A Study on Realistic Audio Sound Generation according to User's Movement in Virtual Reality System

Kwangki Kim
Korea Nazarene University
Wolbongro 48, Seobukgu
Cheonan 31172, Korea
k2kim@kornu.ac.kr

Kiho Kim
Korea Nazarene University
Wolbongro 48, Seobukgu
Cheonan 31172, Korea
isk01259@gmail.com

Jinsul Kim
Chonnam National University
Yongbongro 77, Bukgu
Gwangju 61186, Korea
jsworld@jnu.ac.kr

ABSTRACT

Generally, a virtual reality (VR) system cannot provide the realistic sound formed by the multi-channel audio signals to a user with the stereo headphone environment. In addition, the VR system has a gap between the visual scene and the sound because it supplies the audio signal having only a constant sound scene without respect to the change of the user's position. To solve these problems, we introduce the sound scene control of binaural sound. Binaural sound is a stereo realistic sound that can be generated by convolving a 10.1 channel audio signal with a head related transfer function (HRTF) coefficients which features all the paths from the multi-channel speaker layout to the ears in free space. However, since a binaural sound is generated using a fixed multi-channel layout and HRTF coefficients, the binaural sound has a constant sound scene and cannot reflect the user's movements. So, we apply the sound scene control scheme that modifies the binaural sound to allow the user's movement in the VR system. Initially, a multichannel layout is re-created according to the user's azimuth change and the original multi-channel signal is mapped to a new ultra multi-channel layout using a constant power panning law. Secondly, the sound level of the new multichannel audio signal is controlled by the user's distance change, using the characteristics of the sound level inversely proportional to distance. Finally, the final multi-channel audio signal is convolved with the HRTF coefficients to produce a binaural sound with the controlled sound scene. As a result, the proposed realistic audio sound generation method allows the current VR system to provide true VR service without distinction between the visual scenes and sounds.

Keywords

Virtual Reality; Realistic Audio Sound; HRTF; Sound Scene Control; Ultra Multi-channel Audio

1. INTRODUCTION

Generally, since the audio in the virtual reality (VR) service is provided based on the stereo headphone environment, it is not possible to provide a realistic sound by the 10.1 or more multi-channel audio signals so that there is a gap between a scene and a sound in the VR service. Therefore, we adopted the binaural rendering to provide the realistic audio sound through the stereo headphone environment in the VR service. The binaural rendering is a traditional useful scheme to generate the realistic audio sound by convolving the multi-channel audio signals with the head related transfer function (HRTF) coefficients which are responses that characterizes how an ear receives a sound from a point in space [1-3]. But, since the linear convolution of the binaural rendering in time domain has very high computational complexity, it cannot be implemented in the real time. Therefore, we implemented the binaural rendering in frequency domain and separately performed the binaural rendering in two frequency regions – low and high frequency regions [4]. In addition, in the present stereo headphone-based VR service, the audio signal having only a constant sound scene is always provided regardless of a change in user position (azimuth, distance). Consequently, there is another gap between the sound and the scene in the VR service. To solve this problem, we proposed the sound scene control (SCC) scheme of the binaural sound for the user's azimuth change [5] and the sound level modification for the user's distance change. The proposed method enables the real VR service without the gap between the visual scene and the sound scene despite of the user's free movement in the VR system. This paper consists of as follows. In chapter 2, we describe the simple realistic sound generation method for VR system and its' complexity reduction. In chapter 3, we proposed the realistic sound generation method according to the user's azimuth and distance change. In chapter 4, we give the conclusion and the future work.

2. REALISTIC SOUND GENERATION FOR VR SYSTEM

Traditionally, the multi-channel audio signals are simply down-mixed into the stereo signal and delivered to the user with the stereo headphone or speaker layout. So, the stereo down-mix signal rarely has the multi-channel sound effect and the user cannot experience the realistic audio sound effect by the stereo down-mix signal. To provide the user with the realistic audio sound formed by 10.1 or more multi-channel audio signals through the stereo headphone, as shown in Figure1 and 2, the binaural sound should be generated using HRTF coefficients that characterizes how an ear receives a sound from a point in space. For this purpose, we need to measure the HRTF coefficients, which represents the signal path from each

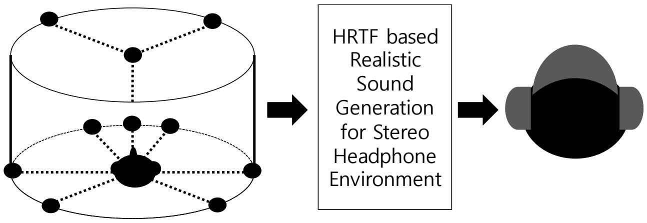

Figure 1. Concept of realistic audio sound generation for VR system

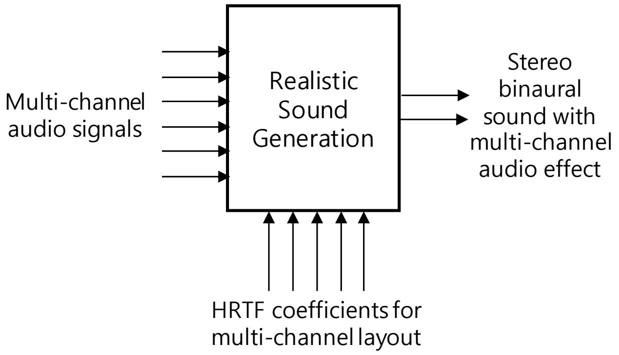
Multi-channel audio signals
Realistic Sound Generation
Stereo binaural sound with multi-channel audio effect
HRTF coefficients for multi-channel layout

Figure 2. HRTF based Realistic audio sound generation for VR system

speaker of 10.1 or more reproduction system to the human ear. By convolving the multi-channel audio signals with the measured HRTF coefficients, it is possible to generate realistic sounds for the stereo headphone environment. If there are N channels reproduction systems, we need N HRTF coefficients representing all signal paths from N channels to the human's left and right ear. Using all HRTF coefficients, the realistic audio sound for the stereo headphone can be calculated as (1).

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

where o_L and o_R are the generated left and right realistic signals in time domain and s_n is the nth channel signal in time domain. h_n^L and h_n^R are the HRTF coefficients from the nth channel to human left and right ear and $\otimes$ is the linear convolution. The linear convolution of the multi-channel audio signals and the HRTF coefficients in time domain cannot be implemented in the real time due to very high computational complexity so that the realistic audio sound should be generated by multiplication in frequency domain as (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 O_L and O_R are the generated left and right realistic signals in frequency domain and S_n is the nth channel signal in frequency domain. H_n^L and H_n^R are the HRTF coefficients from the nth channel to human left and right ear in frequency domain. Also, by using the characteristics that human ears are sensitive to low frequency regions and insensitive to high frequency regions [6], the realistic audio sounds can be separately calculated using (3).

$$O_L = \sum_{n=1}^{N}\left(S_n \cdot H_n^L\right),\ O_R = \sum_{n=1}^{N}\left(S_n \cdot H_n^R\right),\ \text{for low frequency regions}$$
$$O_L = \sum_{n=1}^{N}\left(S_n \cdot \left|H_n^L\right|\right),\ O_R = \sum_{n=1}^{N}\left(S_n \cdot \left|H_n^R\right|\right),\ \text{for high frequency regions} \tag{3}$$

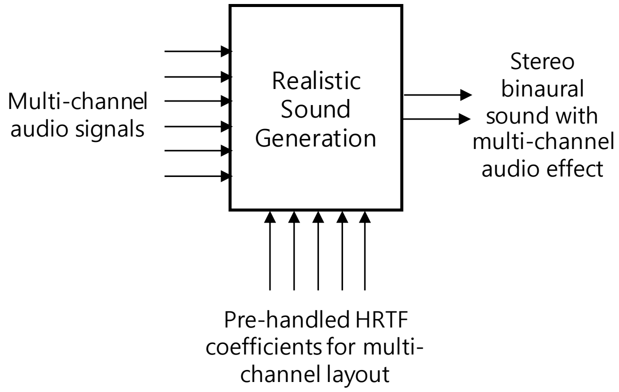
Multi-channel audio signals
Realistic Sound Generation
Stereo binaural sound with multi-channel audio effect
Pre-handled HRTF coefficients for multi-channel layout

Figure 3. Realistic sound generation using pre-handled HRTF coefficients for complexity reduction

In (3), the realistic audio sounds for the low frequency regions are calculated by modifications of both amplitude and phase information while those for the high frequency regions are estimated by modification of only amplitude information. It is because the human ears are insensitive to phase information in the high frequency regions. So, we can generate the realistic audio sound with the same audio quality compared to (2) while minimizing the computational complexity. Consequently, compared to the complexity of (1), (3) can achieve very high reduction rate to be more than 90 %.

3. REALISTIC SOUND GENERATION ACCORDING TO USER'S AZIMUTH AND DISTANCE CHANGE FOR VR SYSTEM

3.1 Realistic Sound Generation according to User's Azimuth Change

In the present stereo headphone-based virtual reality system, the audio signals having only a constant sound scene are played regardless of a change in user's position (azimuth angle and distance). Therefore, there is a gap between visual scenes and sounds in the VR system. To solve this problem, we firstly introduce the SCC of the realistic audio sound generated by the binaural rendering. There are two kinds of the SCC methods. The first method is based on the change of the HRTF coefficients and the second one is based on the constant power panning (CPP) law [7], [8].

At the first method, the original HRTF coefficients are substituted by the new HRTF coefficients corresponding to the user's azimuthal change and the new HRTF coefficients are convolved with the multi-channel audio signals for generating the realistic audio sound with the controlled sound scene. Accordingly, (2) and (3) are updated as followings.

$$O_L = \sum_{n=1}^{N}\left(S_n \cdot H_{n+\theta}^L\right),\ O_R = \sum_{n=1}^{N}\left(S_n \cdot H_{n+\theta}^R\right) \tag{4}$$

$$O_L = \sum_{n=1}^{N}\left(S_n \cdot H_{n+\theta}^L\right),\ O_R = \sum_{n=1}^{N}\left(S_n \cdot H_{n+\theta}^R\right),\ \text{for low frequency regions}$$
$$O_L = \sum_{n=1}^{N}\left(S_n \cdot \left|H_{n+\theta}^L\right|\right),\ O_R = \sum_{n=1}^{N}\left(S_n \cdot \left|H_{n+\theta}^R\right|\right),\ \text{for high frequency regions} \tag{5}$$

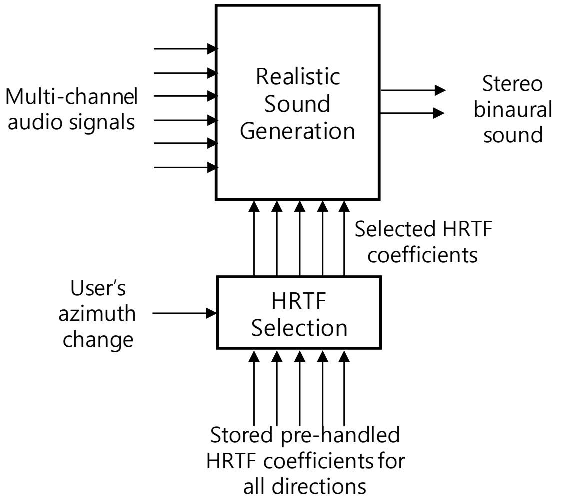
Multi-channel audio signals
Realistic Sound Generation
Stereo binaural sound
Selected HRTF coefficients
User's azimuth change
HRTF Selection
Stored pre-handled HRTF coefficients for all directions

Figure 4. Realistic sound generation using the stored HRTF coefficients for all directions

Here, θ is the user's azimuth change, and $H_{n+\theta}^{L}$ and $H_{n+\theta}^{R}$ are the new left and right HRTF coefficients of the nth channel corresponding to the user's movement, respectively. In the first method, the HRTF coefficients for 360 degrees all directions should be measured using the dummy head in the anechoic room and stored in the memory so that the first method has a constraint that it cannot be implemented in the embedded environment with low memory storage. We propose the second method to solve the memory problem of the first method.

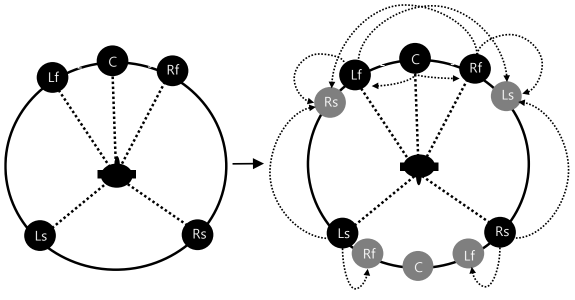

Figure 5. Multi-channel audio signals mapping according to user's azimuth change

At the second method, the original HRTF coefficients are fixed and the multi-channel audio signals are mapped onto the newly figured multi-channel layout according to the user's movement as shown in Figure 5. Then, the realistic audio sound is generated by convolving the new multi-channel audio signals with the original HRTF coefficients. Figure 6 shows the whole procedure of the realistic sound generation using the SCC. Since the HRTF coefficients for the original multi-channel layout are only needed, the memory problem at the first method can be solved. Then, (2) and (3) are also updated as followings.

$$O_L=\sum_{n=1}^{N}\left(S_n^{'}\cdot H_n^{L}\right),\ O_R=\sum_{n=1}^{N}\left(S_n^{'}\cdot H_n^{R}\right)\qquad(6)$$

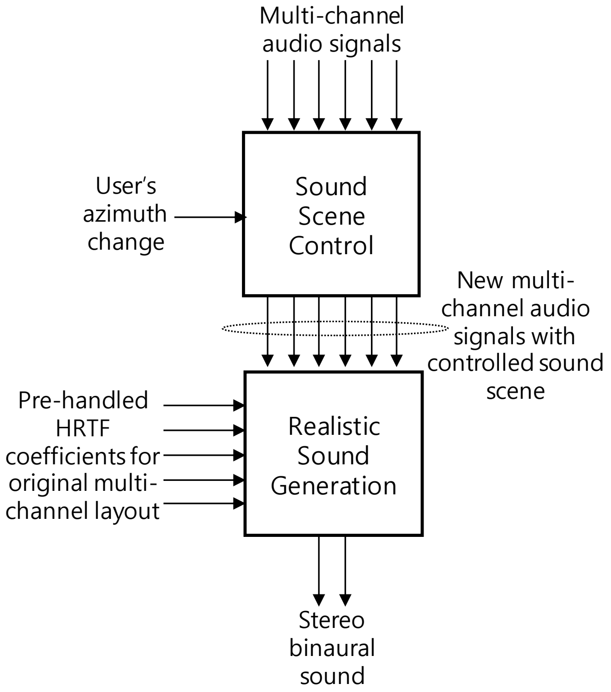
Multi-channel audio signals
User's azimuth change
Sound Scene Control
New multi-channel audio signals with controlled sound scene
Pre-handled HRTF coefficients for original multi-channel layout
Realistic Sound Generation
Stereo binaural sound

Figure 6. Realistic sound generation using sound scene control

$$O_L=\sum_{n=1}^{N}\left(S_n^{'}\cdot H_n^{L}\right),\ O_R=\sum_{n=1}^{N}\left(S_n^{'}\cdot H_n^{R}\right),\ \text{for low frequency regions}$$

$$O_L=\sum_{n=1}^{N}\left(S_n^{'}\cdot\left|H_n^{L}\right|\right),\ O_R=\sum_{n=1}^{N}\left(S_n^{'}\cdot\left|H_n^{R}\right|\right),\ \text{for high frequency regions}\qquad(7)$$

Here, $S_n^{'}$ is a new signal generated by mapping the original multi-channel signals to the new configuration multi-channel layout. As shown in Figure 7, if there is a channel 3 signal between channel 1 and 2 which are newly formed according to the user's movement, the channel 3 signal is mapped to the channel 1 and the channel 2 through CPP to generate new channel 1 signal and channel 2 signal as followings.

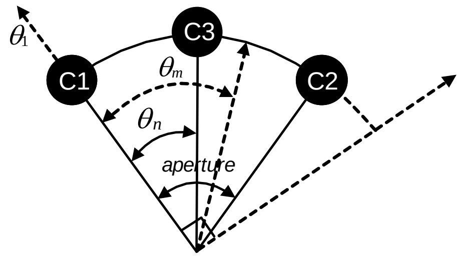

Figure 7. Example of channel signal mapping using CPP technique

$$\theta_m=\frac{\left(\theta_n-\theta_1\right)}{\left(aperture-\theta_1\right)}\times\frac{\pi}{2}\qquad(8)$$

$$S_{C1} = S_{C3} \times \cos\left(\theta_m\right),\ S_{C2} = S_{C3} \times \sin\left(\theta_m\right) \qquad (9)$$

Here, θ_m is the normalization of the azimuth angle (θ_n) of the channel 3 located between the channel 1 and the channel 2 at 90 degrees. θ_1 is the azimuth angle of the channel 1 and $aperture$ is the angle between the channel 1 and the channel 2. S_{C1} and S_{C2} are newly generated signals of the channel 1 and the channel 2 through mapping of existing channel 3 signal (S_{C3}) using the CPP. Consequently, we can generate the realistic audio sound reflecting the user's azimuth change without the additional HRTF coefficients.

3.2 Realistic Sound Generation according to User's Azimuth and Distance Change

To calculate the final realistic audio sound corresponding to the user's azimuth and distance changes, the newly generated multi-channel audio signals as previously described above are firstly adjusted according to the user's distance change and the adjusted multi-channel audio signals are convolved with the fixed HRTF coefficients. Therefore, (6) and (7) are just updated as followings.

$$O_L = \sum_{n=1}^{N}\left(\alpha_n \cdot S_n^{'} \cdot H_n^L\right),\ O_R = \sum_{n=1}^{N}\left(\alpha_n \cdot S_n^{'} \cdot H_n^R\right) \quad (10)$$

$$O_L = \sum_{n=1}^{N}\left(\alpha_n \cdot S_n^{'} \cdot H_n^L\right),\ O_R = \sum_{n=1}^{N}\left(\alpha_n \cdot S_n^{'} \cdot H_n^R\right),\ \text{for low frequency regions}$$
$$O_L = \sum_{n=1}^{N}\left(\alpha_n \cdot S_n^{'} \cdot \left|H_n^L\right|\right),\ O_R = \sum_{n=1}^{N}\left(\alpha_n \cdot S_n^{'} \cdot \left|H_n^R\right|\right),\ \text{for high frequency regions} \quad (11)$$

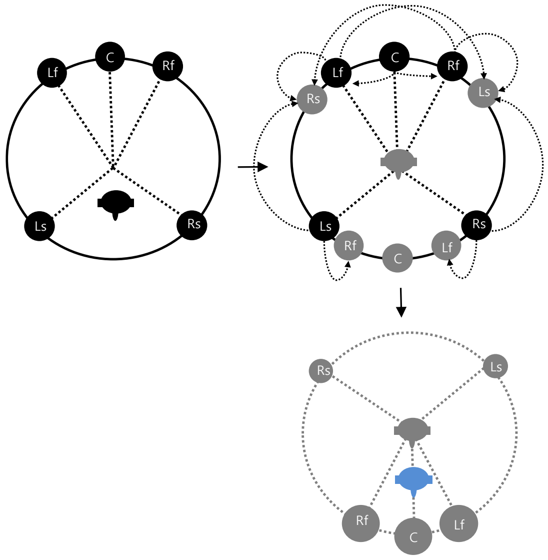

Figure 8. Multi-channel audio signals mapping and level modification according to user's azimuth and distance change

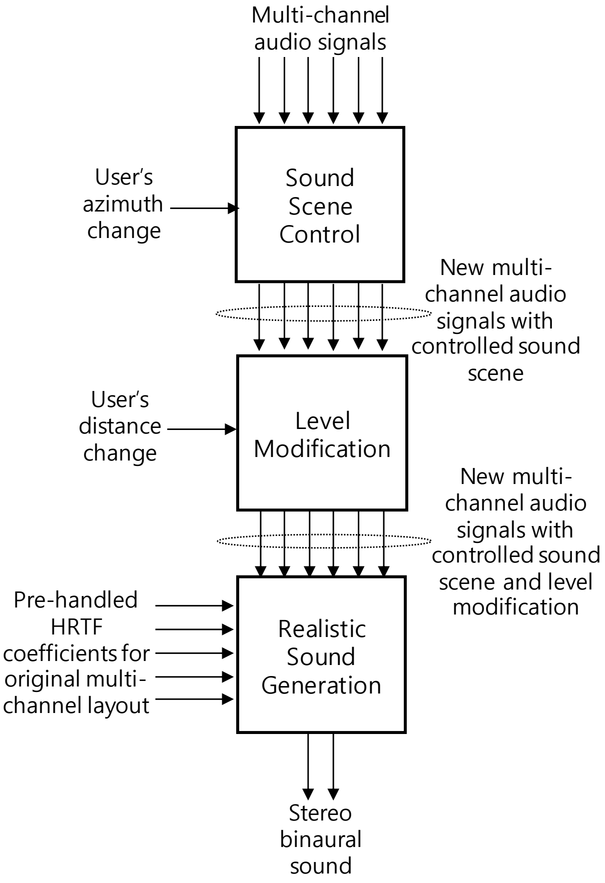
Multi-channel audio signals
User's azimuth change
Sound Scene Control
New multi-channel audio signals with controlled sound scene
User's distance change
Level Modification
New multi-channel audio signals with controlled sound scene and level modification
Pre-handled HRTF coefficients for original multi-channel layout
Realistic Sound Generation
Stereo binaural sound

Figure 9. Realistic sound generation using sound scene control and level modification

Here, α_n is a scale factor for reflecting the distance change of the nth channel and if d_n is the distance change between the newly figured multi-channel layout and the user, α_n is calculated as following.

$$\alpha_n = \begin{cases} \dfrac{1}{\left(1+d_n\right)} & \text{if a user is going to be far away from the n}^{\text{th}} \text{ channel} \\ \left(1+d_n\right) & \text{if a user is going to be close to the n}^{\text{th}} \text{ channel} \end{cases} \quad (12)$$

4. CONCLUSION

In this paper, we presented a technology for generating realistic audio sound that can be applied to the VR service. To apply the realistic audio sound by 10.1 channel or more multi-channel audio signal to the VR service, we proposed the technology to convert the multi-channel audio signals into the realistic sound for the stereo headphone environment and the sound scene control and level modification techniques to generate the realistic audio sound according to the user's position (azimuth, distance) changes. As future works, the subjective listening test will be performed to verify the performance of the proposed schemes and we will apply the proposed schemes to the VR service through the performance improvement.

ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2017R1D1A3B03034951). This research was supported by the National Research Foundation of Korea(NRF) funded by the MSIT(NRF-2018R1A4A1025559). Also, this research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program(IITP-2018-2016-0-00314) supervised by the IITP (Institute for Information & communications Technology Promotion).

5. 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, Jeroen, 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] Kwangki Kim and Jinsul 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] Kwangki 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.

Columns on Last Page Should Be Made As Close As Possible to Equal Length

Authors’ background

Your Name	Title*	Research Field	Personal website
Kwangki Kim	Assistant professor	Audio signal processing	None
Kiho Kim	Bachelor student	Audio signal processing	None
Jinsul Kim	Associate professor	Audio signal processing, Network	None