AUTOENCODING HRTFS FOR DNN BASED HRTF PERSONALIZATION USING ANTHROPOMETRIC FEATURES

Tzu-Yu Chen, Tzu-Hsuan Kuo, and Tai-Shih Chi

Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu 300, Taiwan

ABSTRACT

We proposed a deep neural network (DNN) based approach to synthesize the magnitude of personalized head-related transfer functions (HRTFs) using anthropometric features of the user. To mitigate the over-fitting problem when training dataset is not very large, we built an autoencoder for dimensional reduction and establishing a crucial feature set to represent the raw HRTFs. Then we combined the decoder part of the autoencoder with a smaller DNN to synthesize the magnitude HRTFs. In this way, the complexity of the neural networks was greatly reduced to prevent unstable results with large variance due to overfitting. The proposed approach was compared with a baseline DNN model with no autoencoder. The log-spectral distortion (LSD) metric was used to evaluate the performance. Experiment results show that the proposed approach can reduce LSD of estimated HRTFs with greater stability.

Index Terms— HRTFs, Anthropometry, Autoencoder, DNN, Spatial audio

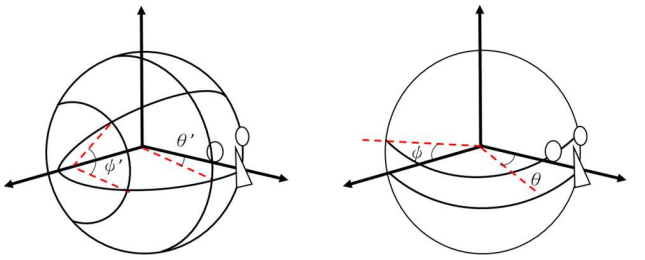
1. INTRODUCTION

Head-related transfer functions (HRTFs) are embedded with binaural cues such as interaural time difference (ITD), interaural level difference (ILD) and spectral modifications caused by pinna, head, and torso. These cues are used by human to localize sound sources. The HRTF is defined as the transfer function of the spatial filter from the sound source to the entrance of the ear canal in the frequency domain. Hence, a virtual audio signal from an arbitrary location can be produced by filtering a non-spatial audio signal using the corresponding HRTF. However, HRTFs are very sensitive to anthropometric features, that is, they are different from person to person such that using other peoples HRTFs would cause direction confusion. Hence, HRTFs need to be personalized, however, directly measuring HRTFs is time-consuming and expensive.

Several methods have been proposed for customizing HRTFs. For instance, the transmitted audio signal can be divided into several blocks, each of which has its own mathematical model with personalized parameters [1,2]. In addition to the model-based methods, methods of selecting approximate HRTFs from a HRTF database for an individual were also proposed in [3,4]. Later on, based on the assumption that HRTFs and anthropometric features share a very similar relation, personalized HRTFs of a new subject were derived by applying his weights in the space of the sparse representation of anthropometric features to stored HRTF templates [5–7]. Besides, statistical analysis, regression analysis, and support vector regression analysis were carried out to find principal anthropometric features in customizing HRTFs [8–10]. Gradually, the concept of neural network was also introduced into this research area. For example, dimensional reduction methods, such as principal component analysis (PCA) and isometric feature mapping (Isomap), were combined with neutral networks to synthesize HRTFs [11–16]. Not surprising, directly using a deep neural network (DNN) to estimate personalized head-related impulse responses (HRIRs) from the anthropometric features was proposed recently in [17].

However, DNNs require lots of data for training or they would suffer from the over-fitting problem to produce poor estimation for unseen condition. Although some public HRTF datasets are available, their sizes are not large due to the time-consuming measuring. Besides, there is no unified feature set for measuring HRTFs such that different datasets record different anthropometric features. Therefore, directly combining different datasets into a large dataset for DNN training is not feasible. To solve this problem, we propose an autoencoder to encode HRTFs from different datasets. By increasing the number of samples of HRTFs, the autoencoder becomes more general and provide good description of the HRTF space. Then, a DNN is used to map the measured anthropometric features to the more general bottleneck features of the autoencoder to decode personalized HRTFs. To evaluate our idea, we compared performance of the proposed method to performance of a baseline DNN system built by following the architecture and parameters in [17]. Note, for the purpose of comparison, the baseline DNN system was built to estimate HRTFs rather than HRIRs.

The rest of this paper is organized as follows. In Section 2, we will describe pre-processing of the data and the mechanism of the proposed method which combines an autoencoder and a DNN. In Section 3, the experiment results of compared systems are given and discussed. Finally, the conclusion and potential future work are given in Section 4.



(a) Interaural coordinate system (b) Vertical coordinate system

Fig. 1. Different coordinate system for allocation of HRTFs measuring points

2. PROPOSED METHOD

We have done some pilot experiments in estimating personalized HRTFs directly using a DNN. However, the estimation error of such a model has a large variance potentially caused by the over-fitting of the DNN model. To mitigate the over-fitting, the training dataset needs to be enlarged and the number of involved estimation parameters needs to be reduced. For these two purposes, we adopted an autoencoder to encode/decode an enlarged HRTF space, which was established by adding more HRTF samples from different datasets, using less number of features (i.e., the bottleneck features of the autoencoder). The proposed method and the pre-processing on the data is described in more details in this section.

2.1. The CIPIC database and pre-processing

The public CIPIC database [18], which is the most popular database in the research field of estimating personalized HRTFs, was used as the primary database in our method. This database contains three sets of data, including ITD, HRIRs, and corresponding anthropometry measurements. Since we focused on estimating magnitude responses of HRTFs, the ITD data was not used in our experiments.

2.1.1. Definition of the coordinate system

The CIPIC database adopts the interaural coordinate system rather than the vertical coordinate system, which is the most used coordinate system when measuring HRIRs in past studies. Both coordinate systems are shown in Fig. 1. The elevation angle of the interaural coordinate system is defined in the range of $-90^\circ \leq \phi' < 270^\circ$ and the azimuth angle is in the range of $-90^\circ \leq \theta' < 90^\circ$ as shown in Fig. 1(a). For our approaches, the azimuth angle was re-defined. The definition of the negative azimuth angle was changed from the left side of the listener to the ipsilateral side of the receiving ear. In contrast, the definition of the positive azimuth angle was changed from the right side of the listener to the contralateral side of the receiving ear. The advantage of this new definition is explained in section 2.2.1.

2.1.2. Deriving HRTFs from HRIRs

To derive magnitude responses of HRTFs, we first applied 512-point FFT on the raw HRIRs in CIPIC database, then smoothed the magnitude spectra using a constant-Q filterbank (Q=8), and finally took the logarithm to produce magnitude HRTFs in dB. For our usage, we retained magnitude HRTFs between 200 Hz and 15 kHz based on the study in [11]. After these processes, each magnitude HRTF was represented by a vector of the length of 173. Since we adopted the sigmoid function as the activation function of the output layer, the log. magnitude of the HRTF was normalized to values between 0 and 1. Without loss of generality, the performance comparison and discussion in this paper are based on experiment results of the left ear on all azimuth angles at the 0° elevation angle.

2.1.3. Anthropometric features

The CIPIC database contains 37 anthropometry measures, including 17 measures related to the torso and the head and 10 measures related to each pinna. The definition of these measurements can be accessed in [18]. Since we were estimating magnitude HRTFs of the left ear, we didn't use the right-pinna-related 10 measures. Hence the input anthropometric feature of our DNN model is a 27-dimensional vector. Each input feature vector was then normalized by following procedures in [17] as

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

where x_i is the $i-th$ feature and μ_i and σ_i are the mean and standard deviation of the $i-th$ feature, respectively. Finally, the $\{x_i', i = 1, 2...27\}$ were used as the input features to the DNN model.

2.2. Architecture of proposed models

The proposed autoencoder and DNN models are shown in Fig. 2. The training processes are indicated by the bold and solid arrows. First, we trained the autoencoder for HRTFs. Then, the bottleneck vector, the encoder of HRTFs, were used as the target when training the DNN model with anthropometric features as input. On the other hand, the dash arrows indicate the test processes, where we generate personalized magnitude HRTFs with anthropometric features of a new subject. During the test phase, the DNN model first produced the bottleneck vector based on anthropometric features of the new subject. Then, the decoder part of the autoencoder decoded the estimated bottleneck vector to produce the estimated magnitude HRTFs. Details of the training and parameters used in each model are described below.

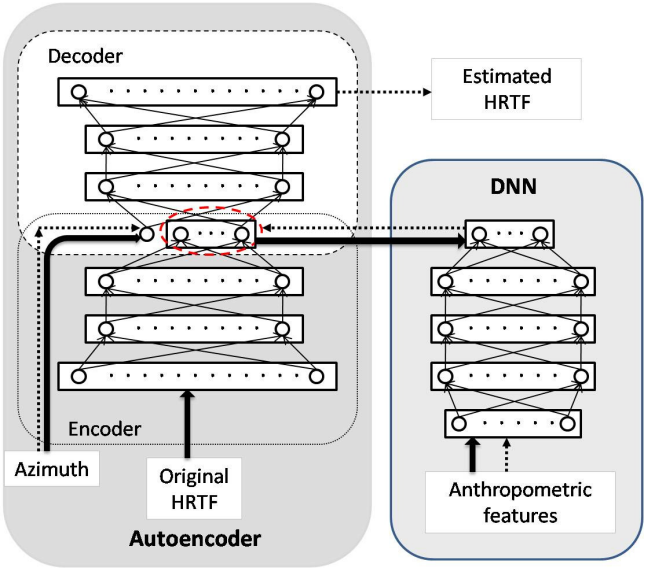


Fig. 2. The architecture of the proposed autoencoder and DNN models and training/test procedures.

2.2.1. Autoencoder settings

In the proposed method, one autoencoder was trained for each elevation angle. As shown in Fig. 2, the azimuth angle was appended to the bottleneck vector as the label of the vector, with its sign defined in Section 2.1.1. Originally, each HRTF requires two parameters, left or right ear and the azimuth angle, to describe the horizontal relative position of the sound source to the receiving ear. After changing the definition of the sign of the azimuth angle as in Section 2.1.1., only the azimuth angle parameter is needed to describe the relative position of the sound source. In addition, the left ear and the right ear now share a single autoencoder, which doubles the training data for each autoencoder. The rationale of our approach is that HRTFs of the two ears are highly symmetric.

Each autoencoder has five hidden layers, including the latent layer (bottleneck layer). We adopted the ReLU function as the activation function for all hidden layers, and the sigmoid function for the output layer. According to [19], we set the width of latent layer to 20 for encoding HRTFs while not losing too much information between different subjects. The width of all other hidden layers was set to 150 for the least reconstruction error. The mean-squared-error (MSE) was adopted as the cost function and the adaptive moment estimation (ADAM) technique with the learning rate of 0.001 was used for optimization during training. Besides, we set the dropout rate to 0.95 due to the condition of over-fitting.

2.2.2. DNN settings

Similar to the method in [17], one DNN model was trained for predicting bottleneck features for one azimuth angle. In other words, in our experiments, we trained one autoencoder for a particular elevation angle and 25 DNN models for 25 azimuth angles at the same elevation. Since HRTFs were encoded by the bottleneck features, the dimension of the output layer of the DNN model was much reduced comparing with the baseline system in [17]. Each DNN contained three hidden layers with width of 40 units. The ReLU function was used as the activation function in all hidden layers and the output layer. Same as in the autoencoder, the MSE was chosen as the cost function and ADAM technique with the learning rate of 0.001 was used for optimization. Dropout rate was set to 0.95 as well.

Table 1. Mapping between the azimuth angle in the vertical coordinate system (θ) and the azimuth angle in the interaural coordinate system (θ').

	Sound source on the left side (0° to 90°)	Sound source on the right side (270° to 360°)
Left ear	$\theta' = -\theta$	$\theta' = 360 - \theta$
right ear	$\theta' = \theta$	$\theta' = \theta - 360$

2.3. Further improvement

2.3.1. Joint training

In our original approach, we trained an autoencoder for dimensionality reduction of the HRTFs at a particular elevation angle. One DNN was trained separately for each azimuth angle. Although this approach has already shown improvement over the DNN-only baseline system, further actions were taken for potential improvement. The first action was fine-tuning weights of the system by joint training. In other words, after the DNN was separately trained, we connected it with the decoder part of the trained autoencoder to fine-tune the weights. However, this means HRTFs of different azimuth angles do not share the same decoder, but having similar decoders, and our system becomes more complex.

2.3.2. Combining HRTF dataset

To increase training data, we added HRTFs of the LISTEN database [20] and the SADIE database [21] in our experiments. Since these two databases use the vertical coordinate system for measuring HRTFs, the mapping of the angles between two coordinate systems is needed. Considering our new definition in Section 2.1.1, the mapping rules of azimuth angles between two coordinate systems at 0° elevation angle are shown in Table 1. For other angles, more complicated mapping rules between two coordinate systems need to

be further derived for both the elevation and the azimuth angles.

3. EXPERIMENT RESULTS

In HRTF related studies, LSD is usually used to evaluate the performance of the proposed method. It is formulated as follows

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

where k is the index of frequency bin; H and $\hat{H}$ are the actual and the predicted HRTFs.

Four different approaches, named Baseline, AutoEn+DNN, Joint-Training, and Dataset-Combined, were evaluated. The Baseline system is a pure DNN model we built by following the architecture and parameters in [17]. The AutoEn+DNN is our original approach, where the autoencoder and DNNs are trained separately. The Joint-Training approach fine-tunes the weights of each DNN and the decoder by further joint training as mentioned in Section 2.3.1. The Dataset-Combined approach further combines HRTFs in other databases for training the autoencoder. In this study, we only used HRTFs from 35 subjects, who have complete anthropometry measurements, in the CIPIC database for evaluation. In addition, we only considered the plane of 0° elevation angle where the simple mapping rules in Table 1 can be applied. In other words, for each approach, we estimated 25 magnitude responses of HRTFs at all azimuth angles at the 0° elevation ($\phi' = 0^\circ$) and computed the LSD. Leave-one-out cross validation was adopted, so we had 25×35 LSD values for each compared approach.

Fig. 3 shows mean and variance of LSD of four compared approaches at the azimuth angle of -65, -30, 0, 30, and 65 degrees. In Table 2, we show the means of 25×35 LSD of each compared approach. From these results, we can observe that both mean and variance at most azimuth angles can be reduced by incorporating an autoencoder for dimensional reduction. Furthermore, joint training the DNN and the decoder part of the autoencoder to fine-tune the weights can further improve the performance slightly. However, the benefit of adding more HRTF data is not clearly shown yet.

4. CONCLUSION AND FUTURE WORK

Although several public HRTF databases are available, none of them is large enough for training DNNs. Therefore, we propose an autoencoder for HRTFs for dimensional reduction to lighten the over-fitting problem by reducing the complexity of the DNN. The data set for training the autoencoder can be doubled by utilizing the quasi-symmetric characteristic of HRTFs. Experiment results show that the proposed approach

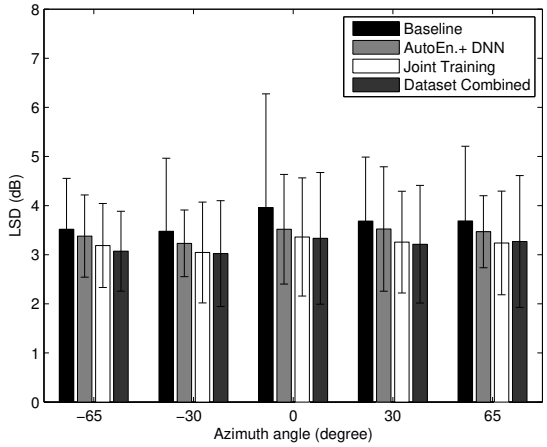


Fig. 3. Mean and variance of LSD of compared approaches at different azimuth angles.

Table 2. Mean LSD of compared approaches over all azimuth angles

Training Scheme	Baseline	AutoEn +DNN	Joint Training	Dataset Comb.
mean LSD	3.705	3.429	3.252	3.246

can reduce both the mean and variance of LSD of different azimuth angle. In other words, the proposed approach can estimate HRTFs more accurately and more stably. The other direct way to mitigate the over-fitting problem is to add more training data. However, measuring HRTFs is very expensive and time consuming. The idea of using autoencoder to combine different HRTF databases to build a universal HRTF codebook is carried out in this study. No benefit is shown yet from the experiment results. One possible reason is the combined dataset is still too small to show any benefit. The other reason is that different measuring environments of different HRTF database, including different measuring distance and different setups of the measuring chamber, etc., might introduce more variance into the combined dataset. Further studies in effectively combining different HRTF databases will be carried out in the future.

5. ACKNOWLEDGEMENT

This research is supported by the Ministry of Science and Technology, Taiwan under Grant No MOST 107-2221-E-009-132-MY3.

6. REFERENCES

[1] C. P. Brown and R. O. Duda, "An efficient hrtf model for 3-d sound," in *Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 1997.

[2] N.A. Gumerov, A.E. O'Donovan, R. Duraiswami, and D. N. Zotkin, "Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation," *The Journal of the Acoustical Society of America*, vol. 127, pp. 370–386, 2010.

[3] X. Liu and X. Zhong, "An improved anthropometry-based customization method of individual head-related transfer functions," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 336–339.

[4] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L.S. Davis, "Hrtf personalization using anthropometric measurements," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2003, pp. 157–160.

[5] 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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4468–4472.

[6] J. He, W. Gan, and E. Tan, "On the preprocessing and postprocessing of hrtf individualization based on sparse representation of anthropometric features," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 639–643.

[7] I. Tashev, "Hrtf phase synthesis via sparse representation of anthropometric features," in *Proceedings of Information Theory and Applications Workshop (ITA)*, Feb 2014, pp. 1–5.

[8] H. Hu, L. Zhou, J. Zhang, H. Ma, and Z. Wu, "Head related transfer function personalization based on multiple regression analysis," in *Proceedings of International Conference on Computational Intelligence and Security*, Nov 2006, vol. 2, pp. 1829–1832.

[9] Q. H. Huang and Q. L. Zhuang, "Hrir personalisation using support vector regression in independent feature space," *Electronics Letters*, pp. 1002–1003, Sep. 2009.

[10] M. Zhang, R. A. Kennedy, T. D. Abhayapala, and W. Zhang, "Statistical method to identify key anthropometric parameters in hrtf individualization," in *Proceedings of Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, May 2011, pp. 213–218.

[11] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein, "A manifold learning approach for personalizing hrtfs from anthropometric features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 559–570, March 2016.

[12] F. Grijalva, L. Martini, S. Goldenstein, and D. Florencio, "Anthropometric-based customization of head-related transfer functions using isomap in the horizontal plane," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4473–4477.

[13] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[14] K. J. Fink and L. Ray, "Individualization of head related transfer functions using principal component analysis," *Applied Acoustics*, vol. 87, pp. 162 – 173, 2015.

[15] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimumphase reconstruction," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637–1647, 1992.

[16] J. C. Middlebrooks and D. M. Green, "Observations on a principal components analysis of headrelated transfer functions," *The Journal of the Acoustical Society of America*, vol. 92, pp. 597–599, July 1992.

[17] C. J. Chun, J. M. Moon, G. W. Lee, N. K. Kim, and H. K. Kim, "Deep neural network based hrtf personalization using anthropometric measurements," in *Audio Engineering Society Convention*, 2017.

[18] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *Proceedings of IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, Oct 2001, pp. 99–102.

[19] H. Fayek, L. van der Maaten, G. Romigh, and R. Mehra, "On data-driven approaches to head-related-transfer function personalization," in *Audio Engineering Society Convention*, Oct. 2017.

[20] O. Warusfel, "The listen hrtf database," 2013.

[21] G. Kearney and T. Doyle, "An hrtf database for virtual loudspeaker rendering," in *Audio Engineering Society Convention*, Oct. 2015.