International Journal of Internet, Broadcasting and Communication Vol.13 No.1 161-167 (2021)
http://dx.doi.org/10.7236/IJIBC.2021.13.1.161

IJIBC 21-1-19

A DNN-Based Personalized HRTF Estimation Method for 3D Immersive Audio

Ji Su Son*, Seung Ho Choi**

*, **Dept. of Electronic and IT Media Engineering, Seoul National University of Science and Technology, Seoul, Korea
*wltn9587@hanmail.net, **shchoi@snut.ac.kr

Abstract

This paper proposes a new personalized HRTF estimation method which is based on a deep neural network (DNN) model and improved elevation reproduction using a notch filter. In the previous study, a DNN model was proposed that estimates the magnitude of HRTF by using anthropometric measurements [1]. However, since this method uses zero-phase without estimating the phase, it causes the internalization (i.e., the inside-the-head localization) of sound when listening the spatial sound. We devise a method to estimate both the magnitude and phase of HRTF based on the DNN model. Personalized HRIR was estimated using the anthropometric measurements including detailed data of the head, torso, shoulders and ears as inputs for the DNN model. After that, the estimated HRIR was filtered with an appropriate notch filter to improve elevation reproduction. In order to evaluate the performance, both of the objective and subjective evaluations are conducted. For the objective evaluation, the root mean square error (RMSE) and the log spectral distance (LSD) between the reference HRTF and the estimated HRTF are measured. For subjective evaluation, the MUSHRA test and preference test are conducted. As a result, the proposed method can make listeners experience more immersive audio than the previous methods.

Keywords: *3D immersive audio, Head-related transfer function (HRTF), Head-related impulse response (HRIR), Deep Neural Network (DNN), notch filter, personalization, Anthropometric measurement*

1. Introduction

In recent years, technologies of immersive media, virtual reality (VR) and augmented reality (AR) have rapidly developed. The significant component for implementing these technologies is 3D immersive audio [2]. The spatial sound from stereo headphones uses head-related transfer function (HRTF). The HRTF is defined as a three-dimensional function from the sound source to the entrance of the ear canal in the frequency domain. Since the HRTFs are very sensitive to anthropometric measurements, each individual has a different HRTFs [3]. However, it is inefficient to measure the individual HRTFs in terms of cost and time. To overcome this problem, mathematical HRTF estimation method based on the measured HRTF have been studied [4,5]. Recently, studies based on deep learning have also conducted to explore the relationship between anthropometric measurements and HRTFs [6,7]. A deep neural network (DNN) model was proposed to estimate the magnitude of the personalized HRTFs by using anthropometric measurements [1]. However, when

Manuscript Received: December. 29, 2020 / Revised: January. 4, 2021 / Accepted: January. 9, 2021
Corresponding Author: shchoi@seoultech.ac.kr
Tel: +82-2-970-6461, Fax: +82-2-979-7903
Professor, Department of Electronic and IT Media Engineering, Seoul National University of Science and Technology, Korea

the sound is synthesized only with the magnitude of the HRTF, it causes the problem of internalization that a sound image is formed inside-head when listening to the spatial sound. Therefore, we devised a new DNN model that can estimate the phase as well as the magnitude of the HRTF.

A previous study refers to one of the prominent features observed in the HRTF which is spectral notches [8]. This notch appears to be generated by time-delayed reflections off the posterior concha wall interfering with sound directly entering the external auditory canal [9]. In addition, previous study has found that the location of the HRTF spectral notch appears at high frequencies as the elevation increases and it has been shown to be important for elevation perception [8]. A method of improving elevation reproduction using this feature was explored through various experiments.

2. Proposed Personalized HRTF Estimation Method

2.1. The CIPIC database and preprocessing

To estimate personalized HRTF, the public HRTF database was provided by The Center for Image Processing and Integrated Computing (CIPIC) of The University of California at Davis [10]. This database included head-related impulse responses (HRIRs), for 45 subjects at 25 different azimuths and 50 different elevations with anthropometric measurements and ear images of each subject. The anthropometric measurements in the CIPIC HRTF database includes data from a total of 45 subjects, each consisting of 17 parameters for head and torso measurements and 10 for pinna measurements.

Since the range of the anthropometric data of 35 people differs between the data, it was normalized using the mean and variance for all training data and used as an input feature for the DNN training model [7]. For normalization, the following equation is used.

$$\overline{z_i} = (1 + e^{-\frac{(z_i - \mu_i)}{\sigma_i}})^{-1} \quad (1)$$

Where z_i is the i-th feature and $\overline{z_i}$ are used as the normalized input features, respectively. And μ_i is the mean and σ_i is the standard deviation of the i-th features. Note that z_i could be the anthropometric measurements in the CIPIC HRTF database which is head, torso and pinna measurements.

2.2. A DNN-based 3D immersive audio reproduction model

The proposed DNN model for estimating personalized HRTF is shown in Figure 1.

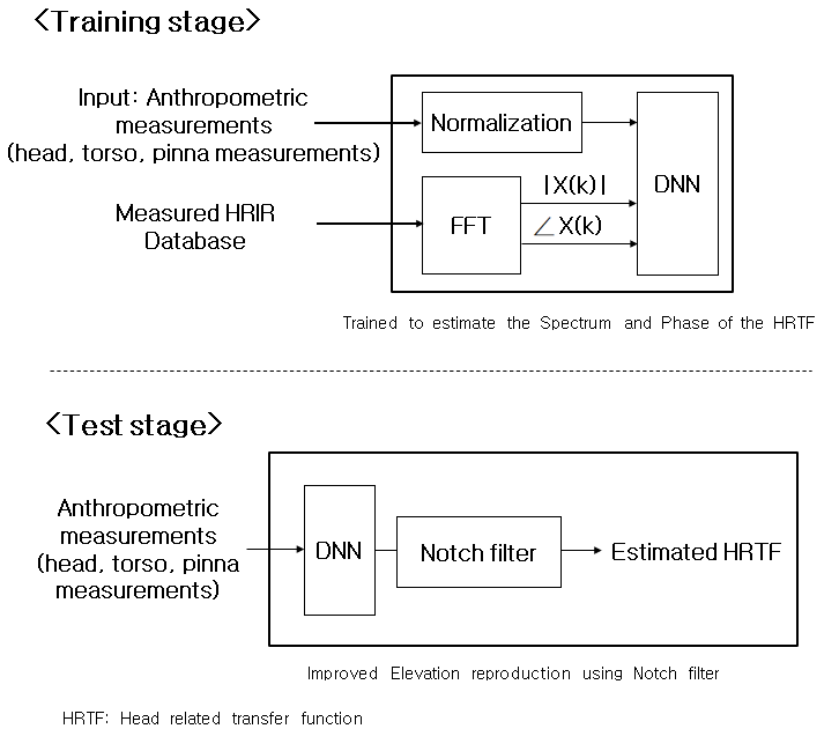

Figure 1. A proposed DNN model

Anthropometric measurement is input features, and the magnitude and phase of HRTF are target features. Two deep neural networks estimate the magnitude and phase of HRTF, respectively. Since HRIR was easy to apply to speech signals, the HRIR was extracted by inverse Fourier transform (IFFT) of the magnitude and phase of the HRTF, which are the output of the DNN model.

After applying the estimated HRTF to synchronize an immersive audio, an appropriate notch filter is used to improve the elevation reproduction performance. One of the prominent features observed in the HRIR including elevation is a spectral notch [8]. This phenomenon occurs because when the sound enters human ears, reflection occurs due to the structure of the ear [9]. It can be observed that the position of this spectral notch is located on the high frequencies as the elevation increases.

Through experiments, it was found that when the amplitude attenuation of the frequency band where the spectral notch is located becomes larger, it is effective in reproducing elevation. Hence, a notch filter was designed according to the HRIR spectral notch position and applied to the HRIR or sound source to improve the elevation reproduction performance. Figure 2 shows an example of the spectrum before and after applied with a notch filter. It shows the spectrum applied with a notch filter to remove the frequency band of the spectral notch observed in the reference HRIR.

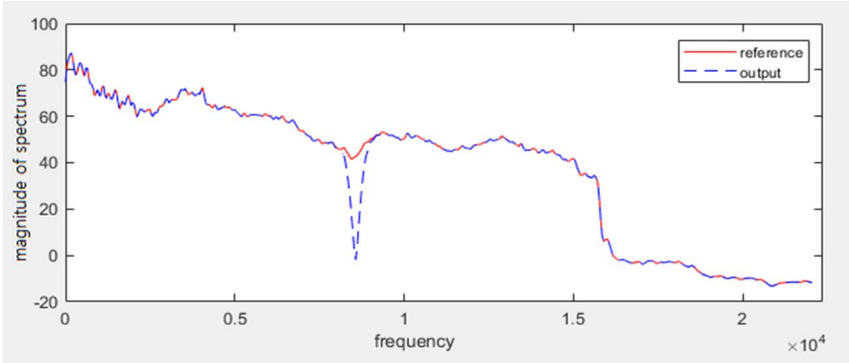

Figure 2. An example of 3D sound spectrum before and after applied notch filter [azimuth angle: 0°, elevation angle: 45° (left ear, subject2), (-: reference, --: output)]

3. Performance Evaluation

The two deep neural networks that estimate the magnitude and phase of the HRTF have the same structure, each consisting of 1 input layer, 5 hidden layers, and 1 output layer. There are 37 input units which are 17 parameters for head and torso measurements and 20 for pinna measurements for both ears. The output nodes are set to 129 to match the magnitude and phase length of the HRTF. The number of nodes of the hidden layer was set to 64, and ReLU (rectified linear unit) was applied as the activation function of each layer.

3.1 Objective evaluation

As mentioned earlier, the RMSE and LSD between the reference HRTF and the estimated HRTF were measured for the objective evaluation. The RMSE which is defined as Eq. (2) was compared between HRIR data.

$$\mathrm{RMSE}(\mathrm{y},\hat{y}) = 20\log_{10}\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}(y(n)-\hat{y}(n))^2} \tag{2}$$

where N is the total length of the HRTF. $y(n)$ means reference HRIR and $\hat{y}(n)$ means the estimated HRIR. The LSD was defined as

$$\mathrm{LSD}(\mathrm{Y},\hat{Y}) = \sqrt{\frac{1}{M}\sum_{k=0}^{M-1}(20\log_{10}\frac{|Y(k)|}{|\hat{Y}(k)|})^2} \tag{3}$$

where $|\mathrm{Y}(k)|$ and $|\hat{Y}(k)|$ are reference and estimated HRTF magnitudes, which were obtained by applying a fast Fourier transform (FFT) to $y(n)$ and $\hat{y}(n)$, respectively. In addition, M is half of the FFT size.

Table 1 shows the comparison results of the LSDs of magnitude, RMSEs of phase and RMSEs of HRIR between reference HRIR and estimated HRTF of the proposed method (Azimuth angle:45°, Elevation angle:

0°). It shows that the LSDs between the reference HRIR and estimated HRTF are as low as about 1. RMSEs of phase and HRIR show low values as well. It can be confirmed that the estimated HRIR has a very similar value to the reference HRIR because the RMSE value of the HRIR by synthesized the estimated magnitude and the estimated phase is close to 0.

Table 1. Comparison of LSDs and RMSEs between reference HRIR and the proposed HRTF

Estimated HRTF (Left ear, Azimuth angle: 45°)						
Method	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Mean
LSD of Magnitude	1.6003	2.0838	1.1185	1.8258	1.4968	1.6250
RMSE of Phase	1.4229	1.4451	2.0767	5.8912	8.5881	3.8848
RMSE of HRIR	0.0375	0.0293	0.0250	0.0346	0.0433	0.0339
Estimated HRTF (Right ear, Azimuth angle: 45°)						
Method	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Mean
LSD of magnitude	1.8440	0.8110	1.5803	1.1287	0.9934	1.2715
RMSE of Phase	2.0138	1.6759	2.3343	1.6545	2.7091	2.0775
RMSE of HRIR	0.1210	0.0619	0.1541	0.0919	0.1821	0.1222

In addition, at the azimuth angle -80° and elevation angle 0°, the mean values of the LSD of magnitude, RMSE of phase and RMSE of HRIR were 1.432, 4.952, 0.0986, respectively.

3.2. Subjective evaluation

The previous method estimates only the magnitude of HRTF based on a DNN using anthropometric data, and obtains a personalized HRTF using zero-phase [2]. We proposed a new method estimates not only the magnitude of HRTF, but also the phase of the HRTF by a DNN model. Through an objective evaluation in 3.1, it was confirmed that the proposed method estimates a personalized HRTF very similar to the reference HRIR.

We carried out subjective evaluations as well. One was the MUSHRA test and preference test was conducted for 5 subjects. The subjects listened to 3D sound applied with the previous method(zero-phase), proposed method (estimating both of the HRTF magnitude and phase) and the reference of HRIRs which were produced by the azimuth angle moving to -80°→15°→45°. After listening, subjects rated the previous and proposed one compared to the reference one, respectively. (from lowest -5 to the highest 0)

Table 2. MUSHRA test result

Method	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Mean
Previous method (zero-phase)	-3	-5	-4	-3	-4	-3.8
Proposed method (Estimated phase)	0	-0.5	0	0	0	-0.1

The preference test was conducted as the second listening experiment. It is a test in which subjects chose what they prefer after listening to the 3D sound applied with the previous and proposed method. All of subjects answered that the 3D sound applied of proposed method was better than the previous one.

In these two listening tests, all listeners answered that the 3D sound which was applied of proposed method was very similar to the reference 3D sound, and preferred the proposed method to the previous method. In particular, all the listeners answered that they experienced internalization of sound when they heard 3D sound based the zero-phase HRTF. On the other hand, it was described that the sound was externalized and immersive when listening to the 3D sound applied with the HRTF (proposed method) estimated the magnitude and phase. As a result, it was confirmed that the proposed method of estimating up to the phase is much better than the previous method.

For the performance evaluation of elevation enhancement using notch filter, a listening experiment was conducted with five subjects. Three subjects answered that the elevation of the output signal was felt more clearly and two subjects replied that they did not notice a significant difference in reference and output.

5. Conclusions

A new method for estimating the magnitude and phase of personalized HRTF was proposed by using the anthropometric measurements of customers as input to the DNN model. In particular, while the previous DNN-based HRTF estimation method used zero-phase, we proposed a new method which is also estimates the phase of HRTF. As a result of objective and subjective performance evaluations, the proposed method can make listeners experience more immersive audio than the previous methods. The contribution of this paper is to academically propose a new method for phase reconstruction of acoustic signals. In addition, practically, the improvement of immersive 3D sound has been achieved. For further research works, in order to increase the performance of the DNN model, a larger amount of training data is required. Alternatively, the performance of DNN can be improved through an appropriate data augmentation technique.

Acknowledgement

This study was supported by the Research Program funded by the SeoulTech (Seoul National University of Science and Technology).

References

[1] T. Chen, T. Kuo and T. Chi, "Autoencoding HRTFS for DNN Based HRTF Personalization Using Anthr opometric Features," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 271-275.
DOI: https://doi.org/10.1109/ICASSP.2019.8683814.

[2] Rumsey, F. (2001). Spatial Audio (1st ed.). Routledge.
DOI: https://doi.org/10.4324/9780080498195

[3] Begault, R.D. 3D Sound for Virtual Reality and Multimedia; Academic Press: Cambridge, MA, USA, 1994.
DOI: https://doi.org/10.2307/3680997

[4] Kistler, D.J.; Wightman, F.L., "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," J. Acoust. Soc. Am. 1992, 91, 1637–1647.
DOI: https://doi.org/10.1121/1.402444

[5] Ngai-Man Cheung, S. Trautmann and A. Horner, "Head-related transfer function modeling in 3-D sound systems with genetic algorithms," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, Seattle, WA, USA, 1998, pp. 3529-3532.

DOI: https://doi.org/10.1109/ICASSP.1998.679630.

[6] Hu, H.; Zhou, L.; Ma, H.; Wu, Z., "HRTF personalization based on artificial neural network in individual virtual auditory space," Appl. Acoust. 2008, 69, 163–172.
DOI: https://doi.org/10.1016/j.apacoust.2007.05.007

[7] Chun, C.J.; Moon, J.M.; Lee, G.W.; Kim, N.K.; Kim, H.K., "Deep neural network based HRTF personalization using anthropometric measurements," In Proceedings of the 143rd AES Convention, New York, NY, USA, 18–21 October 2017. Preprint 9860.
DOI: https://doi.org/10.3390/app8112180

[8] V. C. Raykar and R. Duraiswami, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," J. Acoust. Soc. Am., vol. 118, no. 1, pp. 364-374, July 2005.
DOI: https://doi.org/10.1121/1.1923368

[9] Hebrank, J. and Wright, D. (1974b), "Spectral cues used in the location of sound sources on the median plane," J. Acoust. Soc. Am. 56, 1829–1834.
DOI: https://doi.org/10.1121/1.1903520

[10] V. R. Algazi, R. O. Duda, D. M. Thompson and C. Avendano, "The CIPIC HRTF database," *Proceeding s of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, New Platz, NY, USA, 2001, pp. 99-102.
DOI: https://doi.org/10.1109/aspaa.2001.969552