



Audio Engineering Society

Convention Paper 9860

Presented at the 143rd Convention
2017 October 18–21, New York, NY, USA

This Convention paper was selected based on a submitted abstract and 750-word precis that have been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This convention paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library, <http://www.aes.org/e-lib>. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Deep Neural Network Based HRTF Personalization Using Anthropometric Measurements

Chan Jun Chun¹, Jung Min Moon², Geon Woo Lee², Nam Kyun Kim², and Hong Kook Kim²

¹ Korea Institute of Civil Engineering and Building Technology (KICT), Goyang 10223, Korea

² School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Korea

Correspondence should be addressed to Hong Kook Kim (hongkook@gist.ac.kr)

ABSTRACT

A head-related transfer function (HRTF) is a very simple and powerful tool for producing spatial sound by filtering monaural sound. It represents the effects of the head, body, and pinna as well as the pathway from a given source position to a listener's ears. Unfortunately, while the characteristics of HRTF differ slightly from person to person, it is usual to use the HRIR that is averaged over all the subjects. In addition, it is difficult to measure individual HRTFs for all horizontal and vertical directions. Thus, this paper proposes a deep neural network (DNN)-based HRTF personalization method using anthropometric measurements. To this end, the CIPIC HRTF database, which is a public domain database of HRTF measurements, is analyzed to generate a DNN model for HRTF personalization. The input features for the DNN are taken as the anthropometric measurements, including the head, torso, and pinna information. Additionally, the output labels are taken as the head-related impulse response (HRIR) samples of a left ear. The performance of the proposed method is evaluated by computing the root-mean-square error (RMSE) and log-spectral distortion (LSD) between the referenced HRIR and the estimated one by the proposed method. Consequently, it is shown that the RMSE and LSD for the estimated HRIR are smaller than those of the HRIR averaged over all the subjects from the CIPIC HRTF database.

1 Introduction

The human auditory system plays a fundamental role in receiving spatial information. For example, people can detect the direction of invisible auditory objects by spatial hearing and thus react accordingly. Such spatial hearing collects all horizontal and vertical information. The principle of spatial hearing is based on binaural and monaural cues [1]. Binaural cues imply the difference between two ears. They include the time difference of arrival and the intensity difference between two ears, which are referred to as the interaural time difference (ITD)

and interaural level difference (ILD), respectively. These binaural cues are related to perceiving the horizontal direction of a sound source. According to the duplex theory [2], the ITD is dominant in low frequencies (below 1–2 kHz), and the ILD is dominant in high frequencies (above 1–2 kHz). On the other hand, monaural cues contain the effects of the head, body, and pinna. They modify the magnitude spectrum of the sound source and are strongly related to perceiving the vertical direction of the sound source [3]. Another monaural cue is the reverberant factor, which is defined as the amount of reflection and reverberation relative to the direct

sound. It is usually expressed as the direct-to-reverberant energy ratio (DRR), which is primarily related to perceiving the distance of the sound source [4].

To date, spatial audio processing based on the human auditory system has been produced [5]. Audio rendering, which is a spatial audio processing technique, is used to localize a sound source into an arbitrary position in three-dimensional space. Thus, a listener virtually perceives the sound produced from the localized position. The audio rendering can be conducted in either the binaural or transaural configuration [5].

In the binaural configuration through headphones, a head-related transfer function (HRTF) provides immersive sound by filtering monaural sound [6]. The transfer function from a sound source to a listener's ear can be measured using a dummy head, which mimics a human head's eardrums [7]. This measurement represents the effects of the head, body, and pinna as well as the pathway from a given source position to a dummy head. Therefore, the HRTF should be individually designed because the propagation of sound is characterized differently depending on the head used for the measurement. In other words, when the measured HRTF is employed, the performance often degrades due to individual morphology differences [8]. Accordingly, the HRTF should also be individually measured to increase the localization performance. Unfortunately, it is difficult to measure individual HRTFs for all horizontal and vertical directions. To overcome such difficulties, previous works have mathematically modeled the HRTF [9]. In [10], HRTF personalization was presented by using a sparse representation of anthropometric features. They utilized limb and full body features as anthropometric features. In addition, other features, including hair color, shirt size and shoe size, were utilized.

In this paper, a deep neural network (DNN)-based HRTF personalization method is proposed by also using anthropometric measurements. While it was assumed that the HRTF was mainly associated with head, torso, and pinna information [10], the

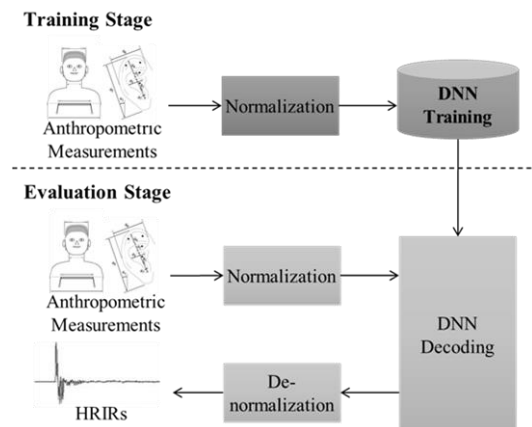


Figure 1. Block diagram of the proposed DNN-based HRTF personalization method.

proposed method utilizes all the features and enables the DNN to selectively train the importance of the features. To this end, the CIPIC HRTF database [11], which is a public-domain database of HRTF measurements for 45 subjects, is first analyzed to generate a DNN model for HRTF personalization. Specifically, the input features for the DNN model are taken as the anthropometric measurements of a left ear, such as the head, torso, and pinna information from the database. In addition, the output labels are taken as the time-domain samples of head-related impulse responses (HRIRs) for a left ear. The performance of the proposed method is evaluated by computing the root-mean-square error (RMSE) and log-spectral distortion (LSD) between the referenced HRIR and the estimated one by the proposed personalization method.

Following this Introduction, Section 2 proposes a DNN-based HRTF personalization method. Then, Section 3 evaluates the performance of the proposed method and compares it with a case that uses the HRIR averaged over all the subjects from the CIPIC database. Finally, Section 4 concludes this paper.

2 DNN-Based HRTF Modeling

Fig. 1 illustrates a block diagram of the proposed HRTF personalization method. The DNN model is trained by using the CIPIC HRTF database, which consists of the HRIRs for 45 subjects. Among all the

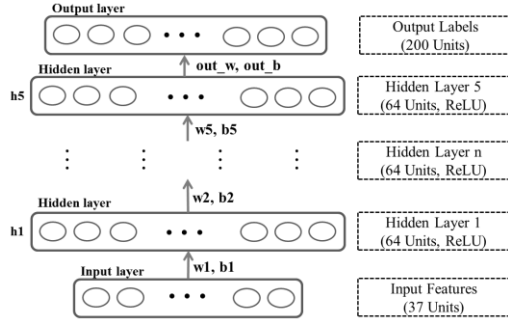


Figure 2. A DNN architecture for the proposed method.

subjects, the data of 35 subjects are taken because they have both HRIRs and anthropometric measurements. In particular, 34 out of 35 subjects are actually used for training the DNN model and the other is used for evaluating the performance of the proposed method.

Fig. 2 shows the DNN architecture of the proposed method for HRTF personalization. As shown in the figure, the DNN model is composed of one input layer, five hidden layers, and one output layer. The number of input units is 37 (17 parameters for the height and circumference measurements and 20 for the pinna measurements). Specific information about anthropometric measurements is described in [11]. The number of output nodes is set to 200. This is because the length of each HRIR in the CIPIC HRTF database is 200 samples, corresponding to a duration of approximately 4.5 msec.

Actually, each input feature for the DNN is normalized by using the mean and variance for all the training data without regarding to the subjects, such as

$$\bar{x}_i = \left(1 + e^{-\frac{(x_i - \mu_i)}{\sigma_i}} \right)^{-1} \quad (1)$$

where x_i is the i -th component of the input feature vector and μ_i and σ_i are the mean and standard deviation for all the training data, respectively. That is, the input features for the DNN are $\{\bar{x}_i\}$ instead of $\{x_i\}$. In addition, the output values are normalized

No. of Hidden Units	No. of Hidden Layers			
	2	3	4	5
32	0.002101	0.002452	0.002041	0.001822
64	0.002094	0.002277	0.002003	0.001758
128	0.002044	0.002311	0.002160	0.001896

Table 1. Comparison of the loss values depending on the number of hidden layers and hidden units for DNN training.

so that they have a dynamic range from 0 to 1. In fact, the HRIR values range from -2 to 2; thus, they are augmented to by 2 and divided by 4.

For unsupervised pre-training, the Xavier initialization technique is utilized [12]. For supervised fine-tuning, the backpropagation method with a minimum-mean-squared-error (MMSE) cost function between the original HRIRs and the estimates of the HRIRs is used [13]. To update the weights in the neural network, the adaptive moment estimation (ADAM) optimization technique with the learning rate of 0.001 is utilized [14]. The dropout technique is also employed to eliminate the overfitting problem [15]. In general, the architecture of the DNN, such as the number of hidden layers, units, and the training parameters, such as the number of iterations and the dropout rate, are set by observing the performance of HRTF personalization.

Table 1 compares the loss values depending on the number of hidden layers and hidden units for DNN training. Consequently, the number of hidden layers and units are set to 5 and 64, respectively. In addition, the number of iterations and the dropout rate are set to 20,000 and 0.9, respectively. In the DNN, a rectified linear unit (ReLU) is used as the activation function for all the nodes, including output nodes. Finally, the output values from the DNN output nodes are multiplied by 4 and reduced by 2 because the output values for training the DNN have a dynamic range from -2 to 2.

3 Performance Evaluation

The performance of the proposed HRTF personalization technique was evaluated by measuring objective differences between the measured HRTF for a given position and the one estimated, for that position. First, the RMSE

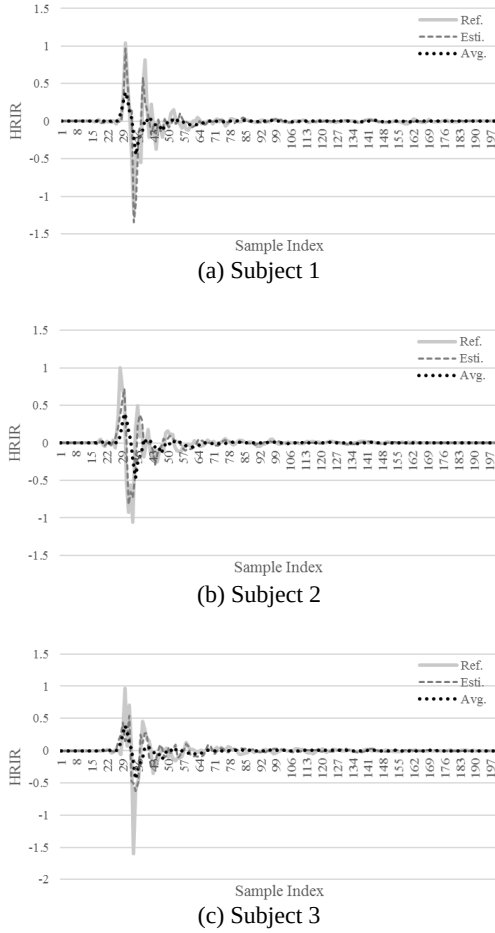


Figure 3. References, estimates, and averages of HRIRs depending on the subjects.

between the original HRIR, $y(n)$, and the estimated HRIR, $\hat{y}(n)$, is defined as

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

where N ($=200$) is the total length of the HRIR. In addition, the log-spectral distortion between $y(n)$ and $\hat{y}(n)$ is

$$LSD(Y, \hat{Y}) = \sqrt{\frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \left(20 \log_{10} \frac{|Y(k)|}{|\hat{Y}(k)|} \right)^2} \quad (3)$$

Measure	Estimated HRIRs	Averaged HRIR
RMSE (dB)	-19.4724±2.0929	-15.5465±0.8204
LSD (dB)	1.5315±0.9682	3.3761±0.3967

Table 2. Results of the average RMSE and LSD of estimated HRIRs and those of averaged HRIR from the CIPIC database.

where $Y(k)$ and $\hat{Y}(k)$ are obtained by applying a discrete Fourier transform to $y(n)$ and $\hat{y}(n)$, respectively. In (3), k_1 and k_2 indicate the minimum and maximum frequencies for the LSD computation, respectively. Here, $k_1 = 1$ (172 Hz) and $k_2 = 116$ (19,982 Hz) when the 256-point fast Fourier transform. Note that the LSD is only calculated in the audible frequency band. In this evaluation, some basic experiments on HRTF personalization were conducted when the HRIRs at azimuth = -45° and elevation = 45° are generated.

As mentioned earlier, the CIPIC database had 45 subjects in total but only 35 with anthropometric data. Thus, the performance evaluation was conducted by leave-one-out cross validation applied to 35 subjects. In other words, 35 trial were performed for training a DNN model and evaluating the RMSE and LSD, and the results were averaged.

To discuss the performance of the proposed method, Fig. 3 shows the references, estimates, and averages of HRIRs. As shown in the figure, the estimates of the HRIRs are more similar to the references than the averages in all subjects. Table 2 compares the RMSE and LSD of HRIRs estimated by the proposed method with those of the HRIR averaged over all the subjects from the CIPIC HRTF database. As shown in the table, the estimated HRIRs has lower RMSE and LSD than the averaged HRIR, which implies that each estimated HRIR is closer to the corresponding referenced HRIR than the averaged HRIR.

4 Conclusion

A DNN-based HRTF personalization method has been proposed in this paper. In particular, the input features were taken as the anthropometric

measurements of a left ear, and the output labels were also taken as the time-domain samples of HRIRs for a left ear. Depending on the number of hidden layers, units, and iterations and the dropout rate, the cost functions were compared. Consequently, the optimal parameters for DNN were experimentally determined. The performance evaluation was conducted by leave-one-out cross validation applied to 35 subjects from the CIPIC HRTF database. By analyzing the RMSE and LSD, it was shown that the RMSE and LSD for the estimated HRIR were smaller than those of the HRIR averaged over all the subjects from the CIPIC HRTF database.

5 Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the government of Korea (MSIP) (No. 2015R1A2A1A05001687), and by the ICT R&D program of MSIP/IITP [R01261510340002003, Development of hybrid audio contents production and representation technology for supporting channel and object based audio].

References

- [1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, MA (1997).
- [2] L. Rayleigh, "On our perception of sound direction," *Philosophical Magazine*, vol. 13, pp. 214–232 (1907).
- [3] S. Roffler and R. Butler, "Factors that influence the localization of sound in the vertical plane," *Journal of the Acoustical Society of America*, vol. 43, no. 6, pp. 1255–1259 (1968).
- [4] A. W. Bronkhorst and T. Houtgas, "Auditory distance perception in rooms," *Nature*, vol. 397, pp. 517–520 (1994).
- [5] F. Rumsey, *Spatial Audio*, Focal Press, Woburn, MA (2001).
- [6] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637–1647 (1992).
- [7] D. R. Begault, *3D Sound for Virtual Reality and Multimedia*, Academic Press, Cambridge, MA (1994).
- [8] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using non-individualized head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123 (1993).
- [9] C. P. Brown and R. O. Duda, "An efficient HRTF model for 3-D sound," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, pp. 298–301 (1997).
- [10] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, pp. 4468–4472 (2014).
- [11] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, pp. 99–102 (2001).
- [12] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, pp. 249–256 (2010).
- [13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 9, pp. 533–536 (1986).
- [14] D. P. Kingma and J. L. Ba, "ADAM: a method for stochastic optimization," in *Proc. 3rd International Conference on Learning Representations*, San Diego, CA, pp. 1–15, (2015).
- [15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958 (2014).