Audio Engineering Society

Convention Paper 10502

Presented at the 150th Convention,Online
2021 May 25–28,

Global HRTF Personalization Using Anthropometric Measures

Yuxiang Wang[1], You Zhang[1], Zhiyao Duan[1], and Mark Bocko[1]

[1]*Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, USA*

Correspondence should be addressed to Yuxiang Wang (ywang310@ur.rochester.edu)

ABSTRACT

In this paper, we propose an approach for global HRTF personalization employing subjects' anthropometric features using spherical harmonics transform (SHT) and convolutional neural network (CNN). Existing methods employ different models for each elevation, which fails to take advantage of the underlying common features of the full set of HRTF's. Using the HUTUBS HRTF database as our training set, a SHT was used to produce subjects' personalized HRTF's for all spatial directions using a single model. The resulting predicted HRTFs have a log-spectral distortion (LSD) level of 3.81 dB in comparison to the SHT reconstructed HRTFs, and 4.74 dB in comparison to the measured HRTFs. The personalized HRTFs show significant improvement upon the finite element acoustic computations of HRTFs provided in the HUTUBS database.

1 Introduction

The Head-related Transfer Function (HRTF), a description of how a human receives sound from various spatial directions [1], is unique to each listener and is vital for accurate virtual acoustic display [2]. Due to its uniqueness, using a generic HRTF for virtual acoustic display may result in compromised results, leading to diffuse or displaced auditory images. Studies have shown that individualized HRTF can improve the localization accuracy and users' immersive experiences [3, 4]. Ideally one would like to have a personalized HRTF for every listener, however, measuring a subject's HRTF requires specialized equipment and is a time-consuming process [1]. Thus, it is desirable to obtain personalized HRTFs without the need for making extensive acoustic measurements.

An individual's HRTF is comprised of a set of acoustic transfer functions that contain both spatial and temporal information. Recent efforts in HRTF personalization from physical appearances [5, 6, 7] mostly were based on data-driven approaches, aided by acoustics modeling. To efficiently link a person's physical features to the features of HRTF, a great amount of dimension reduction is needed, as the data represented in an HRTF has a much higher dimensionality in comparison to anthropometric measurements. To address this issue, some existing work on HRTF personalization focuses only on a small portion of HRTF directions (e.g., frontal or 0 degree elevations) or uses different machine learning models for different azimuths and elevations. In addition, the source location grid used in many pub-

licly available HRTF databases differs from database to database, which limits the format of HRTF prediction results to the given source grids. Therefore, it would appear to be advantageous to include the information of all spatial directions into a single model, regardless of the grid employed in the HRTF training dataset.

An HRTF can be viewed simply as a function defined on a spherical surface. The Head-Related Impulse Response (HRIR) (simply the Fourier transform of the HRTF) for each direction is made between an external sound source to a microphone at the entrance of the ear canal [1, 8]. Employing acoustic reciprocity principle, the same impulse responses would be obtained if the source were at the opening of the subject's ear canal, and receivers were located at each of the sound source locations. To represent such patterns on a spherical surface, spherical harmonics (SH) basis functions are a natural choice.

It has been shown that human listeners are insensitive to smoothing of fine structure in their HRTF. Previous research employed various smoothing methods to validate this observation with perceptual tasks. Kulkarni and Colburn [2] proposed a smoothing method based on coefficient truncation of the Fourier series expansion to the log-magnitude spectrum of HRTFs. Hacihabiboglu et al. [9] proposed wavelet-based spectral smoothing in HRTF filter design, and Romigh et al. [10] designed an efficient spherical harmonic transform based HRTF smoothing representation. These results show that smoothed versions of a listener's HRTF remain perceptually relevant, and suggest the viability of lower-order representations of HRTF's.

Based upon this previous work, we hypothesize that a HRTF reconstructed from a truncated SH expansion are perceptually indistinguishable from the original HRTF. Since SH coefficients efficiently capture the global structure of an HRTF in a low-dimensional representation, we believe it may serve as an effective tool for personalizing HRTF's.

In this paper, we propose an HRTF personalization method for arbitrary directions using a spherical harmonics transform (SHT) representation. We use anthropometric measurements provided by the HUTUBS database [11] and frequency information to predict SH coefficients. Predicted HRTF's were reconstructed for test data taken from the HUTUBS dataset and the deviations from measured HRTF's were assessed.

The remainder of the paper is organized as follows: In Section 2, we present related work on HRTF personalization. We describe our proposed method in Section 3 and experimental details are given in 4. The results and their assessment are given in Section 5, and we conclude our paper in Section 6.

2 Related Work

2.1 Global HRTF representation

Given the high dimensionality of HRTFs and wide variation of anthropometric features across different subjects, it is challenging to predict a person's entire HRTF set, thus the motivation to seek a low-dimensional representation of the HRTF. One may assume that it would be possible to find a low-dimensional representation since there is redundancy in the set of HRTF's [2, 12]. Various approaches to dimension reduction have been explored, such as principal component analysis (PCA) [13, 14] and acoustic pole & zero models [15]. Among these, a truncated spherical harmonics representation seems to provides an intuitive method for representing the most salient features of an HRTF [16, 17, 18].

Specifically, the SHT preserves global spatial features in a low order, compact representation of the entire HRTF set. Following previous work [18], we adopt the SHT representation to achieve dimension reduction.

2.2 HRTF personalization

HRTF personalization using anthropometric measures is an emerging topic of interest and methods using anthropometric parameter matching [19, 20, 21], spectral notches [22], or pinna shape [23] have been described.

Several investigations of HRTF personalization employing machine learning also have been published. The common practice is to learn a low-dimensional representation of HRTF and then to predict that representation for test data using anthropometric features. Other researchers proposed regression algorithms to predict HRTF dimension reduced by PCA [24] and Isomap [25].

Deep learning methods have pushed the limits of HRTF personalization in recent years. Chun et al. [5] proposed a deep neural network (DNN) based method that predicts the head-related impulse response (HRIR) using anthropometric measurements. Lee & Kim [26]

proposed a deep learning approach to personalize HRTFs using anthropometric measurements and images of the ears. Chen et al. [6] proposed a DNN-based approach but used different models for different directions. They first trained an encoder-decoder network to learn the latent representation of a set of HRTFs, then a DNN is trained to map the anthropometric measurements to the latent representation. Finally, the DNN is fine-tuned jointly with the decoder. Miccini & Spagnol [7] extended their method with the use of a convolutional variational autoencoder and includes depth maps of a 3D head model as input. However, the three-stage training in this approach is cumbersome and may lead to suboptimal solutions.

Although some of the work described above has achieved acceptable results for predicting HRTF's, they are limited to HRTF prediction for certain directions. To our knowledge, global HRTF prediction has not been accomplished with a single model. The method in [7] can be extended to make global predictions by using a separate model for each direction, but this would be complex and does not consider the intrinsic connection between different directions. Recently, Zhang et al. [27] proposed HRTF personalization modeling in arbitrary spatial directions based on spatial PCA. As PCA methods may ignore the intrinsic connection between spatial information and undermine the phase warping, a large number of principal components are needed to maintain the variance in reconstructed HRTFs. In contrast, SH-based methods offer more compact HRTF representation. It was shown in [18] that using 4th order SHT (i.e. with 25 SH coefficients), the reconstructed HRTF is perceptually indistinguishable from the original measured HRTF. Therefore, the SH basis apparently is an efficient means to capture the important spatial features of HRTFs.

3 Method

We adopt an SH-based method to extract low-dimensional features to represent HRTF. The HRIR is processed in a perception-inspired way to obtain the HRTF at different frequencies. A deep learning model is designed to predict the SH coefficients using anthropometric measurements and frequency information.

3.1 Feature extraction with SHT

Spherical harmonics (SH) are a set of orthogonal bases for the spherical coordinate system and have been

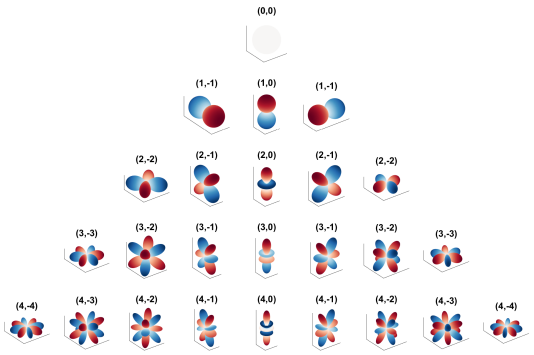


Fig. 1: SH bases up to $L = 4$. Numbers in parenthesis are order l and degree m, where $-l \leqslant m \leqslant l$. Note that the total number of bases is $(L+1)^2$.

widely adopted in the field of spatial audio. The spherical harmonic basis of l-th order and m-th degree at a certain spatial location is computed as

$$Y_l^m(\theta,\varphi) = \sqrt{\frac{(2l+1)}{4\pi}\frac{(l-m)!}{(l+m)!}} P_l^m(\cos\theta) e^{im\varphi}, \quad (1)$$

where θ, φ are the azimuth and elevation angles in the spherical coordinate system. $P_l^m(\cos\theta)$ is associated Legendre polynomial. The real parts of SH bases of the first 4 orders are listed in Figure 1.

The process of SHT is to compute coefficients of each SH basis function, and in practice we follow the method in [18]. The SH coefficients are estimated by solving a system of linear equations (2) using S discretized samples, one for each spatial location $\{\theta_i, \varphi_i\}^S$:

$$\boldsymbol{f} = \boldsymbol{Y}\boldsymbol{c} \quad (2)$$

where

$$\begin{aligned} \boldsymbol{f} &= [f(\theta_1,\varphi_1),\ldots f(\theta_S,\varphi_S)]^T \\ \boldsymbol{c} &= [C_{00}, C_{1-1}, C_{10}, C_{11}, \ldots C_{LL}]^T \\ \boldsymbol{Y} &= [\boldsymbol{y}_{00}, \boldsymbol{y}_{1-1}, \boldsymbol{y}_{10}\boldsymbol{y}_{11}, \cdots \boldsymbol{y}_{LL}] \\ &\text{and} \\ \boldsymbol{y}_{lm} &= [Y_{lm}(\theta_1,\varphi_1),\ldots,Y_{lm}(\theta_S,\varphi_S)]^T \end{aligned} \quad (3)$$

In Eq. (3), $\boldsymbol{f}$ contains the original magnitude values on S spatial directions, $\boldsymbol{c}$ is the desired SH coefficients,

and $\boldsymbol{Y}$ contains SH base values of up to order L at the corresponding source spatial directions. To compute coefficient vector $\boldsymbol{c}$, we used least square fit approach:

$$\boldsymbol{c} = \left(\boldsymbol{Y}^T\boldsymbol{Y}\right)^{-1}\boldsymbol{Y}^T\boldsymbol{f} \tag{4}$$

In our approach, a SHT was performed for each frequency to truncation order $L = 7$, which is higher than the perceptual spatial resolution according to previous related work [18]. For each frequency of a subject's HRTF, a magnitude operation is performed to compensate for the perceptual sensitivity of loudness. We use the real part of the spherical harmonics to perform the SHT on each HRTF magnitude pattern, and obtain the coefficients ($\boldsymbol{c}$ vector) of each SH base. By concatenating SH coefficients of each frequency bin, we obtain a lower-dimensional representation of the HRTF dataset, which is used as the reference training target for deep learning.

3.2 Perception-inspired data preprocessing

To predict a statistical and perceptual viable result, we employ our understanding of the human auditory system in the data processing steps to convert the HRIR's to HRTF's. Specifically, we employ the concept of critical bands that describe the frequency bandwidth of the effective auditory filters of the cochlea [28], which play an important role in auditory masking. Moreover, for loudness perception, the human auditory system responds logarithmically. These perceptual features were incorporated in the data pre-processing step. The original HRTF data set provided the measured impulse response in sofa formats, and the corresponding frequency magnitude value was converted to a dB scale and sampled at the center frequencies of the auditory critical bands.

3.3 Deep learning model design

A deep learning-based model is designed to map the anthropometric measurements to the SH representation of the HRTF. The model structure is illustrated in Figure 2. We consider the frequency and ear as side information and feed them into fully connected (FC) layers to obtain their embedding. The ear, head, and torso measurements are also fed into FC layers to encode that information. We then concatenate these output encodings and use another FC layer to fuse the information. Then the latent encoding is fed into several layers of the 1D convolutional neural network (CNN) to predict the SH coefficients. The training loss is calculated with the mean square error (MSE) of the ground-truth SH coefficients and the predicted SH coefficients.

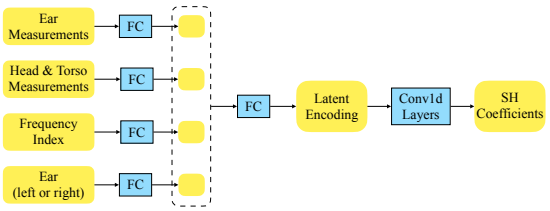


Fig. 2: The data-flow diagram of predicting SH coefficients. The predicted SH coefficients are then used to reconstruct HRTF through inverse SHT.

4 Experiments

4.1 Dataset

We use the HUTUBS dataset [11] to perform our experiments. It has 96 subjects where 93 of them have complete anthropometric measurements, so we use the 93 subjects to construct our dataset. It also provides an acoustic boundary element method simulated version of the HRTF for each subject.

A particular advantage in choosing this database is its 440-point near-uniform sampling scheme that covers the complete spherical surface, which guarantees non-aliasing of the SHT for up to 16th order. The number of subjects is large compared to existing datasets, which would benefit the training of deep learning models. The simulated HRTF in HUTUBS also enables us to investigate whether our prediction results are better than the simulation results in terms of spectral distortion.

4.2 Implementation details

For anthropometric measurements, we normalize the original measures according to the procedure in [6]:

$$\bar{x}_i = \left(1 + e^{-\frac{(x_i - \mu_i)}{\sigma_i}}\right)^{-1} \tag{5}$$

where x_i is the i-th measure of the ear or head & torso measurements, and μ_i and σ_i are the mean and standard deviation across all the training subjects, respectively.

For each subject in the HUTUBS dataset, the ear measurement is a 12-d vector for each ear and the head & torso measurement is a 13-d vector. When estimating the SH coefficients of one ear, we ignore the ear measurement from the other ear. The normalized measurement vectors are then fed into corresponding FC layers. The frequency index is encoded as a one-hot vector to be fed into the FC-Frequency layer. Another one-hot vector indicating whether the anthropometric measurements are from the left or right ear is fed into the FC-EarLR layer. The details of our model and the hyperparameter of the corresponding layers are set as in Table 1.

Table 1: Details of the architecture of the proposed deep network. (Each 1D convolution layer is followed by layer normalization [29] and rectified linear unit (ReLU) activation [30].)

Layers	Kernel	Stride	Output Shape
FC-EarMeasure	/	/	[B, 32]
FC-HeadMeasure	/	/	[B, 32]
FC-Frequency	/	/	[B, 16]
FC-EarLR	/	/	[B, 16]
FC-Fusion	/	/	[B, 256]
Unsqueeze	/	/	[B, 1, 256]
Conv1D-1	7	3	[B, 4, 84]
Conv1D-2	5	2	[B, 16, 40]
Conv1D-3	5	2	[B, 32, 18]
Conv1D-4	5	3	[B, 32, 5]
Conv1D-5	5	2	[B, 64, 1]

We implement our deep learning method with PyTorch. The batch size B is set to 1024. The learning rate is initially set to 0.0005 with 20% decay for every 100 epochs. We train the network for 1000 epochs on a single NVIDIA GTX 1080 Ti GPU. The time cost for each training-evaluation round is around half an hour. Finally, we select the model with the lowest validation loss for evaluation. Our source code is released in `https://github.com/YuriWayne42/hrtf_sht_personalization`.

4.3 Objective evaluation

The commonly used log-spectral distortion (LSD) is adopted to evaluate the performance of our proposed method. The LSD can be formulated as:

$$LSD(\mathrm{H},\hat{\mathrm{H}}) = \sqrt{\frac{1}{SK}\sum_{s}\sum_{k}\left(20\log_{10}\left|\frac{H(s,k)}{\hat{H}(s,k)}\right|\right)^2} \quad (6)$$

where s indicates the spatial location, k indicates the frequency index. S and K are the numbers of spatial locations and frequencies, respectively. $H(s,k)$ and $\hat{H}(s,k)$ denote the magnitude of the ground-truth HRTF and the predicted HRTF, respectively in the linear scale. When the S spatial locations cover the entire discretized space, the LSD evaluates global performance. In our experimental setup, the HRTFs were processed using a dB scale to take account of the perceptual considerations described in Section 3.2. Therefore, our calculation of LSD is the root mean square error of the HRTF.

Instead of directly predicting HRTF, we predict the SH coefficients $\hat{\boldsymbol{c}}$. Hence, we multiply the SH base values $\boldsymbol{Y}$ with the coefficients $\boldsymbol{c}$ to compute the predicted HRTF according to Eq. (2). The ground-truth HRTF for comparison is also calculated from the ground-truth SH coefficients $\boldsymbol{c}$.

Since the HUTUBS dataset is not large enough to have the regular train/val/test split, we adopt leave-one-out cross-validation in our study as employed in [6].

5 Results and discussions

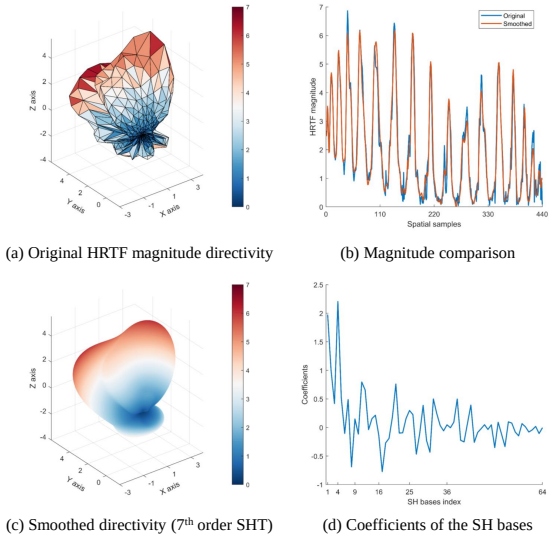
In this section, we evaluate the performance of our approach. We report the HRTF preprocessing results in 5.1, show the HRTF personalization results in 5.2, conduct an ablation study for global HRTF personalization in 5.3, and describe the limitations and future work in 5.4.

To clarify, we will use the following terms for demonstration. The *predicted* HRTF means the HRTF reconstructed from the predicted SH coefficients. The *smoothed* HRTF means the HRTF reconstructed with the ground-truth SH coefficients. The smoothed HRTF is not used during training. The *original* HRTF means the measured HRTF from the dataset.

Note that the LSD values we report are all averaged across subjects (i.e. across different training-evaluation rounds). As in a leave-one-out fashion, we leave one subject for test and use all others to train the deep learning model, for each training-evaluation round.

5.1 HRTF preprocessing results

To demonstrate that the SHT is an effective way to represent HRTFs with global information from all spatial directions, we plotted the SHT result performed on a certain frequency (around 7kHz) of the HRTF magnitude pattern of one subject in the database in Figure 3.



(a) Original HRTF magnitude directivity (b) Magnitude comparison

(c) Smoothed directivity (7th order SHT) (d) Coefficients of the SH bases

Fig. 3: Example of HRTF and SHT reconstruction result in linear scale.

In sub-figure (a), the HRTF pattern was plotted in a spherical coordinate system, assigning the magnitude as the distance from each corresponding source location to the origin. The colormap was also assigned according to the magnitudes. In (b) we show the comparison of magnitudes (original and the reconstruction) across all 440 source locations. In (c), the reconstructed pattern is plotted from the SHT of the 7th order, and (d) shows the results of $(L+1)^2 = 64$ SH coefficients.

Note that the reconstructed spatial formation is in fact smooth compared to the original values due to the truncation of the SHT. The SH truncation order was set to seven, which is still higher than the perceptual viable spatial resolution, according to previous related work [18]. Based on our calculation, the smoothed version introduced minimum spectral distortion, which also validates previous research that this distortion level is perceptual indistinguishable. For each frequency of a subject's HRTF, a magnitude operation is performed to compensate for the perceptual sensitivity of loudness, a set of 64 SH coefficients are produced according to the magnitude layout.

By observing the coefficients of the SH bases across different subjects, we hypothesize that the coefficients follow a normal distribution across subjects under one frequency and one ear condition, expressed in Eq. (7).

$$\boldsymbol{c} \sim N(\mu, \sigma \mid k) \tag{7}$$

We then conduct a normality test [31]. The p values for rejecting the null hypothesis that the SH coefficients come from a normal distribution under different frequencies are all less than 0.05, verifying our hypothesis in Eq. (7). We believe that with this property, it is easier for the deep learning model to predict SH coefficients, compared to directly predicting HRTFs using anthropometric measurements.

5.2 HRTF personalization results

To demonstrate the efficacy of personalization modeling, a comparison of a typical smoothed, predicted, and simulated HRTF is shown in Figure 4, as we examine the frontal direction where $s = (\theta, \varphi) = (0, 0)$.

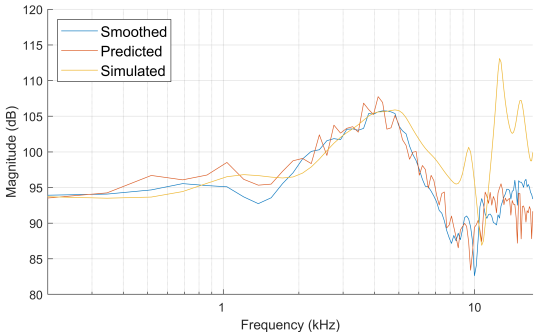


Fig. 4: Comparison of smoothed, predicted and simulated HRTF at the frontal direction, where $s = (\theta, \varphi) = (0, 0)$. Frequency and magnitude axes are both shown in the log scale.

From the figure, we show that our predicted HRTFs follow the trend of the smoothed HRTFs. The simulated version has a similar trend, but deviates significantly from the other two, especially at frequencies above 6 kHz. We calculate the LSD of our predicted HRTF and simulated HRTF, both compared with the smoothed HRTF, for all subjects. The predicted HRTF has 4.06

dB LSD while the simulated HRTF has 7.99 dB LSD. While the acoustic simulation results are affected by the limited resolution of the geometric model and the accuracy of simulation solver, the method explored here is free from this limitation and seems to produce better results.

To verify the global performance of the prediction across all spatial directions, we report the predicted result with the smoothed HRTF, averaged across subjects. Although our method is able to predict HRTFs for arbitrary spatial directions, we evaluate our error according to the measured HRTF source grid. The LSD we achieved is 3.81 dB compared to the smoothed HRTF. Even compared with the original HRTF, the LSD still achieves 4.74 dB. Note that due to the difference in sampling grid in the simulation, we do not include simulated HRTFs in the global comparison.

Since our method can produce personalized results for arbitrary direction, we also examine the LSD performance in different spatial directions. The standard deviation of LSD across all source locations is 0.30 dB. This shows that the performance is robust to direction changes, validating the nature of SHT as a viable global representation method.

To the best of our knowledge, there has been little or no work performing HRTF personalization using the HUTUBS dataset. We notice that [6] is the closest work since they use only anthropometric data to predict HRTFs, although they use a different dataset CIPIC [8]. However, they only perform HRTF personalization for one elevation angle instead of global, and they use different models for different azimuth angles. They achieved 3.25 dB LSD when comparing with smoothed HRTF magnitude spectra with a constant-Q filterbank. Note that the dataset is different, but we believe that our method has a comparable level of performance, and has the advantage of being able to make global HRTF predictions.

5.3 Ablation study

To further verify the effectiveness of using SH coefficients for HRTF prediction, we compare the result of predicting SH coefficients and reconstructed HRTFs, versus directly predict the HRTF without using the SHT. When directly predicting the HRTF, we employ the same model architecture as in Table 1, but we change the output channel of Conv1D-5 to the number of discretized spatial locations, instead of the number of SH bases. The training objective is accordingly changed to the original HRTF, rather than the SH coefficients.

The comparison result is shown in Table 2. The reported LSDs are compared with the original HRTF, not the smoothed HRTF by SHT reconstruction.

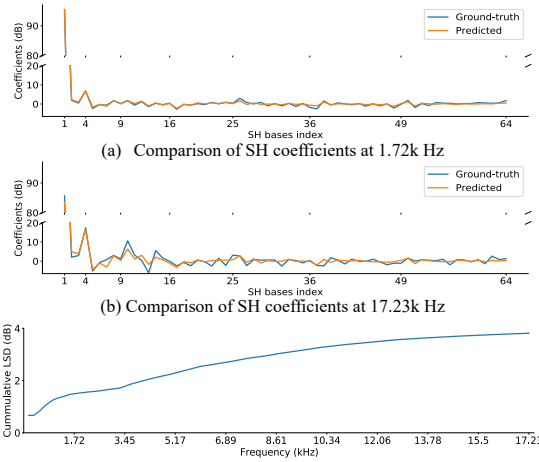
Table 2: Comparison of global HRTF personalization w/ and w/o SHT

Method	w/ SHT	w/o SHT
Global LSD	4.74	6.06

A pairwise statistical t-test across all subjects shows that the difference between our method and directly predicting HRTF is statistically significant, at the significance level of 0.05. From the experiment result, we conclude that SHT benefits the global HRTF prediction.

5.4 Limitations and future work

To investigate the predicted performance across frequencies, in Figure 5, we show the SH coefficients prediction result at a low frequency in (a), and at a high frequency in (b). In (c), we plot the cumulative global LSD up to different frequencies.



(a) Comparison of SH coefficients at 1.72k Hz

(b) Comparison of SH coefficients at 17.23k Hz

(c) Global LSD of reconstructed HRTF magnitude across frequencies

Fig. 5: SH coefficients prediction results and global cumulative LSDs up to corresponding frequencies

Comparing (a) and (b) in Figure 5, the prediction results at low frequency are generally good, while at the high frequency, the predicted SH coefficients deviate

from the ground truth, especially for the high-order basis functions. From sub-figure (c), we can see that the prediction error accumulates as the frequency increased. This may due to the fact that information contained in anthropometric measurements is insufficient for predicting fine structure patterns in HRTFs, especially at high frequencies.

In future work, we believe it would be interesting to investigate alternative representations for subjects' ear, head, and torso features. Given that the anthropometric measurements may not provide enough information to make accurate predictions, it may be worthwhile to explore using each subjects' head mesh to perform HRTF prediction, as it provides much more information than current measurements.

6 Summary

In this paper, we proposed a deep learning model for global HRTF personalization, using a spherical harmonics transform as a compact representation of HRTFs. A leave-one-out validation with the log-spectral distortion metric was used to evaluate the performance of the model. Our results showed that the predicted HRTFs have acceptable error values for all subjects, and were able to produce HRTFs for all directions at the same level of error performance, a benefit of the nature of our feature extraction method. Our predicted HRTFs have smaller errors than the acoustically simulated HRTFs provided by the HUTUBS database. We believe that the work described here is a promising method for future work on global HRTF personalization.

References

[1] Xie, B., *Head-related transfer function and virtual auditory display*, J. Ross Publishing, 2013.

[2] Kulkarni, A. and Colburn, H. S., "Role of spectral detail in sound-source localization," *Nature*, 396(6713), p. 747, 1998.

[3] Hu, H., Zhou, L., Ma, H., and Wu, Z., "HRTF personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, 69(2), pp. 163–172, 2008.

[4] Armstrong, C., Thresh, L., Murphy, D., and Kearney, G., "A perceptual evaluation of individual and non-individual HRTFs: a case study of the SADIE II database," *Applied Sciences*, 8(11), p. 2029, 2018.

[5] Chun, C. J., Moon, J. M., Lee, G. W., Kim, N. K., and Kim, H. K., "Deep neural network based hrtf personalization using anthropometric measurements," in *Audio Engineering Society Convention 143*, Audio Engineering Society, 2017.

[6] Chen, T.-Y., Kuo, T.-H., and Chi, T.-S., "Autoencoding HRTFs for DNN based HRTF personalization using anthropometric features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 271–275, IEEE, 2019.

[7] Miccini, R. and Spagnol, S., "HRTF individualization using deep learning," in *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 390–395, IEEE, 2020.

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

[9] Hacihabiboglu, H., Gunel, B., and Murtagh, F., "Wavelet-based spectral smoothing for head-related transfer function filter design," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Audio Engineering Society, 2002.

[10] Romigh, G. D., Brungart, D., Stern, R. M., and Simpson, B. D., "The role of spatial detail in sound-source localization: Impact on HRTF modeling and personalization." in *Proceedings of Meetings on Acoustics ICA2013*, volume 19, p. 050170, Acoustical Society of America, 2013.

[11] Manoj, D., Robert, P., Jan Joschka, W., Fabian, S., Daniel, V., Peter, G., and Stefan, W., "The HUTUBS head-related transfer function (HRTF) database," 2019.

[12] Kulkarni, A., Isabelle, S., and Colburn, H., "Sensitivity of human subjects to head-related transfer-function phase spectra," *The Journal of the Acoustical Society of America*, 105(5), pp. 2821–2840, 1999.

[13] Sodnik, J., Umek, A., Susnik, R., Bobojevic, G., and Tomazic, S., "Representation of head related

transfer functions with principal component analysis," in *Proceedings of the Annual Conference of the Australian Acoustical Society, NSW*, pp. 603–607, 2004.

[14] Hwang, S., Park, Y., and Park, Y.-s., "Modeling and customization of head-related transfer functions using principal component analysis," in *IEEE International Conference on Control, Automation and Systems*, pp. 227–231, IEEE, 2008.

[15] Haneda, Y., Makino, S., Kaneda, Y., and Kitawaki, N., "Common-acoustical-pole and zero modeling of head-related transfer functions," *IEEE Transactions on speech and audio processing*, 7(2), pp. 188–196, 1999.

[16] Evans, M. J., Angus, J. A., and Tew, A. I., "Analyzing head-related transfer function measurements using surface spherical harmonics," *The Journal of the Acoustical Society of America*, 104(4), pp. 2400–2411, 1998.

[17] Zotkin, D. N., Duraiswami, R., and Gumerov, N. A., "Regularized HRTF fitting using spherical harmonics," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 257–260, IEEE, 2009.

[18] Romigh, G. D., Brungart, D. S., Stern, R. M., and Simpson, B. D., "Efficient real spherical harmonic representation of head-related transfer functions," *IEEE Journal of Selected Topics in Signal Processing*, 9(5), pp. 921–930, 2015.

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

[20] Torres-Gallegos, E. A., Orduna-Bustamante, F., and Arámbula-Cosío, F., "Personalization of head-related transfer functions (HRTF) based on automatic photo-anthropometry and inference from a database," *Applied Acoustics*, 97, pp. 84–95, 2015.

[21] Shu-Nung, Y., Collins, T., and Liang, C., "Head-related transfer function selection using neural networks," *Archives of Acoustics*, 42(3), pp. 365–373, 2017.

[22] Iida, K., Ishii, Y., and Nishioka, S., "Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae," *The Journal of the Acoustical Society of America*, 136(1), pp. 317–333, 2014.

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

[24] Bomhardt, R., Braren, H., and Fels, J., "Individualization of head-related transfer functions using principal component analysis and anthropometric dimensions," in *Proceedings of Meetings on Acoustics 172ASA*, volume 29, p. 050007, Acoustical Society of America, 2016.

[25] Grijalva, F., Martini, L., Florencio, D., and Goldenstein, S., "A manifold learning approach for personalizing HRTFs from anthropometric features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3), pp. 559–570, 2016.

[26] Lee, G. W. and Kim, H. K., "Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear," *Applied Sciences*, 8(11), p. 2180, 2018.

[27] Zhang, M., Ge, Z., Liu, T., Wu, X., and Qu, T., "Modeling of individual HRTFs based on spatial principal component analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, pp. 785–797, 2020.

[28] Grantham, D. W., "Spatial Hearing and Related Phenomena," *Hearing*, p. 297, 1995.

[29] Ba, J. L., Kiros, J. R., and Hinton, G. E., "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[30] Nair, V. and Hinton, G. E., "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning (ICML)*, 2010.

[31] D'AGOSTINO, R. and Pearson, E. S., "Tests for departure from normality," *Biometrika*, 60(3), pp. 613–622, 1973.