INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

Directional Audio Rendering Using a Neural Network Based Personalized HRTF

Geon Woo Lee, Jung Hyuk Lee, Seong Ju Kim, Hong Kook Kim

School of EECS, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Korea
{geonwoo0801, ljh0412, sjkim3357, hongkook}@gist.ac.kr

Abstract

Multi-channel speech/audio separation and enhancement methods are popularly used for many speech/audio related applications. However, these methods may cause a loss of spatial cues, including the interaural time difference and interaural level difference, for further processing of monoaural signals. Thus, listeners may encounter difficulties in understanding the direction of the source signal. We present a directional audio renderer using a personalized HRTF, which is estimated by a neural network that combines DNN and CNN with anthropometric parameters and ear images of the listener. This demonstrated directional audio renderer concept aims to help foster research on audio processing for virtual reality/augmented reality to improve the quality of service of such devices.

Index Terms: directional audio rendering, head-related transfer function, neural network, ear image, anthropometric parameter

1. Introduction

Many speech/audio related applications utilize a multi-channel microphone array in order to separate or enhance target speech or audio signals in noisy and reverberant environments [1]. However, signals processed by multi-channel speech/audio techniques cannot be reproduced with an appropriate sound directivity level. This is attributed to the loss of spatial cues contained in the multi-channel signal, such as the interaural time difference (ITD) and interaural level difference (ILD), in the processed signal. In particular, such a loss is crucial when the processed signal is played out in a virtual reality (VR) or augmented reality (AR) environment, where spatial sound is vital for enhancing the immersive reality when combined with video [2].

Directional audio rendering is a spatial audio processing technique used to localize a sound source to an arbitrary position in 3D-space [3]. Thus, with this method, a listener can perceive a sound generated from a localized virtual position. Among such various techniques, head-related transfer functions (HRTFs) are known to act as a powerful tool for simply generating spatial and immersive sounds [3]. HRTFs are defined by the path from a given sound source to a listener's ear drum through the head, torso, and pinna, and are represented as transfer functions. Therefore, HRTFs differ from person to person due to sound propagation variations of individuals. Thus, applying measured HRTFs from a dummy head or other people can degrade the performance of immersive sound effects. Therefore, HRTFs should be individually designed or measured to take into account the varying sound propagation properties of each user.

This paper presents a directional audio rendering technique using personalized HRTFs estimated by a neural network based on the anthropometric parameters and ear images of the listener.

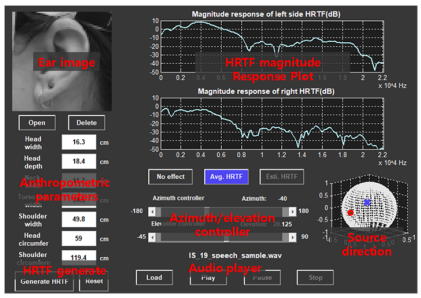

Figure 1: *Snapshot of the simulator, demonstrating directional audio rendering. Red text is not a part of the simulator, but is included for explanation.*

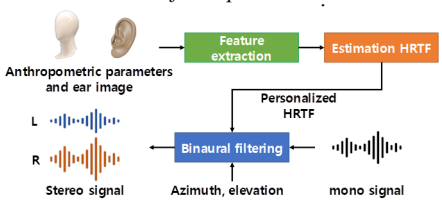

Figure 2: *Schematic diagram of the proposed directional audio rendering technique employing the personalized HRTF estimation.*

Figure 1 shows a snapshot of the simulator used for the directional audio rendering. The simulator consists of four parts; measurements, feature extractions, personalized HRFT estimation, and sound reproduction.

Figure 2 depicts a schematic diagram of the directional audio rendering technique embedded in Figure 1. First, seven anthropometric parameters and an ear image of a listener are measured, and the ear image is then converted to an edge-detected image. Next, in order to estimate the personalized HRTFs of the listener, these feature parameters are used as the input features of the HRTF estimation model constructed by combining CNN and DNN. Finally, the personalized HRTF corresponding to a given azimuth and elevation in degrees is then convoluted with the enhanced monaural signal in order to obtain the stereo signal.

2. Personalized HRTF estimation

Figure 3 illustrates the architecture of the personalized HRTF estimation model using anthropometric parameters and an ear image of a user, where the input parameters in [4] are simplified as seven anthropometric parameters that are all measurable. As shown in the figure, the HRTF estimation model is composed of three neural network modules. Sub-networks A and B use seven anthropometric parameters and one ear image as input features, respectively, and Sub-network C combines the outputs of the two sub-networks to estimate HRTF. In order to train the model, a public HRTF database provided by the Center for Image Processing and Integrated Computing (CIPIC) is used [5].

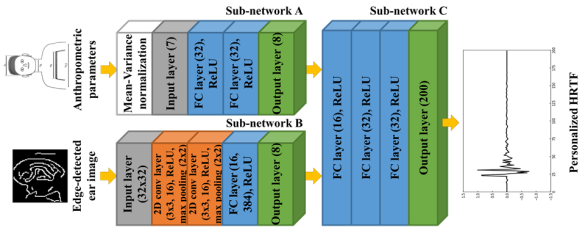

Figure 3: *Architecture of the personalized HRTF estimation model using anthropometric parameters and an ear image.*

2.1. Sub-network A: DNN using anthropometric parameters

Sub-network A is composed of an input layer, two hidden layers, and an output layer. Each hidden layer has 32 nodes with a rectified linear unit (ReLU) activation function. Seven features (head width, head depth, neck width, torso top width, shoulder width, head circumference, and shoulder circumference) are used for the input layer. Note here that there are ten additional anthropometric parameters in the CIPIC HRTF database, yet these parameters are excluded here as they are difficult to measure in real applications. The input features used in this network are normalized by the mean and variance for all training data, without regards to the subjects.

2.2. Sub-network B: CNN using ear image

Sub-network B is composed of an input layer, two 2D-convolution layers followed by a max-pooling layer, and an output layer. Here, each 2D-convolution layer has 3x3 kernels with a stride size of 1. Moreover, ReLU activation is implemented. Edge-detected ear images with a 32x32 resolution are used for the input features of this sub-network, where the edge detection is performed using the Canny edge detection algorithm.

2.3. Sub-network C: DNN for HRTF estimation

Sub-network C consists of a concatenated input layer with 16 nodes, and three hidden layers with 32 nodes each. Note that the output dimension is 200, corresponding to the length of the HRTFs in the CIPIC HRTF database. In addition, the ReLU activation function is applied to all hidden layers, with the exception of the output layer.

2.4. Supervised learning

In order to train the CNN and DNN combination model, Xavier initialization is applied for the initial weight of the model, and the mean square error between the original and estimated target is then selected as the cost function. The adaptive moment estimation optimization is utilized to update the model parameters with a learning rate of 0.001, and dropout is performed with the dropout rate of 0.1. The model is trained for 30,000 epochs with an early stopping technique.

3. Performance evaluation of the personalized HRTF estimation model

The performance of the personalized HRTF estimation model was evaluated by measuring the root mean square error (RMSE) and log spectral distance (LSD) between the measured and estimated HRTF. The performance of the average HRTFs and the CNN-DNN models with different number of anthropometric parameters were then compared. A total of 30 subjects of the CIPIC HRTF database were used for training each neural network, while one subject was used for testing.

Table 1: *Comparison of the average RMSE and LSD of different HRTF models.*

Measure	Azimuth	-135°	-80°	-45°	Avg.
RMSE (dB)	Average HRTF	-20.98	-19.39	-17.32	-19.23
	CNN-DNN (17)	-20.26	-18.60	-16.35	-18.40
	CNN-DNN (7)	-20.42	-18.82	-16.38	-18.54
LSD (dB)	Average HRTF	11.29	4.12	7.41	7.61
	CNN-DNN (17)	5.70	3.11	4.61	4.47
	CNN-DNN (7)	5.82	3.52	5.56	4.97

Thus, 31 cross-validations were performed, with all measurements averaged over all cross-validations.

Table 1 compares the RMSE and LSD of the average HRTF and CNN-DNN models with varying anthropometric parameters, measured at the azimuths of -135°, -80°, and -45° with an elevation of 0°. As shown in the table, the CNN-DNN-based models provided higher RMSEs by 0.83 and 0.69 dB, and lower LSDs by 3.14 and 3.60 dB compared to the average HRTF. The performance differences according to the number of anthropometric parameters were marginal.

4. Demonstration and conclusion

We designed a directional audio rendering simulator implemented using a Matlab-based graphical user interface (GUI) for the interactive demonstration, as depicted in Figure 1. First, the user's ear image is captured using a smart phone. Following this, the image is cropped using a photo editor such that the ear area is exaggerated. This is then inputted to the simulator by clicking "open". Next, the simulator receives the user's seven anthropometric parameters, and both the image and these anthropometric parameters are used for the NN-based HRTF estimation model implemented by Tensorflow. For a given pair of azimuth and elevation, the HRTFs for the left and right ear are estimated in real time, whereby 1,250 directions (25 azimuths and 50 elevations) can be estimated by adjusting the azimuth and elevation slide bars. Finally, audio rendering is performed by applying the estimated HRTFs to the signal that is already loaded. The designed simulator will be interactively demonstrated in a Show & Tell Session, and can contribute to the improvement of the quality of service requirements of VR/AR devices.

5. Acknowledgements

This work was supported in part by the Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2017-0-00255, Autonomous digital companion framework and application), and by GIST Research Institute(GRI) grant funded by the GIST in 2019.

6. References

[1] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–12, 2015.

[2] F. Rumsey, *Spatial Audio*, Focal Press, Woburn, MA, 2001.

[3] D. R. Begault, *3D Sound for Virtual Reality and Multimedia*, Academic Press, Cambridge, MA, 1994.

[4] G. W. Lee and H. K. Kim, "Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear," *Applied Sciences*, vol. 8, no. 11, 2180, 2018.

[5] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 24, New Paltz, NY, Proceedings*, 2001, pp. 99–102.