2023 15th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) | 979-8-3503-2617-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/IHMSC58761.2023.00041

An individualized HRTF model based on random forest and anthropometric parameters

Yuqi Teng, Xiaoli Zhong*
School of Physics and Optoelectronics
South China University of Technology
Guangzhou, China
*Corresponding author, 13710167606@139.com

***Abstract*—Individualized head-related transfer functions (HRTFs) can improve the auditory performance of the listener in the virtual acoustic environment. This study introduced random forest to fit the complex relationship between anthropometric parameters and HRTFs, and further constructed an individualized HRTF model. In terms of the goodness of fit, the determination coefficients of the left and right ears are 90.1% and 91.1%, respectively. Moreover, the fitting performance of the model is better when the sound source location is ipsilateral to the ear as well as at high elevations. The prediction performance of the model was evaluated at typical source locations. Results show that the spectral distortions of the predicted HRTFs increase with frequency, reaching a mean value of 4.74dB.**

Keywords-Individualized HRTF,anthropometric parameters, random forest

I. Introduction

Head-related transfer functions (HRTFs) describe the propagation characteristics of sound wave from the sound source location to the listener, and are the key to virtual sound and its applications. HRTFs vary with individuals, and the use of individualized HRTFs has been highly recommended to improve listener's auditory immersion [1−3].

From views of forming, HRTFs are closely related to anthropometric parameters. Therefore, the method of predicting individualized HRTF using anthropometric parameters has been widely studied. Most of existing studies used linear models between HRTFs and anthropometric parameters [4−8]. Liu et al. obtained the linear regression model of HRTFs with four anthropometric parameters by using multiple linear regression and independent component analysis (ICA) [4]. Rodriguez et al. used principal component analysis (PCA) and linear regression methods to construct a prediction model of pinna-related transfer functions (PRTFs) which is the main component of HRTFs [5]. Nishion et al. used a similar approach, but treated magnitude and phase separately [6−7]. Iida et al. extracted the early HRIRs (the time-domain equivalent of HRTFs) and used linear regression to fit the relationship between anthropometric parameters and early HRIRs [8]. The above methods are effective to a certain extent. However, the relationship between HRTFs and the anthropometric structure is non-linear in essence, thus the existing linear model may not fully describe the complex relationship between HRTFs and the anthropometric structure. Recently, machine learning has been gradually introduced in the field of acoustics [9]. As a supervised learning algorithm, random forest has high accuracy and good robustness. Valente constructed a random forest model, which can predict sound pressure level several kilometers away by measuring simple meteorological data [10]. Acoustic feature extraction and random forest model were used to detect gas pipeline leaks [11]. To our knowledges, random forest hasn't been introduced in HRTF individualization till now.

This study proposed an individualized HRTF model based on random forest and anthropometric parameters. The second section describes the principles of random forest and model construction. The third section evaluates model performance. The final section summarizes the main conclusions.

II. Individualized HRTF model based on Random Forest

A. Principle

Fig. 1 is the framework of this study. Firstly, the model of individualized HRTFs was constructed by using random forest (RF) to explore the non-linear relationship between the measured HRTFs and corresponding anthropometric parameters in the train set. Then, the anthropometric parameters of the subjects in the test set were substituted into the individualized model to obtain the predicted HRTFs. Finally, the predicted HRTFs were compared with the corresponding measured HRTFs to validate the proposed model of individualized HRTFs.

The algorithm of RF combines decision trees and bagging learning strategies. Firstly, multiple subsets are randomly sampled from the train set using the bootstrap method. Each decision tree is trained independently with different subsets to reduce the influence of abnormal data on the model.

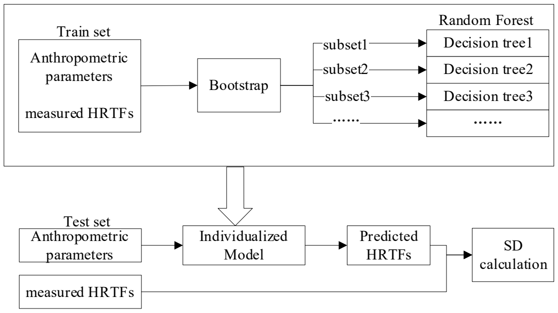

Figure 1. The framework of this study

2157-8982/23/$31.00 ©2023 IEEE
DOI 10.1109/IHMSC58761.2023.00041

Then the prediction variables are randomly sampled, and each decision tree uses different variables for prediction, which reduces the influence of pseudo variables. For each tree, the splitting process follows classification and regression tree (CART) algorithm. Specifically, the difference in mean square error (*ΔMSE*) between two adjacent nodes is selected as the splitting criterion, which is defined as:

$$\Delta MSE = MSE_j - MSE_{j-1}$$
$$MSE_j = \frac{\sum_k (H_{j,k} - \bar{H}_j)}{N_j} \qquad (1)$$

In (1), $H_{j,k}$ represents the *k*th frequency bin of logarithmic magnitude of HRTFs at the *j*th node. $\bar{H}_j$ represents the mean logarithmic magnitude of HRTFs for all frequency bins at the *j*th node. N_j represents the total number of samples at the *j*th node. The calculation of (1) is carried out across all values of all prediction variables, during which the prediction variable and value corresponding to the maximum *ΔMSE* are selected to divide the samples into two child nodes. Once the *MSE* reaches the tolerance value, the splitting stops automatically and the corresponding node becomes a leaf node. The prediction result of one decision tree is the mean logarithmic magnitude of HRTFs for all samples at a certain leaf node, and the prediction result of the RF is the mean value of all decision trees.

B. Database

This study used the HUTUBS HRTF database with 96 subjects [12]. For each subject, there were HRTF data at 440 source locations with the sampling point of 512 and the sampling rate of 44100Hz. The database adopted the counterclockwise spherical coordinate system. Location is represented as (azimuth, elevation). The azimuth ranged from 0° to 360° with an interval of 10° great-circle distance and the elevation ranged from −90° to 90° with an interval of 10°. The database also contained 25 anthropometric parameters of torso, head and pinna for most of subjects. Removing HRTF sets of dummy head as well as repeated measurements, this study used the HRTFs and anthropometric parameters of 90 human subjects.

C. Data Preprocessing

The measured HRIR of a sound source location was transformed into a log-magnitude spectrum by Fourier transform, which was used as the prediction label. An artificial variable *F* was introduced to distinguish the HRTF magnitude at different frequency bins. Considering the low-frequency non-flat characteristics of the loudspeaker and the high-frequency measurement error, the frequency band of HRTFs in this study was limited to 1−12kHz. The 14 out of 25 anthropometric parameters in the database, a_1-a_{14} were selected as the predicted variables, see Table I. Moreover, five area-related parameters (a_{15}–a_{19}) were also proposed as shown in Table I.

TABLE I. ANTHROPOMETRIC PARAMETERS USED IN THE MODEL CONSTRUCTION OF HRTFS

No.	Definition	No.	Definition
a_1	Head width	a_{11}	Pinna width
a_2	Head height	a_{12}	Intertragal incisure width
a_3	Head depth	a_{13}	Pinna rotation angle
a_4	Pinna offset down	a_{14}	Pinna flare angle
a_5	Pinna offset back	a_{15}	Area of cavum concha
a_6	Cavum concha height	a_{16}	Area of cymba concha
a_7	Cymba concha height	a_{17}	Area of fossa
a_8	Cavum concha width	a_{18}	Area of pinna
a_9	Fossa height	a_{19}	Area of intertragal incisure
a_{10}	Pinna height		

The 10 out of the 90 subjects were selected and combined as the test set, while the rest as the training set. As the RF model used the out of bag (OOB) data that had not been sampled to calculate the OOB error, there is no need to construct the verification set. A total of 5120 samples (80 subjects * 64 frequency bins) were generated from 1 to 12 kHz. Each sample contained 21 variables (19 anthropometric parameters, frequency, and HRTF log-magnitude).

D. Hyperparameter Tuning

RF was implemented by MATLAB R2022a. The hyperparameters of RF mainly refer to the number of trees and the number of selected variables. The larger the number of trees is, the smaller the variance of model is. However, the complexity of RF increases rapidly when the number of trees is large. On the other hand, the higher the number of selected variables, the more information the model can synthesize for prediction. However, too many variables may increase the interference of irrelevant variables. Fig. 2 shows the relationship between hyperparameters and determination coefficients (R^2) of the model. The R^2 is used to evaluate the fitting degree of the RF model, and defined as the square of the correlation coefficient between measured HRTF H_m and predicted HRTF H_p obtained from OOB data:

$$R^2 = \frac{\sum_n [(H_p(n) - \bar{H}_p(n)) \times (H_m(n) - \bar{H}_m(n))]^2}{\sum_n (H_p(n) - \bar{H}_p(n))^2 \times \sum_n (H_m(n) - \bar{H}_m(n))^2} \qquad (2)$$

where *n* represents the frequency bin. When the number of selected variables (shorten as Var) is 18 or 20, there is no significant difference in the fitting performance of the model. When the number of trees is more than 500, there is no significant gain from increasing the number of trees. Considering the accuracy and complexity of the model, 18 variables and 500 trees were selected.

E. Fit Goodness of the Individualized HRTF Model

The above individualized HRTF model can be applied to any sound source location. To evaluate how the fitting performance vary with sound source location, we calculated

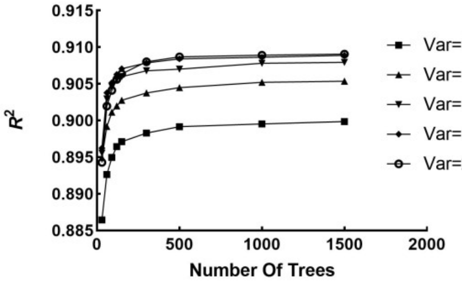

Figure 2. The hyperparameter tuning

TABLE II. THE R^2 OF FIVE TYPICAL SOUND LOCATIONS

R^2	(0°,0°)	(40°,0°)	(320°,0°)	(0°,30°)	(0°, −30°)	Mean
Left ear	90.8%	92.5%	87.2%	94.2%	85.7%	90.1%
Right ear	91.3%	88.6%	92.9%	94.8%	88.2%	91.1%

determination coefficients R^2 of five typical locations, as shown in Table II. Generally, the binaural HRTFs have been well fitted by the proposed individualized HRTF model, and the mean R^2 about five locations reaches 90.1% and 91.1% for the left and right ear, respectively. It is suggested that the individualized HRTF model proposed in this study based on anthropometric parameters and RF can reflect the essential correlation between anthropometric parameters and HRTFs.

In Table II, when the sound source location is (0°,0°), the R^2 of the left and right ears are 90.8% and 91.3%, suggesting a left-right symmetry corresponding to the left-right symmetry of sound transmission path. When the sound source is at (40°,0°), R^2 equals 92.5% for the left ear ipsilateral to the sound source, while only 88.6% for the right ear contralateral to the sound source. Similarly, when the sound source is at (320°,0°), R^2 equals 92.9% for the right ear ipsilateral to the sound source, while is only 87.2% for the left ear contralateral to the sound source. This phenomenon is due to the fact that, the HRTFs contralateral to the sound source are rather complicated because the sound needs to transmit through a complex diffraction path to reach the contralateral ear, so the fitting performance of the contralateral HRTFs is worse than that of the ipsilateral HRTFs. On the other hand, when the sound source is located at the high elevation (0°,30°), the R^2 of left and right ears is 94.2% and 94.8%, which is significantly better than those of the low elevation (0°, −30°) where R^2 is only 85.7% and 88.2%. This phenomenon may be due to the fact that, the reflections from torso and knee at low elevation result in large measurement errors in HRTFs, and the fitting performance of the low-elevation HRTFs is worse than that of the high-elevation HRTFs.

III. PERFORMANCE OF THE INDIVIDUALIZED HRTF MODEL

The 19 anthropometric parameters of the 10 test set subjects were substituted into the model to obtain the predicted HRTF for the sound source located in the ipsilateral ear at above five locations. Spectral distortion *SD* was used as an objective evaluation index to describe the difference between predicted and measured HRTFs, and was defined as:

$$SD(f,\Omega,s) = 20 \times \lg \frac{\left|H_p(f,\Omega,s)\right|}{\left|H_m(f,\Omega,s)\right|} \tag{3}$$

where f represents frequency, Ω represents sound source location, and s represents a certain subject. H_p and H_m represent predicted HRTF and measured HRTF, respectively. The smaller *SD* is, the more similar they are.

Fig. 3 shows the *SD* varying with frequency for a typical subject. The mean *SD* is 0.94dB below 5.7kHz, and a maximum error of 3dB is observed around 2.7kHz. However, when the frequency increases to high frequency above 5.7 kHz, the mean *SD* is 3.8dB, and a maximum error of 11.4dB is observed around 10.2kHz. Moreover, analyses of all subjects show that *SD* begins to rise significantly around 6 kHz. Since at high frequencies, there exist obvious individualized characteristics as well as measurement errors, it is difficult to predict high-frequency HRTFs. Fortunately, the auditory system of humans is insensitive at high frequency, thus the relatively large high-frequency prediction errors may not lead to audible perception.

Table III shows the mean *SD* of predicted and measured HRTFs about frequency for each subject in the test set. The mean *SD* about all locations ranges from 3.98dB to 5.71dB. The best prediction is obtained by subject 9, with the mean *SD* of locations reaching 3.98dB. The worst prediction is obtained by subject 7, with the mean *SD* of locations reaching 5.71dB. The *SD* of subject 7 is as large as 7.77dB at (0°,0°), while as small as 3.60dB at (320°,0°). This suggests that the prediction performance varies significantly with locations. From the perspective of sound location, (0°,30°) and (320°,0°) achieve the best performance with a mean *SD* of 4.48dB and 4.39dB; The front location (0°,0°) has a general performance with a mean *SD* of 4.71dB; (0°, −30°) and (40°,0°) perform the worst with mean *SD* of 5.1dB and 5.0dB. Overall, the mean *SD* about all subjects and all locations is 4.74dB. Therefore, the proposed individualized HRTF model can predict HRTFs close to the measured HRTFs.

To further explore the sources of prediction error of the proposed model, we calculated the relative deviations D_i of *i*th anthropometric parameters a_i between the best-predicted subject (subject 7) and the worst-predicted subject (subject 9) as follows:

$$D_i = \left| \frac{a_{best,i} - a_{worst,i}}{\overline{a_i}} \right| \tag{4}$$

where $\bar{a}_i$ represents the mean value of anthropometric parameter a_i about the dataset. Result shows that there are significant differences between the two subjects in pinna offset down (a_4), pinna flare angle (a_{14}), area of cymba concha (a_{16}), and area of intertragal incisure (a_{19}). The relative deviation of above four parameters reaches 1.03, 0.95, 0.55, and 0.58, respectively. This implies that HRTFs may be more sensitive to these parameters in the individualized HRTF model.

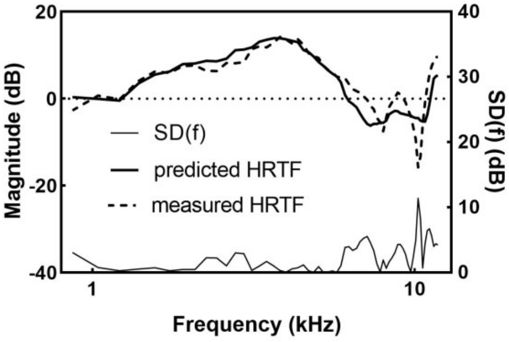

Figure 3. The *SD* of a typical subject

TABLE III. EVALUATION OF PREDICTION RESULTS IN THE TEST SET

SD	**(0°,0°)**	**(40°,0°)**	**(320°,0°)**	**(0°,30°)**	**(0°, −30°)**	**Mean**
Subject 1	4.40	4.31	4.62	2.77	4.52	4.12
Subject 2	5.29	3.63	6.17	4.61	5.35	5.01
Subject 3	3.36	3.36	6.29	5.82	4.17	4.60
Subject 4	3.78	3.47	4.23	4.79	5.08	4.27
Subject 5	4.78	5.43	4.14	7.31	6.29	5.59
Subject 6	3.88	5.89	5.28	6.68	5.22	5.39
Subject 7	7.77	7.02	4.41	5.73	3.60	5.71
Subject 8	5.20	4.33	5.65	4.31	1.76	4.25
Subject 9	4.25	3.90	4.71	4.24	2.77	3.98
Subject 10	4.35	3.50	5.47	3.78	5.13	4.45
Mean	4.71	4.48	5.10	5.00	4.39	4.74

IV. CONCLUSIONS

In this study, a complex relationship between anthropometric parameters and HRTFs was constructed based on anthropometric parameters and random forest, and then an individualized HRTF model was obtained. Individualized HRTFs can be predicted by substituting the anthropometric parameters of a new subject into the model. Results indicate that the proposed model performs well in the five typical locations. Generally, the mean *SD* between predicted and measured HRTFs is 4.74dB. The subject with the best prediction performance is 3.98dB and the subject with the worst prediction performance is 5.71dB. The relative differences in anthropometric parameters between the best- and worst-predicted subjects indicate that pinna offset down, pinna flare angle, area of cymba concha, and area of intertragal incisure are vital to HRTF individualization.

ACKNOWLEDGMENT

This work was supported by Natural Science Foundation of Guangdong Province (Grant No. 2021A1515011871)

REFERENCES

[1] J. Blauert, “Spatial hearing: the psychophysics of human sound localization,” MIT press, 1997.

[2] B. S. Xie, “Head-related transfer function and virtual auditory display,” M. Ross Publishing, 2013.

[3] X. L. Zhong, “Binaural auditory localization and its virtual display,” The press of South China University of Technology, 2023.

[4] X. J. Liu, H. Song and X. L. Zhong, “A hybrid algorithm for predicting median-plane head-related transfer functions from anthropometric measurements,” Applied Sciences., vol.9(11), 2019, doi: 10.3390 / app 9112323..

[5] M. A. Ramirez, and S. G. Rodriguez, “HRTF individualization by solving the least squares problem,” Audio Engineering Society Convention 118., May. 2005, Preprint:6438.

[6] T. Nishion, Y. Nakai, K. Takeda, and F. Itakura, “Estimating head related transfer function using multiple regression analysis,” The Journal of the Institute of Electronics, Information and Communication Engineers., vol.J84-A(3), pp. 260−268, 2001.

[7] T. Nishino, N. Inoue, K. Takeda, and F. Itakura, “Estimation of HRTFs on the horizontal plane using physical features,” Applied Acoustics., vol.68(8), pp. 897−908, 2007.

[8] K. Iida, H. Shimazaki, and M. Oota, “Generation of the individual head-related transfer functions in the upper median plane based on the anthropometry of the listener’s pinnae,” 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), IEEE press, October. 2018, doi: 10.1109/GCCE.2018.8574603.

[9] K. McMullen, and Y. Wan, “A machine learning tutorial for spatial auditory display using head-related transfer functions,” The Journal of the Acoustical Society of America., vol.151(2), pp. 1277-1293, 2022.

[10] D. Valente, “Data-driven prediction of peak sound levels at long range using sparse, ground-level meteorological measurements and a random forest,” The Journal of the Acoustical Society of America., vol.134(5), pp. 4159-4159, 2013.

[11] F. Ning, Z. Cheng, D. Meng, and J. Wei, “A framework combining acoustic features extraction method and random forest algorithm for gas pipeline leak detection and classification,” Applied Acoustics., vol.182, pp. 108255, 2021.

[12] F. Brinkmann, et al, “A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses,” Journal of the Audio Engineering Society., vol.67(9), pp. 705-718, 2019.