# Exploring redundancy of HRTFs for fast training DNN-based HRTF personalization

Tzu-Yu Chen, Po-Wen Hsiao, and Tai-Shih Chi
Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan 300, R.O.C.
E-mail: {as8815179, d8412121}@gmail.com, tschi@mail.nctu.edu.tw

***Abstract*—A deep neural network (DNN) is constructed to predict the magnitude responses of the head-related transfer functions (HRTFs) of users for a specific direction and a specific ear. Using the CIPIC HRTF database (including 25 azimuth angles and 50 elevation angles for both ears), we trained 2500 DNNs to predict magnitude responses of all HRTFs of a user. To reduce training time, we propose to use the final weights of the trained DNN of a nearby direction as the initial weights of the current DNN under training since magnitude responses of the HRTFs are smoothly changing across nearby directions. Analysis of variance (ANOVA) was performed to show that the proposed training scheme produces equivalent magnitude responses of HRTFs as the standard training scheme with random initial weights in terms of the log-spectral distortion (LSD) measure. Meanwhile, the proposed training scheme can dramatically reduce training time by more than 95%.**

## I. Introduction

Head-related transfer functions (HRTFs) model the scattering effects by the ears, the head, and the torso of a user to an audio wave travelling from the sound source to the ear canal. The HRTFs play an important role in a spatial audio system since filtering a non-spatial audio signal by HRTFs transforms it to a spatial audio signal virtually from specific directions. However, using HRTFs of another person would inevitably produce off-tuned spatial perception such as front-back confusion. To avoid spatial confusion, personalized HRTFs are needed for each user. Although measuring directly can have the most accurate HRTFs [1][2], it unfortunately requires time-consuming and expensive procedures.

Several methods have been proposed to synthesize personalized HRTFs. For instance, a mathematical model, whose parameters can be adjusted to fit a particular individual, was built in [3] to approximate the phenomenon of sound waves incident the ear. Methods of selecting appropriate HRTFs from a database for a particular individual were proposed in [4][5]. Based on the assumption that HRTFs data and the anthropometry features share a similar relation, a sparse representation of a given subject's anthropometry features were derived and applied to the HRTFs in [6][7] to synthesize personalized HRTFs.

Conventional methods try to find the linear relation between anthropometry features and HRTFs. As machine learning techniques widely utilized in various research areas during the past decade, it was also adopted to approximate the non-linear relation between anthropometry measurements and HRTFs. For instance, dimensionality reduction techniques, such as Isomap, and principal component analysis (PCA), were used to combine with machine learning techniques for synthesizing personalized HRTFs [8][9][10][11][12][13]. Before the era of deep neural network (DNN), statistical analysis, regression analysis and support vector regression were also adopted to estimate important parameters for synthesizing HRTFs or the HRTFs directly [14][15][16]. Then, of course, a DNN was recently proposed to synthesize head-related impulse responses (HRIRs) [17].

When filtering an audio wave, multiplying in the frequency domain requires less computation than convolving in the time domain. Besides, it has been shown that HRTFs possess quasi-linear phases which can be well captured by the head size [18]. Therefore, we focus on HRTFs rather than HRIRs and estimate phase responses and magnitude responses of HRTFs separately. Similar to the approach in [17], we construct a DNN to predict the magnitude response of the HRTF in each direction with a specific ear. Using the CIPIC HRTF database [1], which includes measurements from 25 azimuth angles, 50 elevation angles and both ears, we need to train 2500 (25x50x2) DNNs to predict magnitude responses of overall HRTFs of a user. To save the training time, the similarity of magnitude responses of HRTFs between adjacent directions was considered when initializing the DNNs. In this paper, we propose a training scheme in which final weights of a trained DNN of the adjacent direction are used as the initial weights of the current DNN. In this way, the DNNs for predicting magnitude responses of HRTFs converge much faster during the training phase.

The rest of th paper is organized as follows. In Section 2, we will illustrate the preprocessing conducted on the CIPIC database, the DNN model in details, and the proposed training scheme. In Section 3, we will compare the proposed training scheme with the baseline training scheme using random initials. Finally, we give conclusions in Section 4.

## II. Proposed Method

We first briefly introduce the database we used and the preprocessing for constructing DNN models. Later, we illustrate each DNN model and the characteristics of HRTFs that inspire our proposed training scheme.

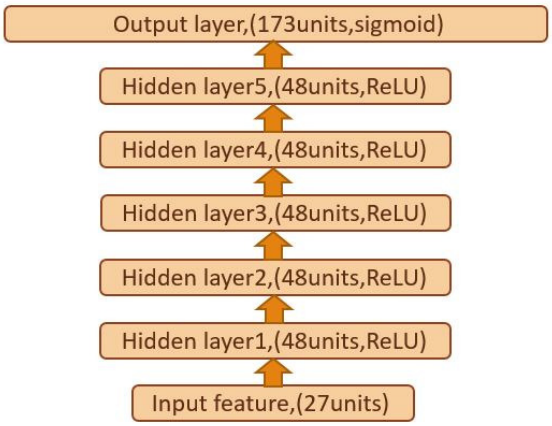



Fig. 1. The DNN model for predicting magnitude responses of HRTFs of a specific direction.

### A. Training data and preprocessing

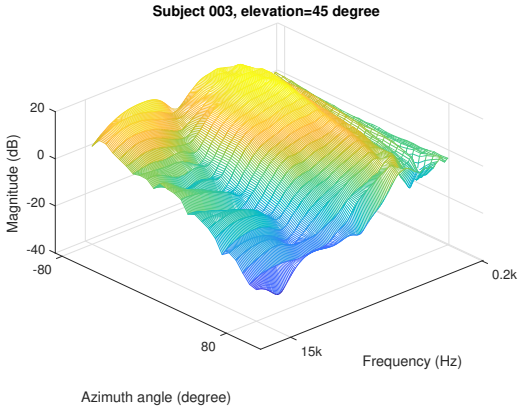
In this work, we used the CIPIC HRTF database [1]. This database contains HRTFs of 45 subjects. We only selected 35 subjects who have complete anthropometry measurements. Each subject's data contains HRIRs of 1250 directions for each of the two ears, and 37 anthropometry features (17 features of torso and head, and 10 features of each pinna). Before using the features to train the DNN model, we conducted some preprocessing on the data.

*1) Anthropometry features:* Without loss of generality, we trained DNNs to predict the magnitude responses of HRTFs of the left ear in this paper. The input feature set to our DNN model is a 27-dimensional vector which includes 10 measurements of the left pinna (cavum concha height, cymba concha height, cavum concha width, fossa height, pinna height, pinna width, intertragal incisure width, cavum concha depth, pinna rotation angle, and pinna flare angle) and 17 measurements of the torso and head. Firstly, we normalized each input feature by following procedures in [17] as

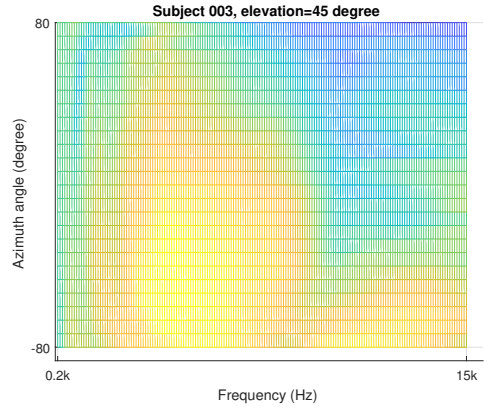
$$x_i' = \left(1 + e^{-\frac{(x_i - \mu_i)}{\sigma_i}}\right)^{-1} \quad (1)$$

where $x_i$ is the $i-th$ feature and $\mu_i$ and $\sigma_i$ are the mean and standard deviation of the $i-th$ feature, respectively. We adopted $\{x_i', i = 1, 2...27\}$ as the input features to each DNN model.

*2) HRIRs and HRTFs:* Using the tool provided by the CIPIC database, we conducted 512-point FFT on HRIRs, smoothed the magnitude responses using a constant-Q filter bank (Q=8) and took logarithm on the results to produce magnitude responses of HRTFs in dB. Following procedures in [13], we also retained HRTFs between 200 Hz and 15 kHz, hence, the magnitude response of each HRTF was comprised of 173 points. Since we adopted the sigmoid function as the activation function of the output layer in our DNN models, we also normalized the log. magnitude of HRTFs to values between 0 and 1. Without loss of generality, we used magnitude responses of HRTFs of the left ear from all directions with the $45^\circ$ elevation angle in simulations.




(a)




(b)

Fig. 2. Log. magnitude of HRTFs of the subject 003 in the CIPIC database at the elevation angle of 45 degree ($\phi = 45^\circ$) shown in the 3D plot (panel (a)), and in the 2D plot (panel (b)). Note that the plots show the original values before being normalized to values between 0 and 1 for DNN training.

### B. DNN model

For a specific subject, his HRTFs vary across different directions. For a specific direction, the HRTFs vary across different subjects. In this work, we followed the DNN approach in [17] which uses the DNN to characterize variations of HRIRs between different subjects for a specific direction. Therefore, to predict overall HRTFs of a new subject using the CIPIC database, we need to construct 2500 (25 azimuth angles, 50 elevation angles and both ear) DNN models separately.

The architecture of the DNN model we used to estimate magnitude responses of HRTFs for a specific direction is shown in Fig. 1. The model consists of 5 hidden layers and an output layer. Each hidden layer has 48 units with ReLU as the activation function. The activation function of output layer

is the sigmoid function. During training, mean-squared-error (MSE) is chosen as the cost function and the adaptive moment estimation (ADAM) technique with the learning rate of 0.001 is used for optimization. Besides, to avoid over-fitting, we set the dropout rate to 0.9. The termination conditions for the training process were set as

$$\mathbb{E}\{\, MSE_i \,\}_{i=n-4\sim n} - \mathbb{E}\{\, MSE_i \,\}_{i=n-9\sim n-5} \geq \epsilon \quad and \quad \mathbb{E}\{\, MSE_i \,\}_{i=n-9\sim n-5} - \mathbb{E}\{\, MSE_i \,\}_{i=n-14\sim n-10} \geq \epsilon \tag{2}$$

where $\epsilon$ is $5 \times 10^{-6}$, $i$ is the iteration index, and $n$ is the current iteration.

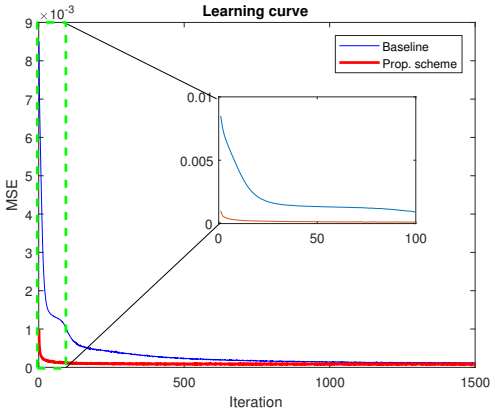



Fig. 3. Learning curves of compared DNN training schemes for estimating magnitude responses of HRTFs for the direction of ($\phi = 45°$, $\phi = -65°$). The green dotted box is reploted in the middle of the figure.

### C. Proposed training scheme

As mentioned above, we need to construct 2500 DNNs for predicting overall HRTFs of a user. A slight change in the dataset will result in re-training the 2500 DNNs, which is time-consuming. Fig. 2 shows sample magnitude responses of HRTFs measured at the same elevation angle ($\phi$) of 45 degree of a particular subject in the CIPIC database. As can be seen, these magnitude responses change smoothly along azimuth angles ($\theta$). Since the magnitude responses of HRTFs look similar across nearby azimuth angles, we can use the trained weights of the DNN for the previous azimuth angle to initialize the weights of the DNN for the current azimuth angle to save lots of training time. For instance, in our experiments, we first trained a DNN to estimate magnitude responses of HRTFs for the direction of ($\phi = 45°$, $\theta = -80°$). Later, we used the resulting weights to initialize the DNN for the direction of ($\phi = 45°$, $\theta = -65°$). This process was repeatedly conducted across azimuth angles to estimate magnitude responses of HRTFs for all directions.

## III. Experiment Results

In this section, we will compare the baseline (randomly initialized DNN) with the DNN by our proposed training scheme. We use the learning curve and actual training time to demonstrate the efficiency of the proposed scheme. Besides, the log spectral distortion (LSD) measure combined with the analysis of variance (ANOVA) test is used to evaluate the fidelity of prediction from the DNN with the proposed training scheme.

TABLE I
AVERAGE NUMBERS OF ITERATIONS NEEDED FOR CONVERGENCE BY THE COMPARED TWO TRAINING SCHEMES FOR VARIOUS DIRECTIONS AT 45° ELEVATION ANGLE

| Azimuth angle | $-65°$ | $-35°$ | $-10°$ | $10°$ | $35°$ | $65°$ |
|---|---|---|---|---|---|---|
| Baseline | 9105 | 9291 | 9528 | 9742 | 10134 | 12560 |
| Prop. scheme | 78 | 64 | 53 | 37 | 21 | 23 |

### A. Learning curve

Fig. 3 shows the learning curves of compared DNN training schemes for estimating magnitude responses of HRTFs for the direction of ($\phi = 45°$, $\phi = -65°$). Data of 34 randomly selected subjects were used to train the DNNs and data of the remaining subject was used for test. The red line is the learning curve of our approach, which used the trained weights of the DNN at ($\phi = 45°$, $\phi = -80°$) for initialization. The figure clearly shows the proposed approach converges much faster than the baseline approach. In addition, we conducted leave-one-out cross validation and averaged the results. The average numbers of iterations needed for convergence are listed in Table I for several directions. Results in the table show the DNN with the proposed training scheme needs less and less training time when more and more directions are scanned. In average, the proposed training scheme can dramatically reduce the iteration number by more than 99% if trained weights of a nearby direction are available for initialization.

### B. Training time

All DNNs were trained using a server with the NVIDIA GeForce GTX 1080 video card. The 1607 MHz GPU has 2560 NVIDIA CUDA cores and is equipped with 8 GB GDDR5X memory. Table II shows training time needed to estimate magnitude responses of HRTFs for 25 directions at the same elevation angle $\phi = 45°$ by the compared training schemes. In our proposed scheme, the first DNN per elevation angle is still randomly initialized such that training time only drops about 95%, not 99%, compared with the baseline scheme.

TABLE II
TRAINING TIME NEEDED FOR CONVERGENCE BY THE COMPARED TRAINING SCHEMES

| Method | Baseline | Prop. scheme |
|---|---|---|
| Training time | 1059.9 sec | 43.72 sec |

### C. Fidelity evaluation

In the research field of HRTF estimation, LSD is usually used to evaluate the performance of the proposed method. It is formulated as follows

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

where k is the index of frequency bin. However, during the pre-processing, we already took 20 $log_{10}(.)$ to calculate the log. magnitude of HRTFs. As a result, we used the following equation to calculate LSD for our settings. Here, $Y(k) = 20\ log_{10}H(k)$.

$$LSD(\mathrm{Y},\hat{\mathrm{Y}}) = \sqrt{\frac{1}{k_2-k_1+1}\sum_{k=k_1}^{k_2}\left(Y(k)-\hat{Y}(k)\right)^2} \quad (4)$$

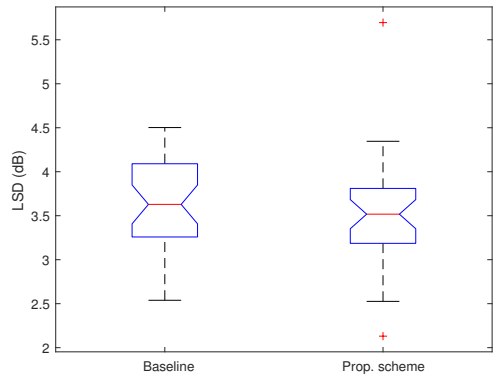



Fig. 4. Analysis of variance (ANOVA) of two compared training schemes with the p-value of 0.2956.

For each training scheme, we trained 25 DNN models to estimate magnitude responses of HRTFs from all azimuth angles at the 45-degree elevation ($\phi = 45°$), and then computed the average value of LSD over all azimuth directions. We also conducted leave-one-out cross validation and obtained 35 average LSD values for each training scheme. Fig. 4 shows the box plot of the two compared training schemes. The red line is the median value (3.6287 for the baseline scheme and 3.5176 for our proposed scheme, respectively) and the boxes indicate quarter quantile. We also obtained the p-value of 0.2956 by ANOVA test, which indicates the two approaches did not produce significant different results. In other words, the proposed training scheme produces results with similar fidelity as results from the baseline training scheme.

## IV. Conclusions

For personalizing HRTFs, we constructed a DNN for each direction and thousands of DNNs are needed to cover all directions. With possible frequent re-training due to changes in the dataset, long training time becomes troublesome. By exploring the redundancy of HRTFs among nearby azimuth angles, we propose a training scheme to speed up the training process using the trained DNN weights of the adjacent azimuth angle to initialize the DNN of the current azimuth angle. Simulation results show that the proposed training scheme can reduce training time by more than 95% without degrading the fidelity of the prediction results comparing with the baseline DNN approach. In addition to long training time, the small sizes of available HRTF dataset cause another concern for DNN-based approaches. In the future, we will work on modifying the DNN model, which can be used to combine different HRTF dataset to enlarge the training set, to produce prediction results with high fidelity.

## Acknowledgment

This research is supported by the Ministry of Science and Technology, Taiwan under Grant No MOST 105-2221-E-009-152-MY2.

## References

[1] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, pp. 99–102, 2001.

[2] W. G. Gardner and K. D. Martin, "Hrtf measurements of a kemar," *Acoustical Society of America Journal*, vol. 97, pp. 3907–3908, 1995.

[3] C. P. Brown and R. O. Duda, "An efficient hrtf model for 3-d sound," *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 4 pp.–, 1997.

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

[5] D. Y. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "Hrtf personalization using anthropometric measurements," *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, pp. 157–160, 2003.

[6] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt, "Hrtf magnitude synthesis via sparse representation of anthropometric features," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4468–4472, 2014.

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

[8] 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*, pp. 1637–1647, 1992.

[9] K. J. Fink and L. Ray, "Tuning principal component weights to individualize hrtfs," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 389–392, 2012.

[10] ——, "Individualization of head related transfer functions using principal component analysis," *Applied Acoustics*, vol. 87, pp. 162 – 173, 2015.

[11] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

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

[13] 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*, vol. 24, no. 3, pp. 559–570, 2016.

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

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

[16] Q. Huang and Q. Zhuang, "Hrir personalisation using support vector regression in independent feature space," *Electronics Letters*, vol. 45, pp. 1002 – 1003, 2009.

[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," *Audio Engineering Society Convention 143*, 2017.

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