



Head-related Transfer Function Reconstruction with Anthropometric Parameters and the Direction of the Sound Source

Deep Learning-Based Head-Related Transfer Function Personalization

Dongdong Lu¹ · Xiangyang Zeng¹ · Xiaochao Guo² · Haitao Wang¹

Received: 9 June 2020 / Accepted: 20 October 2020 / Published online: 5 November 2020
© Australian Acoustical Society 2020

Abstract

An accurate head-related transfer function can improve the subjective auditory localization performance of a particular subject. This paper proposes a deep neural network model for reconstructing the head-related transfer function (HRTF) based on anthropometric parameters and the orientation of the sound source. The proposed model consists of three subnetworks, including a one-dimensional convolutional neural network (1D-CNN) to process anthropometric parameters as input features and another network that takes the sound source position as input to serve as a marker. Finally, the outputs of these two networks are merged together as the input to a third network to estimate the HRTF. An objective method and a subjective method are proposed to evaluate the performance of the proposed method. For the objective evaluation, the root mean square error (RMSE) between the estimated HRTF and the measured HRTF is calculated. The results show that the proposed method performs better than a database matching method and a deep-neural-network-based method. In addition, the results of a sound localization test performed for the subjective evaluation show that the proposed method can localize sound sources with higher accuracy than the KEMAR dummy head HRTF or the DNN-based method. The objective and subjective results all show that the personalized HRTFs obtained using the proposed method perform well in HRTF reconstruction.

Keywords Head-related transfer function · Personalization · Point-net network · Deep neural network (DNN) · Convolutional neural network (CNN)

1 Introduction

In recent years, technologies for virtual reality and augmented reality have undergone rapid development. Virtual stereo, as an important component of virtual reality that can

be played back through either speakers or headphones, has also attracted the attention of researchers [1]. For headphone playback, the head-related transfer function plays a major role.

The head-related transfer function (HRTF), or the equivalent head-related impulse response (HRIR) in the time domain, describes how sound is filtered by the head, torso, and pinnae of a listener as it propagates from the source to the listener's eardrum in free space [2]. The listener obtains a sense of space based on the interaural time differences (ITDs) and interaural level differences (ILDs) at various positions. Since the head, torso, and auricle structures of different listeners are different, to achieve better auditory perception, each listener's HRTF needs to be personalized [3].

Researchers have proposed a variety of HRTF personalization methods, including measurement methods [4], anthropometric parameter matching methods [5–7], numerical methods [8], and anthropometric parameter regression

✉ Xiangyang Zeng
zengxy@nwpu.edu.cn

Dongdong Lu
1157835731@qq.com

Xiaochao Guo
guoxiaochao1965@163.com

Haitao Wang
wht@nwpu.edu.cn

¹ Northwest Polytechnical University, No. 127 Youyixi Road, Beilin District, Xi'an 710072, Shaanxi, People's Republic of China

² Air Force Medical Center of FMMU, Beijing 100142, People's Republic of China

methods [9–12]. Among them, anthropometric parameter regression methods have received the most attention.

Regarding the selection of anthropometric parameters, different combinations of anthropometric parameters have been chosen in different studies, which can naturally have a great impact on the ultimate customized model. To date, there is still no clear conclusion regarding which anthropometric parameters to choose for modeling.

In recent years, deep learning has played an increasingly significant role in multiple industries, such as manufacturing, finance, and medical care. This is mainly due to the rapid advancement of big data capabilities, which make it easier to obtain various types of data. Thus, as existing model systems and optimization methods continue to advance, the performance of models can be continually improved. In addition, deep learning makes it easy to extract relevant features, allowing complex mathematical models to be described in terms of simple inputs and outputs.

Deep learning considers inputs and outputs individually but does not consider the relationship between the inputs and outputs. Numerous studies have used machine learning to describe the relationship between anthropometric parameters and the HRTF. Hu et al. [13] used a backpropagation (BP) neural network to reveal the relationship between anthropometric parameters and the HRTF. This model first reduces the dimensionality of the HRTF, then filters out unrelated anthropometric parameters, and finally uses the BP network to realize the regression of the anthropometric parameters with respect to the HRTF. Chun et al. [14] proposed a deep neural network (DNN) model to achieve a direct connection between anthropometric parameters to HRIRs instead of extracting both the anthropometric parameters and the features of the HRTF. The operational steps of that model were simplified by virtue of the intent to reveal the relationship between the anthropometric parameters and the HRIR. This DNN model was further optimized through integration with a convolutional neural network (CNN), as proposed by Lee et al. [15]. However, this integrated model only convolves the anthropometric parameters of the pinnae; moreover, the model is built in only one direction, meaning that 1250 models in total are needed to obtain HRTFs for all directions in the CIPIC database.

In this paper, an acoustic model that determines the head-related impulse responses (HRIRs) based on both the anthropometric parameters and the direction of the sound source as training parameters is proposed as an alternative to an acoustic model that needs to be separately trained for each orientation of the sound source. The anthropometric parameters used in this model are three-dimensional parameters obtained through measurement. When sound waves are transmitted from different directions, the anthropometric parameters that serve to filter the sound waves are also different; therefore, the HRIRs are significantly related to

Table 1 The number of HRTF measurements at different elevations and azimuth intervals

ϕ	$\pm 40^\circ$	$\pm 30^\circ$	$\pm 20^\circ$	$\pm 10^\circ$	0°	50°	60°	70°	80°
$\Delta\theta$	$6^\circ/7^\circ$	6°	5°	5°	5°	8°	10°	15°	30°
N	57	61	73	73	73	47	31	25	13

both the anthropometric parameters and the location of the sound source. Existing regression methods for human physiological HRTF parameters also support this conclusion.

To reduce the complexity of existing models arising from the need to build a different model for each sound-source direction, we propose a deep learning model. The model relies on a point network [16] and takes the anthropometric parameters in combination with the direction of the sound source as input, which greatly reduces the number of neural networks that need to be built (the CNN + DNN method requires 1250 individual networks) and shortens the time spent on model training. The proposed model uses deep learning to mine the relationship between the anthropometric parameters and the HRTF, thereby achieving HRTF personalization in a single network.

The performance of the proposed method is evaluated both objectively and subjectively. As an objective measure, the root mean square error (RMSE) between the reference and estimated HRTFs is calculated.

2 Database Used

All the researches in our study were conducted with the Chinese pilots HRTF database [17]. The database contains the measured HRTF of 63 subjects in 723 different directions. Table 1 describes the direction selected of the HRTF database. In Table 1, the ϕ represents the elevation, which ranges from -40° to 80° . The $\Delta\theta$ represents the interval of azimuth. In the elevation of -40° and 40° , the $\Delta\theta$ of $6^\circ/7^\circ$ means the azimuth selected with an interval of 6° and 7° in turn, while in other elevation, the $\Delta\theta$ is a value, which means the interval is constant. The N represents the number of HRTF direction in different elevation.

Different from the CIPIC database [4], which uses distance and angle to quantify anthropometric parameters, in the Chinese pilots database, we quantify the subject's anthropometric parameters by using a series of three-dimensional coordinates of anthropometric characteristic points (such as the apex of the head and the tragus point).

For the three-dimensional anthropometric parameters, we selected 93 feature points, which consist of 50 feature points of the ears (25 for each ear), 33 feature points of the face, and 10 feature points of the torso. Figure 1 shows some of the anthropometric characteristic points.

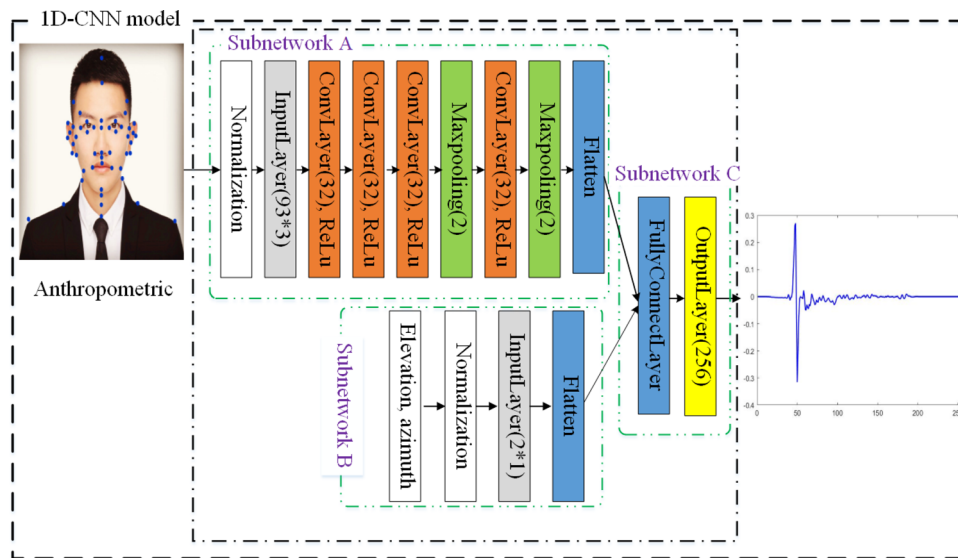


Fig. 1 Block diagram of a deep neural network for the HRIR estimation method based on three-dimensional anthropometric data and the position of the sound source (partial physiological parameters are marked on the figure Anthropometric)

Also, there were 69 anthropometric parameters in the Chinese pilots database extracted from the 93 anthropometric parameter points which included 37 features for head and torso in addition to 32 features for ears.

3 Proposed HRTF Personalization Model

3.1 Neural Network Architecture

This section proposes a method for estimating personalized HRTFs. The detailed structure of the proposed model employed in this experiment is summarized in Fig. 1. In this model, the anthropometric parameters and the elevation and azimuth of the sound source are used as input features, and the HRIRs are the target features. As shown in Fig. 1, the proposed neural network is composed of three subnetworks. The first subnetwork, referred to as “subnetwork A”, is a CNN that takes anthropometric measurements as input features to represent the relationship between the anthropometric measurements and the HRTF. The second subnetwork, referred to as “subnetwork B”, is a feature representation network based on the position of the sound source. The outputs of these two subnetworks are then combined to form the input to another DNN, referred to as “subnetwork C”, to estimate personalized HRTFs. Thus, the neural network is composed of three parts.

Subnetwork A, the first subnetwork, takes the anthropometric data as input. The first stage normalizes the anthropometric parameters, followed by three convolutional layers and then a max-pooling layer. The next stage comprises a convolutional layer and a max-pooling layer, which is fol-

lowed by a flattening layer. In the CNN realization employed here, 32 filters were used. Each filter had a kernel size of 1 cell, a pool size of 2 cells, and a stride of 1. The rectified linear unit (ReLU) activation function was applied for each layer because this function is known to be effective for solving the gradient vanishing problem [18]. Moreover, since the ranges of anthropometric measurements differ between measurements, measurements with a smaller range may have a reduced influence on the learning process. Thus, each input feature for the CNN was normalized with respect to the mean and variance of all training data regardless of the subject from whom the data were obtained, using the sigmoid function:

$$\bar{z}_i = \frac{1}{1 + e^{-z_i}} \tag{1}$$

where z_i and \bar{z}_i are the i th components of the input and normalized feature vectors, respectively. Note that z_i can represent any measured anthropometric parameters.

Subnetwork B takes the azimuth and elevation of the sound source as its input; these parameters are also normalized using the sigmoid function. The normalization layer is followed by a flattening layer. In the proposed network, the azimuth and elevation are mainly used as markers, so subnetwork B does not perform substantial processing.

The outputs of subnetwork A and subnetwork B are combined to form the input to subnetwork C. This subnetwork consists of hidden “dense” feedforward layers and uses the sigmoid function as the activation function.

The proposed network constructs left and right HRTF, respectively.

3.2 Supervised Learning

Because well-initialized weights can enable a low initial cost and fast convergence, during training, we used the Xavier initialization technique to initialize the weights of all layers, and all biases were initialized to zero [19]. Meanwhile, the mean square error (MSE) between the original target and the estimated target was selected as the cost function for the network [20]. We utilized adaptive moment estimation (“Adam”) optimization for the backpropagation algorithm and set the moment decay rate to 0.9, with a learning rate of 0.01 [21]. Finally, we trained the model for 62 epochs, with 398 samples in each epoch. To prevent overfitting, the dropout technique was applied with a probability of 0.7 [22].

4 Performance and Evaluation

In this paper, the performance of the proposed personalized HRTF method is evaluated on the basis of objective and subjective experiments. For the objective experiment, the root mean square error (RMSE) between the reference and estimated HRTFs was measured. Moreover, the performance of the proposed method was compared with that of other HRTF personalization methods: (1) the DNN-based method [15] and (2) a database matching method [6]. The CNN-DNN method [14] requires the ear anthropometric parameters to be presented as images; however, the Chinese pilots database does not provide ear images, making this method unsuitable for application to this database, whereas the DNN and database matching methods are both suitable for this purpose. For the DNN method, the two-dimensional anthropometric parameters in the Chinese pilots database were selected, and the model presented in [15] was selected for training.

The database matching method was implemented in MATLAB version of R2015b, while the DNN method and the proposed method were implemented in Anaconda Python 3.7.0.

The total number of HRTF samples in the Chinese pilots database is $63 * 723 = 45,549$. The data from three subjects were selected as the verification data. Among the remaining 60 subjects, the data from $60 * 70\% = 42$ subjects were selected as the training data, and the data from the remaining $60 * 30\% = 18$ subjects were selected as the test data. Therefore, it was necessary to construct 21 models to verify the HRTFs for all subjects. For the DNN method, since each model is established for a single direction, obtaining the HRTFs for all sound-source positions required the construction of 723 DNN models. Thus, in terms of the number of models to be trained, the proposed method requires fewer models.

4.1 Objective Experiment

The spectral distortion (SD) error metric was employed to evaluate the precision of the synthesis of personalized HRTFs. This metric is defined as follows:

$$SD^{(d)}(H, \hat{H}) = \sqrt{\frac{1}{N} \sum_{n=1}^N \left(20 \log_{10} \frac{\|H^{(d)}(n)\|}{\|\hat{H}^{(d)}(n)\|} \right)^2} \quad (2)$$

where $H^{(d)}$ is the original HRTF in direction d , $\hat{H}^{(d)}$ is the matched HRTF at direction d , $SD^{(d)}(H, \hat{H})$ is the corresponding SD, and n is the index of the frequency bin, with $N = 129$ being the total number of frequency bins.

Then, the root mean square error (RMSE) was used to calculate the mean SD for multiple directions:

$$SD(H, \hat{H}) = \sqrt{\frac{1}{D} \sum_{d=1}^D (SD^d(H, \hat{H}))^2} \quad \text{dB} \quad (3)$$

where, D is the number of directions.

Figure 2 compares the SD values of the personalized HRTFs by subject as obtained using different HRTF estimation methods, measured at $(0^\circ, 0^\circ)$, $(0^\circ, 45^\circ)$, and $(0^\circ, 165^\circ)$. Moreover, Table 2 compares the average over all HRTFs in the Chinese pilots database for the same direction.

As shown in Fig. 2 and Table 2, the DNN-based method [DNN (69)] and the proposed method achieved RMSEs that were 2.77 dB and 3.81 dB lower, respectively, than that of the database matching method. Because the performance of database matching depends on the size of the database, the SD is larger. When the database capacity increases, the performance of database matching can also be enhanced to a certain extent. Compared with that of the DNN method [15], the SD of the proposed method is reduced by 1.04 dB. The primary reason is that the proposed method uses a 1D-CNN to process the anthropometric parameters, thereby building internal relationships between the anthropometric parameters.

Although the proposed method reduces the SD by only 1.04 dB compared to the DNN method, the proposed method has a smaller model size and requires less reconstruction time than either the DNN method [14] or the CNN + DNN method [15].

The proposed method can obtain the HRTFs for all positions in the database using only one model, whereas both the DNN method [14] and the CNN+DNN method [15] require a separate model to be built for each position. (For the Chinese pilots database, 723 models need to be built.) Thus, the proposed method greatly reduces the number of models needed and reduces the storage space required for model packaging.

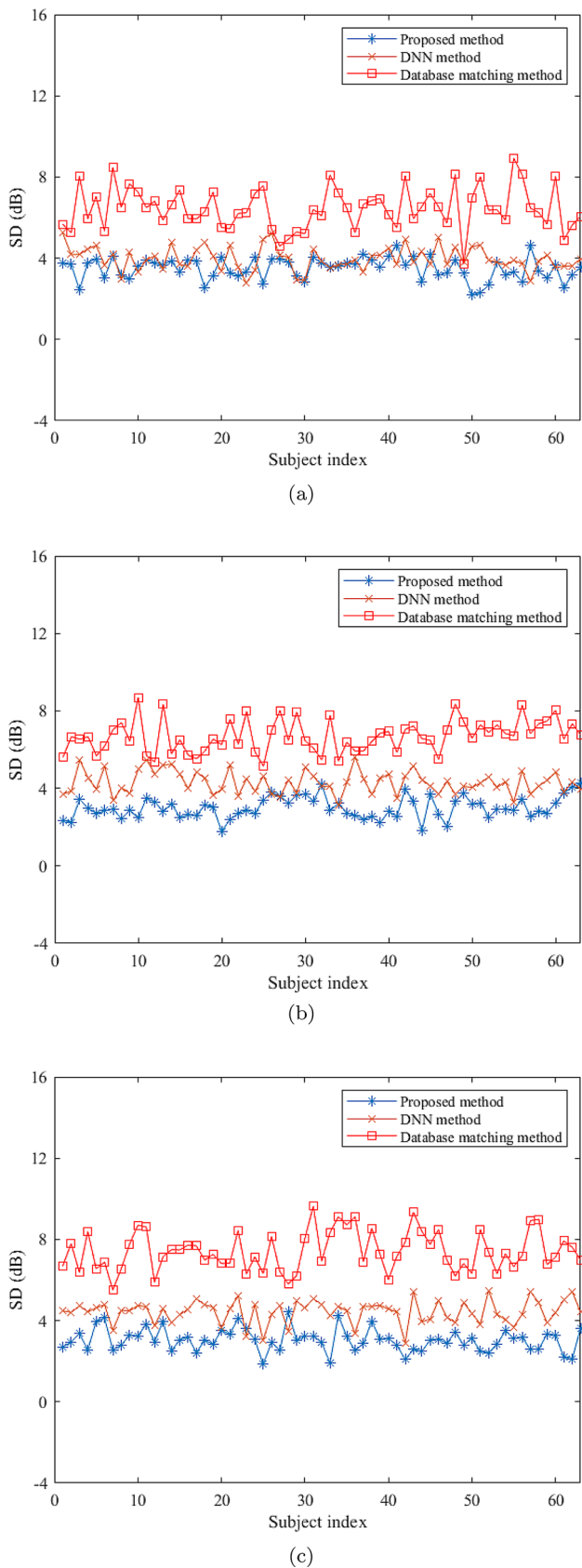


Fig. 2 The SDs for individual subjects achieved with different personalization methods for different directions **a** (0°, 0°); **b** (0°, 45°) and **c** (0°, 165°)

Table 2 RMSEs for different personalization methods at (0°, 0°), (0°, 45°), and (0°, 165°)

Direction	Database matching	DNN(69)	Proposed method
(0°, 0°)	6.75	4.05	3.63
(0°, 45°)	6.92	4.45	3.11
(0°, 165°)	7.54	4.41	3.05
Average	7.08	4.31	3.27

In terms of model application, the proposed model needs only the physiological parameters and sound source direction as input to reconstruct the given subject’s HRTF in the given direction, whereas both the DNN method [14] and the CNN+DNN method [15] require first finding the corresponding model and then reconstructing the HRTF, which increases the time required for reconstruction.

4.2 Subjective Experiment

In this section, we present a subjective evaluation based on a localization experiment involving five male participants without any auditory disease. In the localization experiment, three types of HRTFs were tested: the personalized HRTF obtained using the proposed method, the personalized HRTF obtained using the DNN method and the HRTFs of the KEMAR dummy head. A total of 8 directions were selected on the horizontal plane: 0°, 45°, 80°, 135°, 180°, 225°, 280°, and 350°. The test stimuli were produced via the convolution of white noise with the HRTF in the determined direction.

During the experiment, the test stimuli were played back to the participants through headphones. Before the test, experimental questionnaires were handed out to all the participants. All test azimuths were listed for each test stimulus in the questionnaire. The participants were asked to select the most accurate azimuth according to their own judgment after hearing each stimulus. During the test, each stimulus was presented three times in a row. Then, there was a 5-s time interval during which the participants were asked to enter their judgments about the presented stimulus in the questionnaire. Then, the next stimulus was presented in the same way until all test stimuli had been presented using all three tested HRTF types. The presentation order of the stimuli was randomized using an altered Latin square scheme. Thus, the possible bias caused by the order effect and sequential dependencies could be minimized.

Figures 3, 4, 5 illustrate the sound localization results for all five subjects, showing the azimuths as judged by the participants versus the target azimuth for the KEMAR dummy head HRTFs and the personalized HRTFs. In the figure, the main diagonal solid line in each panel represents correct judgment; a point lying on this line indicates that the subjects’ judgment was exactly the same as the target azimuth. The

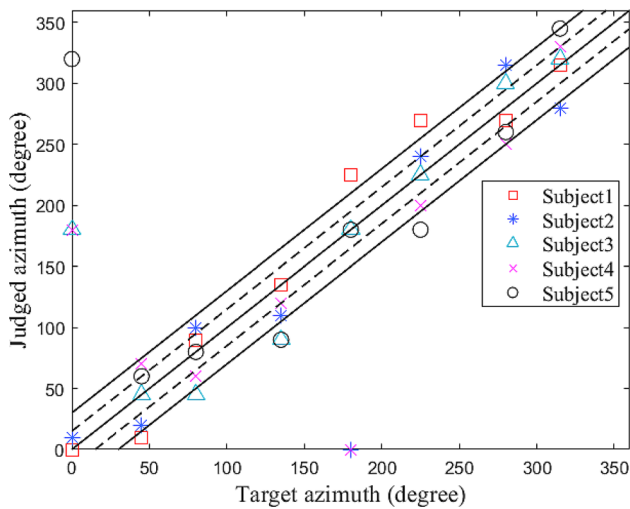


Fig. 3 Results of the localization test with the HRTF of the KEMAR dummy head

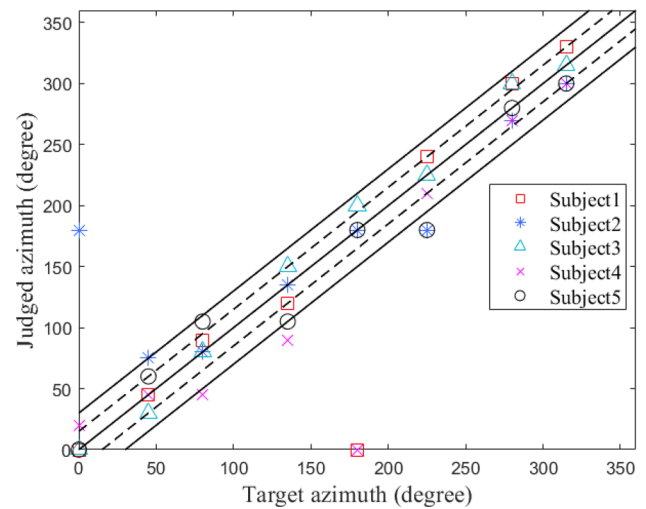


Fig. 5 Results of the localization test with the HRTF generated by the proposed method

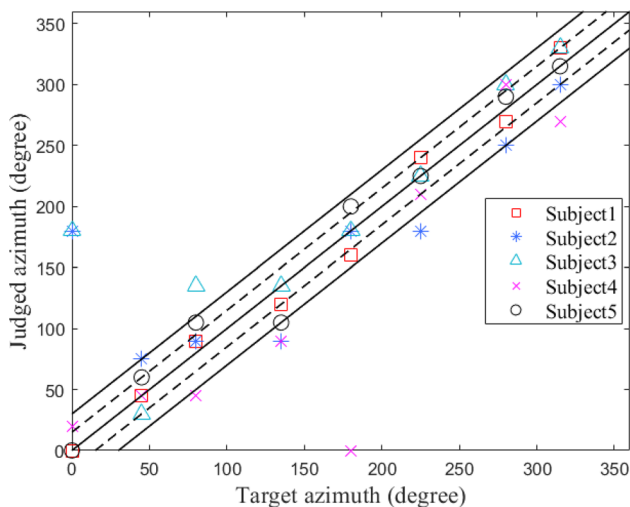


Fig. 4 Results of the localization test with the DNN-based HRTF

upper and lower off-diagonal solid lines represent 30° margins of error, and the upper and lower off-diagonal dashed lines represent 15° margin of error. It can be observed that the personalized HRTFs show slightly better localization performance.

Table 3 compares the accuracy within specific margins, such as $\pm 15^\circ$ and $\pm 30^\circ$, achieved in the localization experiment with the KEMAR dummy head HRTF, the DNN-based HRTF, and the proposed method. As shown in this table, compared with the KEMAR dummy head HRTF, the proposed HRTF estimation method achieved higher average accuracies by 25% and 17.5% for the $\pm 15^\circ$ and $\pm 30^\circ$ margins, respectively. Similarly, the accuracy of the proposed method was improved by 10% and 5% compared to the DNN-based HRTF method.

4.3 Data Augmentation

The number of subjects in the Chinese pilots database is 63, while the number of subjects in the CIPIC database is 31. Even though the database capacity of the Chinese pilots database is twice that of the CIPIC database, in the deep learning contest, this amount of data is still not large. When the training database is small, this can easily result in over-fitting of the model [23]. The processing of a CNN can be invariant with respect to translation, viewpoint, size, illumination, or a combination of the above [24]. Subnetwork A in the proposed model is a CNN; therefore, the data used as input to this subnetwork can be subjected to data augmentation processing. Accordingly, the data were enhanced by applying two types of operations: shifting and rotating.

Table 4 compares the RMSEs for individual subjects achieved using the proposed HRTF estimation method with and without data augmentation, as measured at $(0^\circ, 0^\circ)$, $(0^\circ, 45^\circ)$, and $(0^\circ, 165^\circ)$. The average RMSE decreased with data augmentation. Shifting augmentation reduced the RMSE of the proposed method by 0.2 dB, while rotation augmentation reduced the average RMSE of the proposed method by 0.16 dB.

4.4 Suitability for Application to the SYMARE Database

The proposed method was also applied to the SYMARE database [25], which differs from the Chinese pilots database in terms of the anthropometric parameters and the direction of the sound source.

The SYMARE database contains morphological data from 10 subjects obtained through magnetic resonance imaging (MRI), which is different from the format of the

Table 3 Comparison of average accuracies within specific margins (15° and 30°) for different HRTF estimation methods

	KEMAR dummy head HRTF (%)	DNN-based HRTF (%)	Proposed HRTF (%)
Accuracy within 15°	40	55	65
Accuracy within 30°	65	77.5	82.5

Table 4 Comparison of the average RMSEs for the proposed HRTF estimation method with and without data augmentation at $(0^\circ, 0^\circ)$, $(0^\circ, 45^\circ)$, and $(0^\circ, 165^\circ)$

Direction	Proposed method	Proposed method + shift	Proposed method + rotation
$(0^\circ, 0^\circ)$	3.63	3.21	3.24
$(0^\circ, 45^\circ)$	3.11	2.98	3.00
$(0^\circ, 165^\circ)$	3.05	3.04	3.08
Average	3.27	3.07	3.11

anthropometric parameters in the Chinese pilots database. To obtain physiological parameters equivalent to those in the Chinese pilots database, we imported the .ply file under the Meshes/HighResolution/HeadTorsoandEars directory in the SYMARE database into MATLAB and manually obtained the data.

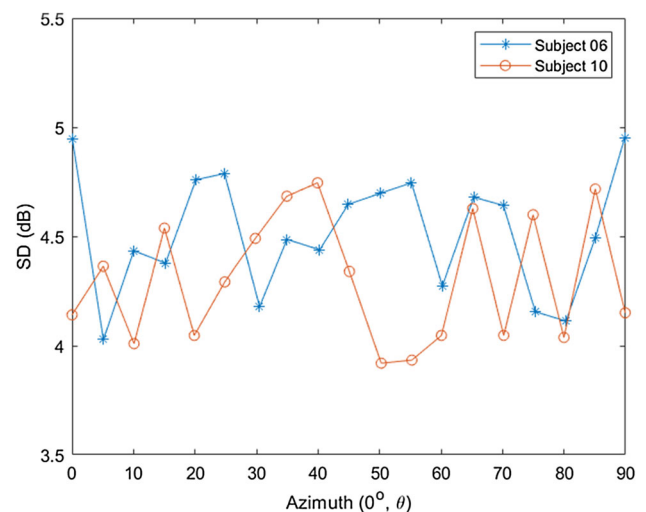
We obtained the three-dimensional anthropometric parameters of subjects 06 and 10 from the morphological data files HTE06.ply and HTE10.ply and input these three-dimensional anthropometric parameters and the direction of the sound source into the trained model to reconstruct the corresponding HRTF data. For the test, we chose a sound-source elevation of 0° and azimuth angles ranging from 0° to 90° in intervals of 5° .

Figure 6 shows the SDs for subject 06 and subject 10 in the direction $(0^\circ, \theta)$. The average SD in Fig. 6 is 4.51 dB, which is greater than the average SD reported for the proposed method in Table 4.

The reasons for the poor performance of the proposed model on the SYMARE database are as follows: (1) Additional deviations were introduced when manually extracting the physiological characteristic parameters, and (2) the environmental differences and differences in the measurement methods between the measurements performed to construct the SYMARE database and the Chinese pilots database result in certain differences between the characteristics of the two databases. However, it should be noted that the average value in Fig. 6 is still less than the maximum spectral distortion of 5.7 dB of the method proposed by Nishino [26], which can achieve rough localization.

5 Conclusion

In this paper, a personalized HRTF estimation method is proposed on the basis of a deep neural network taking anthropometric measurements and the direction of the sound source

**Fig. 6** The SDs obtained for subject 06 and subject 10 in the SYMARE database in the direction $(0^\circ, \theta)$ for the left ear using the trained model

as input. In particular, while the conventional DNN-based method attempts to estimate HRTFs from anthropometric data such as head, torso, and pinna measurements, in the proposed method, these anthropometric parameters are replaced with point parameters. These anthropometric parameter points are processed by a 1D-CNN. The neural network used in the proposed method is composed of three subnetworks. The first one is a 1D-CNN, which processes the anthropometric parameters, and the second one normalizes the location of the sound source. The outputs of these two subnetworks are then merged to form the input to another DNN to estimate a personalized HRTF.

The performance of the proposed personalized HRTF estimation method was evaluated through objective and subjective tests. In the objective test, the RMSEs between the measured and estimated HRTFs were evaluated. For the subjective test, a sound localization experiment was performed. The performance of the proposed method was compared

with that of a database matching method and the DNN (69) method. The objective evaluation showed that the RMSE of the proposed method was decreased by 3.80 dB and 1.04 dB, respectively, compared to those of the two methods considered for comparison. In addition, the subjective evaluation showed that the proposed method provided a 5% higher localization accuracy than the DNN-based method.

Moreover, shift or rotation is used to enhance the data in the Chinese pilots database, and the RMSE of the enhanced data reduced by 0.2 dB and 0.16 dB, respectively. Furthermore, the SYMARE database was used to verify the trained model, and the performance showed that the trained model is suitable for application to the SYMARE database.

Although, compared with the other methods, the proposed model has the lowest average SD, this method still has some limitations. The proposed network uses a 1D-CNN network to process anthropometric parameters and thus requires the anthropometric parameters to be quantified in the form of one-dimensional data; consequently, it is not applicable to databases such as the CIPIC database and the MIT database. In addition, although the number of subjects in the database used in this article is twice that in the CIPIC database, in the context of deep learning, this database size is still limited.

Acknowledgements Thanks for the support of the National Natural Science Foundation of China (11774291) and the Natural Science Foundation of Shaanxi Province of China (2018JM6020).

References

- Rumsey, F.: Spatial Audio. Focal Press, Woburn, MA, USA (2001)
- Blauert, J.: Spatial Hearing, Revised edn. MIT, Cambridge, MA (1997)
- Wenzel, E.M., Arruda, M., Kistler, D.J., et al.: Localization using non-individualized head-related transfer functions. *J. Acoust. Soc. Am.* **94**, 111–123 (1994). <https://doi.org/10.1121/1.407089>
- Algazi, V.R., Duda, R.O., Thompson, D., Avendano, C.: In the CIPIC HRTF database, workshop on applications of signal processing to audio and acoustics, pp. 99–102 (2001)
- Zeng, X.Y., Wang, S.G., Gao, L.P.: A hybrid algorithm for selecting head-related transfer function based on similarity of anthropometric structures. *J. Sound Vib.* **329**, 4093–4106 (2010)
- Lu, D.D., Zeng, X.Y., Guo, X.C., et al.: Personalization of Head-Related Transfer Function Based on Sparse Principle Component Analysis and Sparse Representation of 3D Anthropometric Parameters. *Acoust. Aust.* (2019). <https://doi.org/10.1007/s40857-019-00169-y>
- Torres-Gallegos, E.A., Orduña-Bustamante, F., Arámbula-Cosío, F.: Personalization of head related transfer function(HRTF) based on automatic photo-anthropometry and inference from a database. *Appl. Acoust.* **97**, 84–95 (2015)
- Katz, B.F.G.: Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *Acoust. Soc. Am.* **110**, 2440–2448 (2001)
- Spagnol, S., Geronazzo, M., Avanzini, F.: On the relation between pinna reflection patterns and Head-Related Transfer Function Features. *IEEE Trans. Audio, Speech Lang. Process.* **21**, 508–519 (2013). <https://doi.org/10.1.1.706.9105>
- Spagnol, S., Avanzini, F.: Frequency estimation of the first pinna notch in Head-Related Transfer Functions with a linear anthropometric model. In: *Proceeding 18th International Conference Digital Audio Effects (DAFx-2015)*, Trondheim, Norway, pp. 231–236 (December, 2015)
- Shahnawaz, M., Bianchi, L., Sarti, A., Tubaro, S.: In Analyzing notch patterns of head related transfer functions in CIPIC and SYMARE databases, In: *European Signal Processing Conference*, pp. 101–105 (2016)
- Bilinski, P., Ahrens, J., Thomas, M.R.P., Tashev, I., Platt, J.: In HRTF magnitude synthesis via sparse representation of anthropometric features, In: *International Conference on Acoustics Speech and Signal Processing*, pp. 4468–4472 (2014)
- Hu, H., Zhou, L., Ma, H., Wu, Z.: HRTF personalization based on artificial neural network in individual virtual auditory space. *Appl. Acoust.* **69**(2), 163–172 (2008)
- Chun, C., Moon, J., Lee, J., et al.: Deep neural network based HRTF personalization using anthropometric measurements. In: *Audio Engineering Society Convention 143*, Audio Engineering Society (2017)
- Lee, G.W., Kim, H.K.: Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear. *Appl. Sci.* **8**(11), 2180 (2018). <https://doi.org/10.3390/app8112180>
- Wu, Z., Song, S., Khosla, A., et al.: 3d shape-nets: A deep representation for volumetric shapes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1912–1920 (2015)
- In: Guo, X., Xiong, D., Wang, Y., et al.: Head-Related Transfer Function Database of Chinese Male Pilots. *Proceedings of the 16th International Conference on MMESE*, Xi'an, China, pp. 3–11 (October 2016)
- Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: *Proceeding of the 14th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, USA, pp. 315–323 (April 2011)
- Glorot, X., Bengio, Y.: In Understanding the difficulty of training deep feed-forward neural networks, In: *International Conference on Artificial Intelligence and Statistics*, pp. 249–256 (2010)
- Rumelhart, D., Hinton, G., Williams, R.: Learning representations by back-propagating errors. *Nature* **323**(6088), 533–536 (1986)
- Kingma, D., Ba, J.: ADAM: A method for stochastic optimization. In: *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, pp. 1–15 (May 2015)
- Srivastava, N., Hinton, G., Krizhevsky, A., et al.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
- Frank, J., Harrell, E.: *Regression Modeling Strategies*; Springer: Berlin/Heidelberg. Germany (2006). <https://doi.org/10.1007/978-1-4757-3462-1>
- Simard, P.Y., Steinkraus, D.W., Platt, J.: In Best practices for convolutional neural networks applied to visual document analysis, In: *International Conference on Document Analysis and Recognition*, pp. 958–963 (2003). <https://doi.org/10.1109/ICDAR.2003.1227801>
- Jin, C.T., Guillon, P., Epain, N., et al.: Creating the Sydney York morphological and acoustic recordings of ears database. *IEEE Trans. Multimed.* **16**(1), 37–46 (2014)
- Nishino, T., Inoue, N., Takeda, K., et al.: Estimation of HRTFs on the horizontal plane using physical features. *Appl. Acoust.* **68**, 897–908 (2007)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.