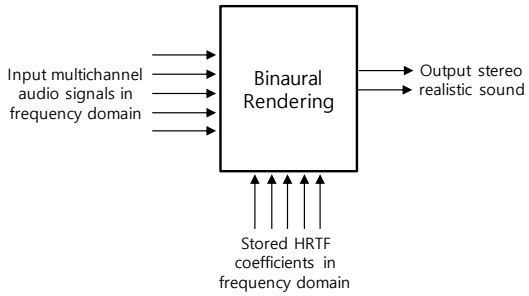

Fig. 2. Stereo Realistic Sound Generation through Binaural Rendering in Frequency Domain.

B. Sound Scene Control of Stereo Realistic Sound Reflecting Azimuth Change in VR

Although the conventional binaural rendering was useful for the VR system, it could not reflect the user's azimuth change. Therefore, the sound scene control of the stereo realistic sound was proposed [11, 12]. When the azimuth angle of the user changed in the 5.1 channel reproduction environment, the direction in which the 5.1 channel signal is transmitted to the user or the azimuth angle of the 5.1 channel reproduction environment also changed. So, the existing HRTF coefficients should be replaced by new HRTF coefficients corresponding to the azimuth angle of the new 5.1 channel reproduction environment. The binaural rendering with the sound scene control could generate the realistic sound with the substituted HRTF coefficients and the 5.1 channel audio signal according to the user's azimuth change as in (4).

$$\begin{bmatrix} O_{L,\theta_{hm}}(k) \\ O_{R,\theta_{hm}}(k) \end{bmatrix} = \begin{bmatrix} H^{Left}_{C-\theta_{hm}}(k) & H^{Right}_{C-\theta_{hm}}(k) \\ H^{Left}_{Lf-\theta_{hm}}(k) & H^{Right}_{Lf-\theta_{hm}}(k) \\ H^{Left}_{Rf-\theta_{hm}}(k) & H^{Right}_{Rf-\theta_{hm}}(k) \\ H^{Left}_{Ls-\theta_{hm}}(k) & H^{Right}_{Ls-\theta_{hm}}(k) \\ H^{Left}_{Rs-\theta_{hm}}(k) & H^{Right}_{Rs-\theta_{hm}}(k) \end{bmatrix}^T \times \begin{bmatrix} S_c(k) \\ S_{Lf}(k) \\ S_{Rf}(k) \\ S_{Ls}(k) \\ S_{Rs}(k) \end{bmatrix}, \text{ for } \begin{cases} 0 \le k \le M-1 \\ 0^\circ \le \theta_{hm} \le 360^\circ \end{cases} \tag{4}$$

where $O_{L,\theta_{hm}}(k)$ and $O_{R,\theta_{hm}}(k)$ the output left and right realistic sound with controlled sound scene according to the user's head movement. $H^{Left}_{X-\theta_{hm}}(k)$ and $H^{Right}_{X-\theta_{hm}}(k)$ are the substituted HRTF coefficients corresponding to the user's azimuth change from the arbitrary channel X to human's left and right ear. θ_{hm} is the angle of the user's head movement. Here, if the angle of any channel X minus θ_{hm} is negative, the final azimuth of any channel is the calculated angle plus 360 degrees. Fig. 3 shows an example of the user's azimuth change in the 5.1 multi-channel speaker layout. Since the angle of the user's azimuth change is 90 degrees, the angle of existing 5.1 multi-channel speaker layout is rearranged as shown in Fig. 3 and the HRTF coefficients are substituted to reflect the rearranged multi-channel speaker layout. The binaural rendering generates the stereo realistic sound with the controlled sound scene using the 5.1 multi-channel audio signals and the substituted HRTF coefficients as in (5). Fig. 4 shows the overall procedure of the sound scene control of the realistic sound based on the substitution of the HRTF coefficients.

$$\begin{bmatrix} O_{L,90^\circ}(k) \\ O_{R,90^\circ}(k) \end{bmatrix} = \begin{bmatrix} H^{Left}_{270^\circ}(k) & H^{Right}_{270^\circ}(k) \\ H^{Left}_{240^\circ}(k) & H^{Right}_{240^\circ}(k) \\ H^{Left}_{300^\circ}(k) & H^{Right}_{300^\circ}(k) \\ H^{Left}_{160^\circ}(k) & H^{Right}_{160^\circ}(k) \\ H^{Left}_{20^\circ}(k) & H^{Right}_{20^\circ}(k) \end{bmatrix}^T \times \begin{bmatrix} S_c(k) \\ S_{Lf}(k) \\ S_{Rf}(k) \\ S_{Ls}(k) \\ S_{Rs}(k) \end{bmatrix}, \text{ for } 0 \le k \le M-1 \tag{5}$$

Although the above explained sound scene control scheme of the realistic sound can successfully generate the stereo realistic sound with the controlled sound scene, it needed very high data amount of the stored HRTF coefficients as 23.6 Mbytes. Therefore, the embedded system with the low memory storage could not implement the sound scene control of the realistic sound with the substitution of the HRTF coefficients. Accordingly, the CPP based sound scene control of the realistic sound was introduced [11-15]. The CPP based sound scene control scheme fixed the HRTF coefficients of the 5.1 multi-channel speaker layout and it mapped the existing multi-channel audio signals onto the new 5.1 multi-channel speaker layout rearranged by the user's head movement. Fig. 5 shows an example of the mapping of the multi-channel audio signals to the newly formed 5.1 multi-channel speaker layout according to the user's head movement. The 5.1multi-channel speaker layout is newly created around the user's new front, and the existing 5.1 multi-channel audio signals are mapped onto the new speaker layout using the CPP technique. The binaural rendering is performed as in (6) using the mapped 5.1 multi-channel signals and the HRTF coefficients of the 5.1 multi-channel speaker layout to generate stereo realistic sound with the controlled sound scene according to the user's head movement.

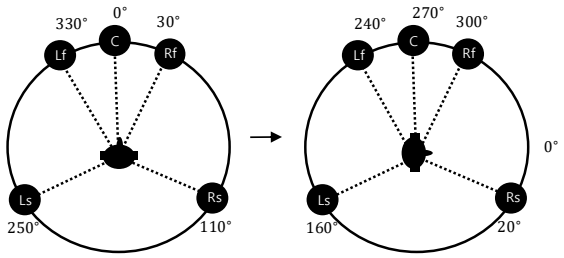

Fig. 3. Example of the user's Azimuth Change in the 5.1 Multi-channel Speaker Layout.

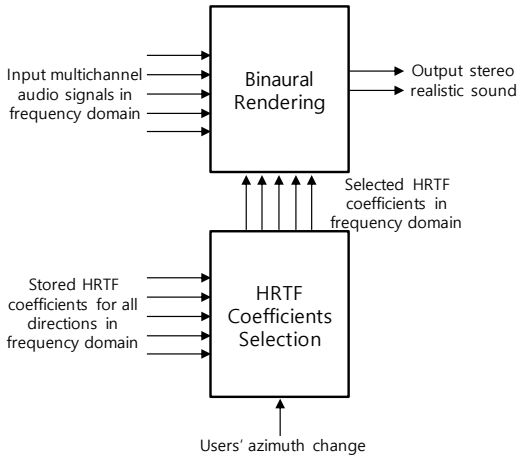

Fig. 4. Overall Procedure of the Sound Scene Control of the Realistic Sound based on the Substitution of the HRTF Coefficients.

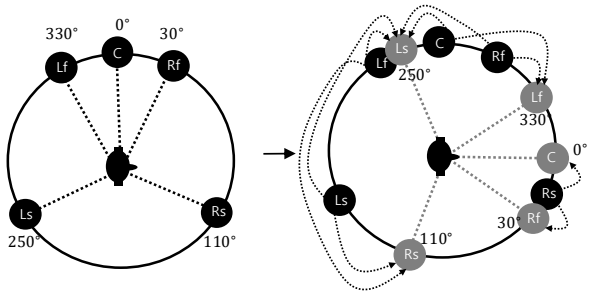

Fig. 5. Example of the Signal Mapping to the Newly Formed 5.1 Multi-Channel Speaker Layout According to the user's Head Movement.

$$\begin{bmatrix} O_L(k) \\ O_R(k) \end{bmatrix} = \begin{bmatrix} H_C^{Left}(k) & H_C^{Right}(k) \\ H_{Lf}^{Left}(k) & H_{Lf}^{Right}(k) \\ H_{Rf}^{Left}(k) & H_{Rf}^{Right}(k) \\ H_{Ls}^{Left}(k) & H_{Ls}^{Right}(k) \\ H_{Rs}^{Left}(k) & H_{Rs}^{Right}(k) \end{bmatrix} \times \begin{bmatrix} S_C^m(k) \\ S_{Lf}^m(k) \\ S_{Rf}^m(k) \\ S_{Ls}^m(k) \\ S_{Rs}^m(k) \end{bmatrix}, \text{ for } 0 \le k \le M-1 \tag{6}$$

where $S_X^m(k)$ is a newly generated signal of any channel X through the mapping of the 5.1 multi-channel audio signals to the newly formed 5.1 multi-channel speaker layout. For the explanation of the signal mapping using the CPP method [14, 15], let's assume that there are two channel speakers (C1 and C2) and any channel (C3) lays in between two channel speakers after the user's head movement as shown in Fig. 6. Then, a signal of channel C3 is mapped onto the channel C1 and C2 using (7) and (8).

$$\theta_{norm} = \frac{(\theta_3 - \theta_1)}{(aperture - \theta_1)} \times \frac{\pi}{2}, \; aperture = |\theta_2 - \theta_1| \tag{7}$$

$$\left.\begin{aligned} S_1^m(k) &= S_1(k) + S_3(k) \times \cos(\theta_m) \\ S_2^m(k) &= S_2(k) + S_3(k) \times \sin(\theta_m) \end{aligned}\right\}, \text{ for } 0 \le k \le M\text{-}1 \tag{8}$$

Here, θ_{norm} is the normalized angle of azimuth of C3 laid in between C1 and C2, and $aperture$ is the reference angle between C1 and C2. θ_1, θ_2 and θ_3 are the azimuth of C1, C2, and C3. $S_1(k)$, $S_2(k)$ and $S_3(k)$ are the signal gains of C1, C2, and C3. $S_1^m(k)$ and $S_2^m(k)$ are the new signal gains of C1 and C2 after the mapping of $S_1(k)$ using the CPP. The signal mapping using the CPP method is applied to the entire 5.1 multi-channel signals to create the new 5.1 multi-channel audio signals with the newly formed 5.1 multi-channel speaker layout according to the user's head movement. Finally, the stereo realistic sound with the controlled sound scene can be generated through the binaural rendering with the new 5.1 multi-channel audio signals and the existing 5.1 multi-channel HRTF coefficients. Fig. 7 shows the overall procedure of the sound scene control of the realistic sound based on the CPP method.

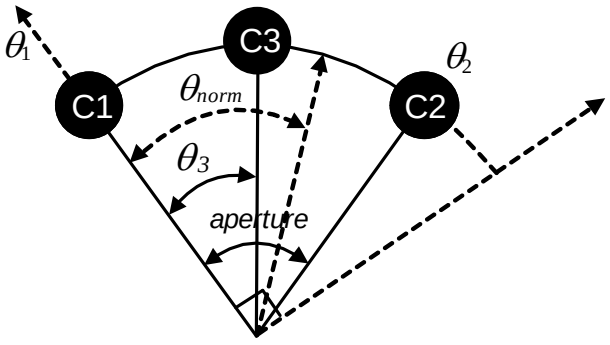

Fig. 6. Example of the Signal Mapping using the CPP.

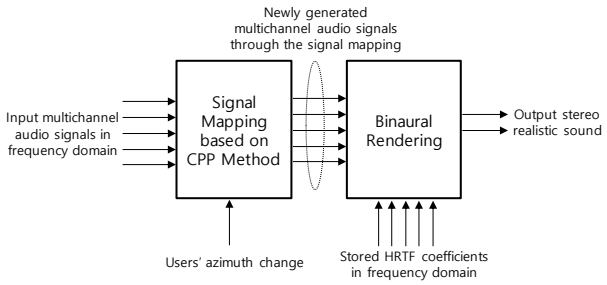

Fig. 7. Overall Procedure of the Sound Scene Control of the Realistic Sound based on the CPP Method.

III. Proposed Stereo Realistic Sound Generation for Free Movement in VR

In the VR service, the user moves freely in the VR space, so the VR sound must be adjusted according to the VR scene. Namely, the VR service allows the user's head movement and the position change in the VR space and the VR sound in the VR service should reflect the user's free movement. However, since the previously explained sound scene control based on the HRTF coefficients and the CPP method only focused on the modification of the stereo realistic sound scene according to the user's azimuth change, the previous realistic sound could not imply the user's distance change. Therefore, there is still a mismatch between the VR scene and the VR sound when the user freely moves in the VR space and the overall performance of the VR service can be severely degraded. To allow the user's free movement in the VR space and reduce the performance degradation of the VR, the realistic sound generation method based on a complex plane for the user's movement tracking is proposed. The user's position change effected on both changes of the distance and the azimuth between the user and the speaker while the user's head movement only effected on the azimuth change between the user and the speaker. Therefore, the proposed method separately handled the user's position change and the head movement. Namely, the distance and the azimuth between the user and the speaker layout by the user's position change were firstly measured, and then the final azimuth between the user and the speaker by considering two azimuths caused by the user's position change and the head movement was determined. The signal level was modified using the calculated distance between the user and the speaker while the sound scene of the realistic sound was controlled using the measured azimuth. The detail of the realistic sound generation using the calculated distance and azimuth between the user and the speaker is given in the below.

For the calculation of the distance and the azimuth between the user and the speaker by the user's position change, it is assumed that the 5.1 multi-channel speaker playback system located in the complex plane as shown in Fig. 8 and the user moved freely on the complex plane. Based on the assumption, the distance and the azimuth between the user and the speaker by the user's position change could be estimated because the user and the speaker were considered as two points in the complex plane. Meanwhile, as the azimuth measurement method could vary according to the user's location on the complex plane, the distance and azimuth between the user and the speaker were measured based on the user's location divided

into four areas around the speaker as shown in Fig. 9 and 10. Meanwhile, the four areas around the speaker in the complex plane are summarized in Table I. After setting four areas for all speakers in the 5.1 multi-channel speaker layout as shown in Fig. 9, the distance and the azimuth between the user and the speaker were calculated in each area as shown in Fig. 10. In Fig. 10, $a + jb$ and $c + jd$ are the position of any speaker X in the 5.1 multi-channel speaker layout and the user in the complex plane, respectively. Table II summarizes the calculation of the distance and azimuth between the user and the speaker for four areas around the speaker according to the user's position change.

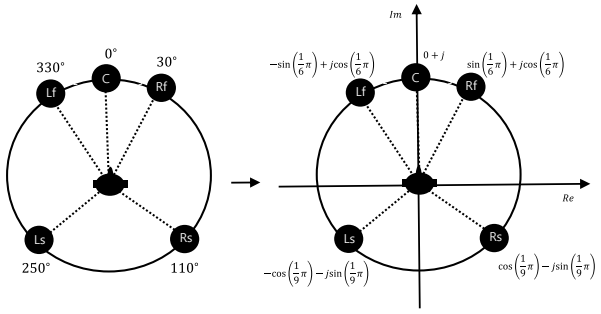

Fig. 8. 5.1 Multi-channel Speaker Placed in the Complex Plane.

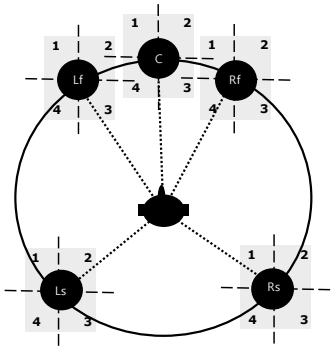

Fig. 9. Four Areas for All Speakers in the 5.1 Multi-channel Speaker Layout.

TABLE I. FOUR AREAS AROUND THE SPEAKER EXPRESSED IN THE COMPLEX PLANE COORDINATE

Area	Complex plane coordinate
Area 1	$Re < 0,\ Im \geq 0$
Area 2	$Re \geq 0,\ Im > 0$
Area 3	$Re > 0,\ Im \leq 0$
Area 4	$Re \leq 0,\ Im < 0$

For the realistic sound generation by reflecting the user's position change and the head movement, the 5.1 multi-channel audio signals were firstly scaled using the measured five distance values between the moved user and the 5.1 multi-channel speaker. Then, the scaled 5.1 multi-channel audio signals were mapped onto the new multi-channel speaker layout using not only the estimated five azimuths between the moved user and the 5.1 multi-channel speaker but also the azimuth according to the user's head movement. Based on the inverse square law that the sound intensity is inversely proportional to the distance from the source [16, 17], the scaled 5.1 multi-channel audio signals were calculated using the estimated five distance values. Moreover, because all the distances between the user and the 5.1 multi-channel speaker layout are equal to one, the scaled 5.1 multi-channel audio signals were calculated as in (9).

$$S_C^s(k) = \frac{1}{r_C} S_C(k),\ S_{Lf}^s(k) = \frac{1}{r_{Lf}} S_{Lf}(k),\ S_{Rf}^s(k) = \frac{1}{r_{Rf}} S_{Rf}(k),$$
$$S_{Ls}^s(k) = \frac{1}{r_{Ls}} S_{Ls}(k),\ S_{Rs}^s(k) = \frac{1}{r_{Rs}} S_{Rs}(k)\ for\ 0 \leq k \leq M-1 \quad (9)$$

where $S_X^s(k)$ is the scaled signal of any channel X and r_X is the estimated distance between the user and any channel X. Because the virtual 5.1 channel speaker layout was set around the moved user, the scaled 5.1 multi-channel audio signals were mapped onto the virtual speaker layout using the measured five azimuths and the azimuth by the head movement. The final azimuths could be determined for the signal mapping as in (10).

$$\left.\begin{aligned} \theta_f^C &= \theta_p^C - \theta_{hm} \\ \theta_f^{Lf} &= \theta_p^{Lf} - \theta_{hm} \\ \theta_f^{Rf} &= \theta_p^{Rf} - \theta_{hm} \\ \theta_f^{Ls} &= \theta_p^{Ls} - \theta_{hm} \\ \theta_f^{Rs} &= \theta_p^{Rs} - \theta_{hm} \end{aligned}\right\} for\ 0 \leq \theta_{hm} \leq 2\pi \quad (10)$$

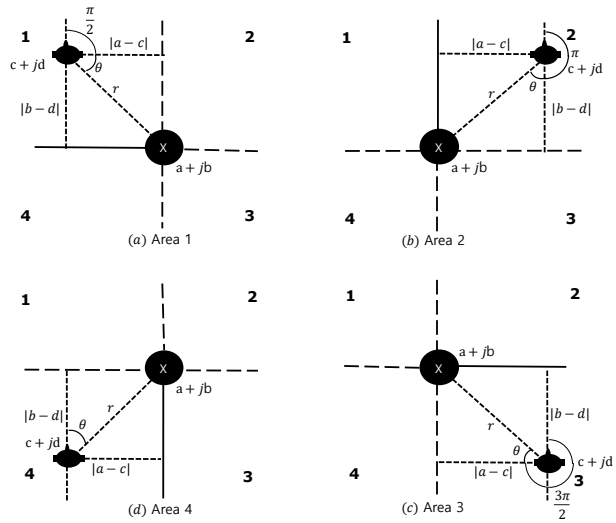

Fig. 10. Calculation of the Distance and the Azimuth in Four Areas.

TABLE II. SUMMARY OF THE CALCULATION OF THE DISTANCE AND THE AZIMUTH BETWEEN THE USER AND THE SPEAKER

Area	Azimuth (θ_p)	Distance (r)
Area 1	$\tan^{-1}\left(\frlac{\lvert b-d\rvert}{\lvert a-c\rvert}\right)+\frac{\pi}{2}$	$\sqrt{(a-c)^2+(b-d)^2}$
Area 2	$\tan^{-1}\left(\frac{\lvert a-c\rvert}{\lvert b-d\rvert}\right)+\pi$	
Area 3	$\tan^{-1}\left(\frac{\lvert b-d\rvert}{\lvert a-c\rvert}\right)+\frac{3\pi}{2}$	
Area 4	$\tan^{-1}\left(\frac{\lvert a-c\rvert}{\lvert b-d\rvert}\right)$	

where θ_f^X is the final azimuth of any channel X for the signal mapping and θ_p^X is the estimated azimuth between the user and the any speaker X according to the user's position change. Here, if θ_f^X has the minus value, θ_f^X is $\theta_f^X + 2\pi$. After the signal mapping using the final azimuth as in (10), the final realistic sound could be obtained to allow the user's free movement including the head azimuth change as in (11).

$$\begin{bmatrix} O_L(k) \\ O_R(k) \end{bmatrix} = \begin{bmatrix} H_C^{Left}(k) & H_C^{Right}(k) \\ H_{Lf}^{Left}(k) & H_{Lf}^{Right}(k) \\ H_{Rf}^{Left}(k) & H_{Rf}^{Right}(k) \\ H_{Ls}^{Left}(k) & H_{Ls}^{Right}(k) \\ H_{Rs}^{Left}(k) & H_{Rs}^{Right}(k) \end{bmatrix} \times \begin{bmatrix} S_C^{s,m}(k) \\ S_{Lf}^{s,m}(k) \\ S_{Rf}^{s,m}(k) \\ S_{Ls}^{s,m}(k) \\ S_{Rs}^{s,m}(k) \end{bmatrix}, \text{ for } 0 \le k \le M-1 \quad (11)$$

where $S_X^{s,m}(k)$ is the generated signal of any channel X of the virtual 5.1 multi-channel speaker layout formed by the user's position and head movement through the signal scaling as in (9) and the signal mapping using the final azimuth as in (10). Fig. 11 shows the overall procedure of the proposed the realistic sound generation for the user's free movement in the VR.

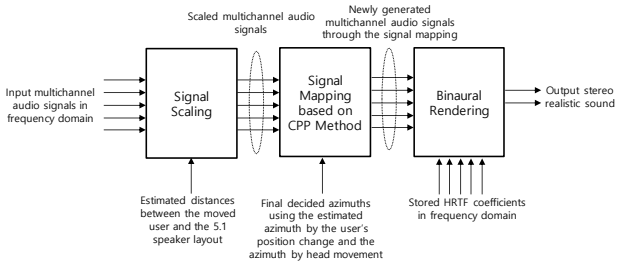

Fig. 11. Overall Procedure of the Proposed the Realistic Sound Generation.

IV. Results and Discussion

To validate the performance of the proposed realistic sound generation method for the user's free movement in the VR, the subjective listening test was performed. Three audio contents were used for the test and listed in Table III. For simplification and clarification of the test, the realistic audio sound only using the left and the right channel signals was separately generated according to the user's position change as shown in Fig. 12. Here, it was assumed that there was no user's head movement. Five listeners participated in the test, and they evaluated the azimuth and distance of the generated realistic audio sound at the changed user's position compared to those of the realistic audio sound at the original position. Meanwhile, the azimuths and distances between the user and the speakers were theoretically calculated in each position as in Table IV.

TABLE III. Test Materials

Material	Description
Item1	Ambience
Item2	Music (back: direct)
Item3	Pathological

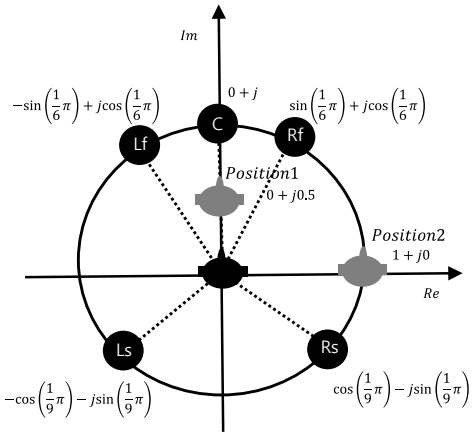

Fig. 12. Two Position Changes for the Test.

TABLE IV. Estimated Distance and the Azimuth between the Given User's Position and the Speaker for the Test

Position	Ch.	Selected Area	Azimuth (θ_p)	Distance (r)
1	C	Area 4	$\tan^{-1}\left(\frac{\lvert 0-0\rvert}{\lvert 1-0.5\rvert}\right)=0$	$\sqrt{(0-0)^2+(1-0.5)^2}=0.5$
	Lf	Area 3	$\tan^{-1}\left(\frac{\lvert\cos(\pi/6)-0.5\rvert}{\lvert-\sin(\pi/6)\rvert}\right)+\frac{3\pi}{2}$ $\approx\frac{30.62\pi}{18}(\approx 306^\circ)$	$\sqrt{(-\sin(\pi/6))^2+(\cos(\pi/6)-0.5)^2}$ ≈ 0.62
	Rf	Area 4	$\tan^{-1}\left(\frac{\lvert\sin(\pi/6)\rvert}{\lvert\cos(\pi/6)-0.5\rvert}\right)$ $\approx\frac{5.38\pi}{18}(\approx 53.8^\circ)$	$\sqrt{(\sin(\pi/6))^2+(\cos(\pi/6)-0.5)^2}$ ≈ 0.62
	Ls	Area 2	$\tan^{-1}\left(\frac{\lvert-\cos(\pi/9)\rvert}{\lvert-\sin(\pi/9)-0.5\rvert}\right)+\pi$ $\approx\frac{22.81\pi}{18}(\approx 228^\circ)$	$\sqrt{(-\cos(\pi/9)-0)^2+(-\sin(\pi/9)-0.5)^2}$ ≈ 1.26
	Rs	Area 1	$\tan^{-1}\left(\frac{\lvert-\sin(\pi/9)-0.5\rvert}{\lvert\cos(\pi/9)\rvert}\right)+\frac{\pi}{2}$ $\approx\frac{13.19\pi}{18}(\approx 132^\circ)$	$\sqrt{(\cos(\pi/9)-0)^2+(-\sin(\pi/9)-0.5)^2}$ ≈ 1.26
2	C	Area 3	$\tan^{-1}\left(\frac{\lvert 1-0\rvert}{\lvert 0-1\rvert}\right)+\frac{3\pi}{2}=\frac{7\pi}{4}(=315^\circ)$	$\sqrt{(0-1)^2+(1-0)^2}\approx 1.41$
	Lf	Area 3	$\tan^{-1}\left(\frac{\lvert\cos(\pi/6)\rvert}{\lvert-\sin(\pi/6)-1\rvert}\right)+\frac{3\pi}{2}$ $=\frac{10\pi}{6}(=300^\circ)$	$\sqrt{(-\sin(\pi/6)-1)^2+(\cos(\pi/6)-0)^2}$ ≈ 1.73
	Rf	Area 3	$\tan^{-1}\left(\frac{\lvert\cos(\pi/6)\rvert}{\lvert\sin(\pi/6)-1\rvert}\right)+\frac{3\pi}{2}$ $=\frac{11\pi}{6}(=330^\circ)$	$\sqrt{(\sin(\pi/6)-1)^2+(\cos(\pi/6)-0)^2}$ $=1$
	Ls	Area 2	$\tan^{-1}\left(\frac{\lvert-\cos(\pi/9)-1\rvert}{\lvert-\sin(\pi/9)\rvert}\right)+\pi$ $=\frac{26\pi}{18}(=260^\circ)$	$\sqrt{(-\cos(\pi/9)-1)^2+(-\sin(\pi/9)-0)^2}$ ≈ 1.97
	Rs	Area 2	$\tan^{-1}\left(\frac{\lvert\cos(\pi/9)-1\rvert}{\lvert-\sin(\pi/9)\rvert}\right)+\pi$ $=\frac{19\pi}{18}(=190^\circ)$	$\sqrt{(\cos(\pi/9)-1)^2+(-\sin(\pi/18)-0)^2}$ ≈ 0.35

Fig. 13 and 14 are the subjective listening test results. In addition, Table V shows the error rate of the proposed method in the azimuth and the distance evaluation, respectively. The test results show that the proposed method could rather successfully generate the realistic sound according to the user's position change because the desired azimuth and distance overlap the confidence intervals of the evaluated ones in most test items. Nevertheless, the confidence intervals of the evaluated azimuth and distance are very wide, so the listening

test results also show that the performance of the proposed method may be rather poor. It is because the proposed method used only the HRTF coefficients of the 5.1 multi-channel speaker layout, namely, the proposed method did not have the sufficient resolution of the HRTF coefficients to generate the realistic audio sound according to the user's free movement in the VR. Therefore, it is necessary to improve the proposed method to generate realistic sound by utilizing HRTF coefficients of the 10.1 or more playback environment.

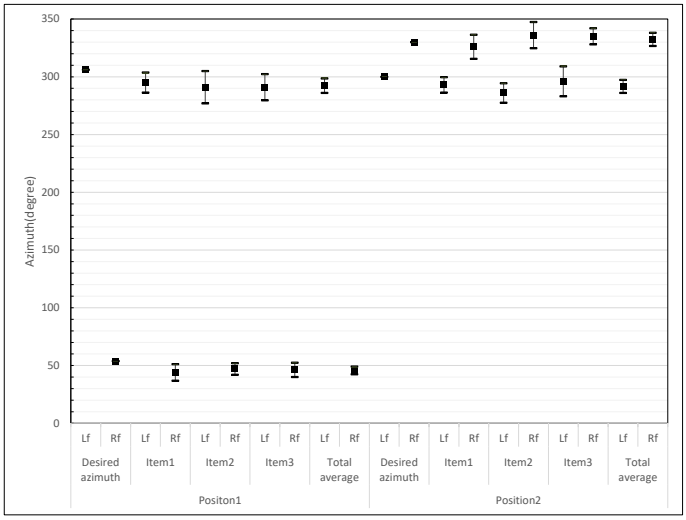

Fig. 13. Subjective Listening Test Result for Evaluation of Azimuth Change.

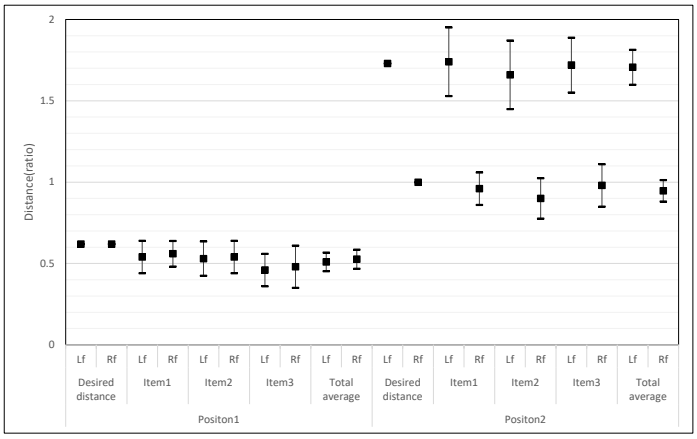

Fig. 14. Subjective Listening Test Result for Evaluation of Distance Change.

TABLE V. ERROR RATE OF THE AZIMUTH AND THE DISTANCE EVALUATION (%)

Position	Channel	Azimuth	Distance
Position1	Lf	4.8	2.3
	Rf	8.1	8.7
Position2	Lf	4.6	1
	Rf	1.2	2.4
Overall		4.7	3.6

V. CONCLUSION

The realistic audio is essential to enjoy the realistic services such as VR, but there is a limitation that multi-channel audio playback environment is involved for the realistic audio. Although the binaural rendering could solve this limitation to provide the realistic sound in the stereo headphone playback environment, there was a problem that the binaural rendering alone could not reflect the user's free movement in the VR. Therefore, there was the mismatch between the visual scene and the audio sound in the VR, so the performance of the VR was degraded. In this paper, the complex plane based stereo realistic sound generation method was proposed to allow the user's free movement such as the position change and the head azimuth change in the VR. In the proposed method, the variations of the azimuth and distance between the user and the speaker were reflected according to the user's movement in the stereo realistic sound generated by the binaural rendering. The subjective listening test results showed that the proposed method could generate the realistic audio sound that successfully reflected the user's free movement only with less than 5 % error rate of the azimuth and the distance evaluation. In spite of the good performance of the proposed method, the performance improvement of the proposed method through the increase of the resolution of the HRTF coefficients remains as a future work because the proposed method only had the HRTF coefficients of the 5.1 multi-channel speaker layout and it caused the error of the azimuth and the distance evaluation.

REFERENCES

[1] B. Gardner and K. Martin, HRTF Measurements of a KEMAR Dummy Head Microphone, MIT Media Lab Perceptual Computing -technical Report #280, 1994.

[2] J. Breebaart et al., "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] J. Breebaart, L. Villemoes, K. Kjörling, "Binaural rendering in MPEG Surround," EURASIP Journal on advances in signal processing, 2008, pp. 1-14.

[4] K. Kim, 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, 2014, pp. 525-526.

[5] K. Kim, "A study on complexity reduction of binaural decoding in multi-channel audio coding for realistic audio service," Contemporary Engineering Sciences, Vol. 9, no. 1, pp. 11-19, 2016.

[6] W. Bailey, B. Fazenda, "The effect of visual cues and binaural rendering method on plausibility in virtual environments," In Audio Engineering Society Convention 144. Audio Engineering Society, 2018.

[7] M. Zaunschirm, M. Frank, F. Zotter, "Binaural rendering with measured room responses: First-order ambisonic microphone vs. dummy head," Applied Sciences, Vol.10, no. 5, 2020.

[8] W. Li, L. Yang, X. Leng, W. Liu, G. Dai, "Application of virtual reality technology in Wushu education," the International Journal of Electrical Engineering & Education, doi.org/10.1177/0020720920940583, 2020.

[9] H. Jialiang, Z. huiying, "Mobile-based education design for teaching and learning platform based on virtual reality," International Journal of Electrical Engineering & Education, doi.org/10.1177/0020720920928547, 2020.

[11] G. Zhang, "Design of virtual reality augmented reality mobile platform and game user behavior monitoring using deep learning," International Journal of Electrical Engineering & Education, doi.org/10.1177/0020720920931079, 2020.

[12] 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.

[13] K Kim, J Kim, "A Study on Realistic Audio Sound Generation according to User's Movement in Virtual Reality System," Proceedings of the 2019 4th International Conference on Intelligent Information Technology, ACM, 2019.

[14] E. Zwicker and H. Fastl, Psychoacoustics, Springer-Verlag, Berlin, Heidelberg, 1999.

[15] V. Pulki, "Virtual sound source positioning using vector base amplitude panning," Journal of Audio Engineering Society, vol. 45, pp. 456-466, 1997.

[16] 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.

[17] Sound intensity as a function of distance from a small source, https://ocw.upc.edu/webs/42254/Acustica_EN/Bloc2/Fitxes/T07_05_Intensitat_i_distancia.htm.

[18] Inverse-Square law, https://en.wikipedia.org/wiki/Inverse-square_law