HRTF PERSONALIZATION MODELING BASED ON RBF NEURAL NETWORK

Lin Li, Qinghua Huang

School of Communication and Information Engineering, Shanghai University, Shanghai, China

ABSTRACT

A tensor is used to describe head-related transfer functions (HRTFs) dependent on frequencies, sound directions and anthropometric parameters. It can represent the multi-dimensional structure of measured HRTFs. To construct a personalization model, high-order singular value decomposition (HOSVD) is firstly applied to extract individual core tensor as the outputs of the model. Some important anthropometric parameters are selected by Laplacian score and correlation analysis between all measured parameters and the individual core tensor. They act as the inputs of the personalization model. Then a nonlinear model is constructed based on radial basis function (RBF) neural network to predict individual HRTFs according to the measured anthropometric parameters. Compared with back-propagation (BP) neural network method, simulation results demonstrate the better performance for predicting individual HRTFs in the midsaggital plane at high elevations.

Index Terms**—**Head-related transfer function, tensor, Laplacian score, radial basis function neural network

1. INTRODUCTION

Head-related transfer function (HRTF) is the core to generate virtual three-dimensional (3D) auditory. It is an acoustical transfer function defined as the ratio of the sound pressure at listener's eardrum to that at the central of head with the listener absent. Its corresponding time domain representation is head-related impulse response (HRIR). In fact, HRTF changes with frequencies, sound directions and individual physiological structure. The tiny difference of anthropometric shape and size can create a significant influence on HRTFs for sound location. Perceptual distortions may occur in spatial hearing using generic HRTFs without the individual difference. Therefore, it is necessary to personalize HRTFs.

Synthesis of ideal 3D acoustic environment for a person needs a lot of measured HRTFs with special instruments. It is time consuming and difficult to implement. Therefore it is not practical and economical for applications. There are some theoretical calculation methods such as snowman model [1] and boundary element method (BEM) [2]. However, they have a large amount of calculation. Hu et al. proposed partial least squares regression (PLSR) [3] and back-propagation neural network (BPNN) [4] to build the relation between some anthropometric parameters and HRTFs. Due to high dimension of HRTFs, principal component analysis (PCA) was applied to get individual weight coefficients and basis vectors. However, it requires a vectorization process of the original HRTFs. Reshaping a multi-dimensional HRTF into a vector obviously breaks the inherent structure and may bring the loss of potential correlation in the original dataset. To overcome the weakness caused by PCA, Grindlay et al. [5] put forward a multilinear (tensor) framework for HRTFs. They exploited the HRTFs natural structure factoring by N-mode singular value decomposition (SVD) and obtained better reconstruction performance.

Eleven concatenate measurements were selected for HRTFs customization, which ignore the difference of these anthropometric parameters [5]. Anthropometric shape and size of a person has different influence on HRTFs [6]. Xu et al. [7] found that many parameters of the ear significantly correlated with the magnitudes of HRTFs at the high frequencies, while the neck shows weak correlation with HRTFs at the whole frequency band. Therefore some measurements of anthropometric parameters are unnecessary. A reasonable selection of anthropometric measurements is important for personalized HRTFs customization. Hu et al. [4] used correlation analysis to select a few parameters. Parameter cross-correlation only describes the linear relationship between them. It fails to evaluate the significance of a single parameter.

In this paper, we aim to construct a nonlinear model for individual HRTFs prediction. To keep the inherent interaction of multiple variables to HRTFs, a tensor is used to represent HRTFs. High-order singular value decomposition (HOSVD) is applied to generate the individual core tensor with lower dimension as the outputs of the nonlinear mapping. The inputs are a few anthropometric parameters selected in consideration of the local geometric structure and global information in parameters data space. RBF neural network is applied to learn the nonlinear relation between the selected parameters and the individual core tensor. The rest of this paper is organized as follows. Section 2 describes the proposed algorithm. Section 3 gives the simulation results. The conclusions are given in the last section.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

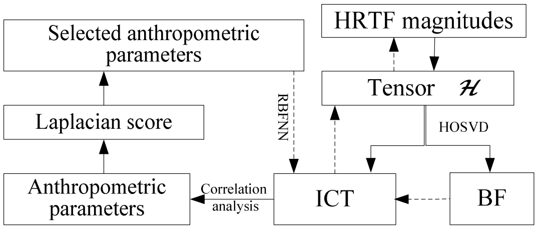

Fig. 1. Proposed HRTFs individualization method.

2. INDIVIDUAL HRTFS CUSTOMIZATION

The goal of the customization is to predict individual HRTFs efficiently via simple anthropometric measurements. Our proposed personalization method seeks to build the nonlinear relation between some selected anthropometric parameters and the individual core tensor representing HRTFs from different sound directions. The schematic diagram is shown in Fig. 1. ICT denotes individual core tensor and BF represents basis function.

2.1. Dimension reduction of HRTF tensor

Firstly, each HRIR is transformed into its corresponding complex HRTF by fast Fourier transform (FFT). HRTFs of different subjects can be described in a 3-mode tensor including sound directions, frequencies, and subjects. Using HOSVD, a core tensor can be obtained as the outputs of the nonlinear personalization model. It keeps the structure of original multi-dimensional data.

A tensor $\mathcal{H} \in \mathbb{R}^{D \times F \times P}$ denotes HRTF magnitudes of P subjects at D sound directions. Each HRTF has F frequencies. The dimension of $\mathcal{H}$ is very high, so it is necessary to reduce its dimension. In order to extract the individual core tensor, the dimensions of frequency and direction should be reduced and the mode of subject is unchanged. HOSVD is the extension of conventional SVD for higher-order tensor decomposition [8]. $\mathcal{H}$ can be decomposed by HOSVD in the first and second mode space as

$$\mathcal{W} = \mathcal{H} \times_1 \mathbf{U}^{(1)^T} \times_2 \mathbf{U}^{(2)^T} \tag{1}$$

where $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ are the new basis matrices. There are two steps for dimension reduction of HRTF tensor. One is the calculation of the transform matrix $\mathbf{U}^{(q)}(q=1,2)$ and the other is the computation of the ICT. $\mathbf{U}^{(q)}$ is the left singular matrix of the n-mode unfolding matrix of $\mathcal{H}$. $\tilde{\mathbf{U}}^{(q)}$ called basis function is the truncated matrix of $\mathbf{U}^{(q)}$. By multiplying the basis functions, $\mathcal{H}$ is projected into $\tilde{\mathcal{W}} \in \mathbb{R}^{D^{'} \times F^{'} \times P}(D^{'} < D,\ F^{'} < F)$ called the ICT. Figure. 2 shows the decomposition of HRTF tensor approximate reconstruction of the original HRTFs is through

$$\tilde{\mathcal{H}} = \tilde{\mathcal{W}} \times_1 \tilde{\mathbf{U}}^{(1)} \times_2 \tilde{\mathbf{U}}^{(2)} \tag{2}$$

with controllable error.

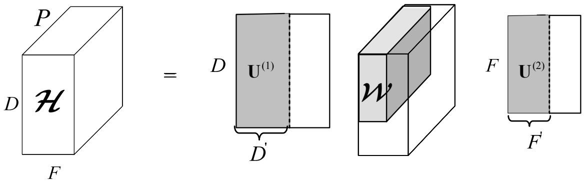

Fig. 2. The decomposition and dimension reduction of HRTF tensor.

2.2. Selection of anthropometric parameters

Secondly, each listener has his specific anthropometric shape and size. We can take following procedure to select the anthropometric parameters as the inputs of the nonlinear personalization model. First, correlation analysis is performed between all the measured anthropometric parameters and the individual core tensor $\tilde{\mathcal{W}}$. It is desirable to delete some parameters with minor correlation coefficients for the selection in the subsequent process. Then in order to avoid the unbalanced selection of similar parameters, k-means is applied to cluster these parameters into m classes. Due to the limitation of cross-correlation analysis, we consider the intrinsic properties of the parameter space to evaluate them by Laplacian score [9]. These parameters of each class are arranged into an ascending sequence according to its Laplacian score, respectively.

Suppose there are P subjects and K parameters of each subject. They can be denoted by a matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_P]$. a_{kp} represents the kth parameter of the pth subject. In order to model the local geometric structure of anthropometric parameters, we construct a nearest neighbor graph G with P nodes $\mathbf{a}_i (i = 1,2,\cdots,P)$. If $\mathbf{a}_i$ and $\mathbf{a}_j$ are close, we put an edge between node i and j with a weight

$$S_{ij} = \frac{\|\mathbf{a}_i - \mathbf{a}_j\|^2}{t} \tag{3}$$

where t is a known constant. Otherwise, $S_{ij} = 0$. $\mathbf{g}_k = [a_{k1}, a_{k2}, \cdots, a_{kP}]$ consists of the kth parameter of P subjects. For a representative parameter, it is reasonable to minimize the following object function

$$L_k = \frac{\sum_{ij}(a_{ki} - a_{kj})S_{ij}}{Var(\mathbf{g}_k)} \tag{4}$$

where $Var(\cdot)$ is the variance. L_k is the Laplacian score of the kth parameter. It captures the local structure and global information. For each parameter, its score is computed to reflect its locality preserving power.

After the above selection process, n parameters with lower scores and significant influence on HRTFs are selected as the final inputs of the personalization model. Without loss of generality, $\mathbf{a}=[a_1,a_2,\cdots,a_n]$ denotes the n selected parameters in the following section.

2.3. HRTF personalization using RBF neural network

When the individual core tensor and a few selected anthropometric parameters are determined, a nonlinear regression model can be learned by RBF neural network. Thus HRTF of a new subject can be predicted via some simple anthropometric parameter measurements. RBF neural network can approximate nonlinear mapping directly from the input and output data with a simple topological structure. A three-layer RBF neural network is applied to learn the intricate relation between the individual core tensor and the selected anthropometric parameters. After the regression model is learned from the training data, the individual HRTF for a new subject can be predicted by his anthropometric parameter measurements.

The input vector is n parameters $\mathbf{a}=[a_1,a_2,\cdots,a_n]$ of a subject. So the input dimension of network is n. The entries of the individual core tensor $\tilde{\mathcal{W}}$ act as the outputs of RBF neural network. The hidden layer has s nodes. The personalization modeling maps the n parameters into a high dimensional space as follows

$$w_i = f_i(\mathbf{a}) = \sum_{j=1}^{s}\Phi_j(\|\mathbf{a}-\mathbf{c}_j\|)\cdot v_{ij} \qquad 1\le i\le N \tag{5}$$

where $\mathbf{c}_j \in \boldsymbol{R}^n$ is the data center vector determined by k-means. $\|\cdot\|$ denotes the Euclidean norm. v_{ij} is a weight calculated by recursive least square (RLS). In general the Gaussian function is chosen as the nonlinear function $\Phi(\cdot)$. w_i is the element of the individual core tensor $\tilde{\mathcal{W}}$ and its number is $N=D^{'}*F^{'}$. After learned from some training data, a fixed structure of RBF neural network is obtained. It can be used to predict $w_i^{(new)}$ for a new listener via corresponding measured anthropometric parameters.

$$w_i^{(new)} = \sum_{j=1}^{s}\Phi_j(\|\mathbf{a}_{new}-\mathbf{c}_j^{(new)}\|)\cdot v_{ij} \qquad 1\le i\le N \tag{6}$$

3. SIMULATIONS

The HRTF personalization is based on a large number of HRTF measurements. The CIPIC database provides high spatial resolution HRIR measurements of 45 different subjects [10]. It contains measured HRIRs for both left and right ears at 1250 directions including 25 azimuths and 50 elevations. Each HRIR is sampled at 44.1 kHz sampling rate and truncated into 200 points. Azimuths vary from -80° to 80° and elevations range from -45° to $+230.625^\circ$.

In the simulations, the fixed azimuth 0° in the mid-saggital plane is chosen to model. The HRIR of each subject was transformed into a HRTF by a 256-point FFT. Total HRTFs of 30 subjects act as training samples denoted by a tensor $\mathcal{H}\in\mathbb{R}^{D\times F\times P}$, where D is the number of elevations (50), F is the number of the frequencies (128), and P is 30 for the training subjects. Five subjects are used for testing and cross-validation method is used.

Table 1. Spectral distortion of HRTF reconstruction at different elevations.

Elevation ($^\circ$)	PCA SD (dB)	HOSVD SD (dB)
−45	9.84	3.85
−16.875	7.34	2.56
11.25	6.66	1.33
39.375	7.73	3.95
67.5	6.54	2.24
95.625	6.23	3.20
180	9.24	4.26
280.125	8.13	2.47

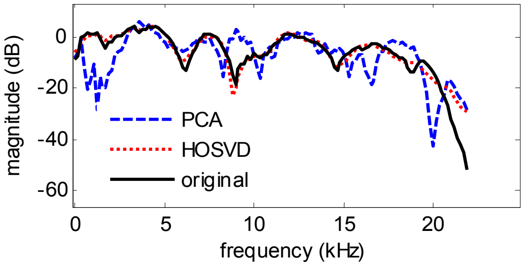

Fig.3. HRTF reconstruction for subject 153.

First of all, HOSVD was applied to the tensor $\mathcal{H}$. We keep the HRTFs in a data reduction of 87% and obtained the individual core tensor $\tilde{\mathcal{W}}\in\mathbb{R}^{10\times20\times30}$. Compared with PCA, reconstruction with basis functions by these two methods is illustrated in Table 1 and Fig. 3. The spectral distortion (SD) is used to measure the reconstruction performance and subsequent regression performance. It is computed in dB as follows

$$\mathrm{SD}=\sqrt{\frac{1}{F}\sum_{i=1}^{F}\left(20\lg\frac{|\mathcal{H}(f_i)|}{|\hat{\mathcal{H}}(f_i)|}\right)^2} \tag{7}$$

where $\mathcal{H}(f_i)$ and $\hat{\mathcal{H}}(f_i)$ denote the ith frequency of the measured and reconstructed HRTF, respectively. HOSVD captures more representative information.

Then correlation analysis [4] is done between $\tilde{\mathcal{W}}$ and 27 anthropometric parameters. The result demonstrates x_5, x_{11}, d_2, and θ_1 with the average of absolute coefficients less than 0.14. Then using the similarity of 27 anthropometric parameters, they were clustered into 3 classes by k-means. These parameters of each class were arranged by Laplacian score into an ascending sequence. They are given in Table 2.

Table 2. Parameters arranged by its Laplacian score in an ascending order.

Class 1	x_2 , x_{11} , x_{16} , x_3 , x_{12} , x_9
Class 2	d_7 , d_2 , θ_2 , d_1 , d_6 , d_3 , x_5 , d_4 , d_8 , θ_1 , d_5 , x_4 , x_8 , x_1 , x_6 , x_7, x_{13} , x_{10}
Class 3	x_{15} , x_{17} , x_{14}

The half of class 1 and class 2 are reserved. Due to x_{15} , x_{17} , and x_{14} in class 3 with large correlation coefficients, all of them are reserved. On the contrary, we delete x_5 , x_{11} , and d_2 in the remaining anthropometric parameters due to their weak correlation with the ICT.

In this way, 12 parameters including x_2 , x_{14}~x_{17}, d_1 , d_3 , d_4 , d_6~d_8 , and θ_2 are finally selected as the inputs of the RBF neural network. x_{14}~x_{17} are important measurements for the construction of HRTF estimating models [7]. But they are not included in the selected parameters of [3].

These two steps yielded the inputs and outputs of RBF neural network. In reference [4], 16 hidden units can achieve the lowest sum-square error value. In this paper, there are also 16 neurons in the hidden layer of RBF neural network. Its prediction performance is compared with BPNN. The SD performance of BP and RBF for subject 153 at all elevations in the midsaggital plane is shown in Fig. 4. Here the $\hat{\mathcal{H}}(f_i)$ in (7) is the individual HRTF. It can be seen that our proposed method has achieved smaller SD in elevations ranged from 123.75° to 230.625°. In the Fig. 5, the discrepancy between the original HRTFs and individual HRTFs may be caused by the information loss in dimension reduction and the inherent defect of RBF neural network. In general, they approximate the original HRTFs.

4. CONCLUSIONS

In this paper, HOSVD extracts the individual core tensor from the original HRTFs. Anthropometric parameters are selected by Laplacian score and correlation analysis. Then a nonlinear regression model for individual HRTFs is constructed by RBF neural network. Simulation results demonstrate that HOSVD has better reconstruction performance than PCA for high-dimensional HRTF data. And the prediction of individual HRTFs by RBF neural network is more accurate than BP for high elevations. However, there is larger spectral distortion in the spectral notch. In the future, we will improve it and use the proposed HRTF customiza-tion for listening experiments.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation (61001160) and Innovation Program of Shanghai Municipal Education Commission (12YZ023) of China.

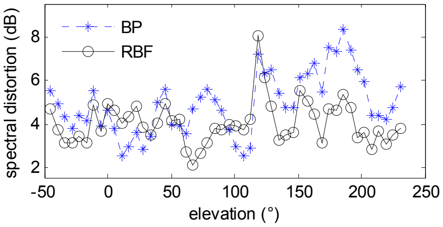

Fig. 4. Prediction performance of BP and RBF networks.

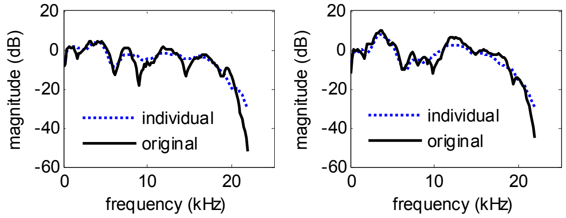

Fig. 5. Individual HRTF for subject 153 at $(0^{\circ}, -45^{\circ})$ and $(0^{\circ}, -22.5^{\circ})$.

REFERENCES

[1] Gumerov N, Duraiswami R, and Tang Z, “Numerical study of the influence of the torso on the HRTF,” in *IEEE Conf. on Acoutics, Speech and Signal Processing*, pp.1965-1968, 2002.

[2] Algazi V, Duda R, Duraiswami R, Gumerov N, and Tang Z, “Approximating the head-related transfer function using simple geometric models of the head and torso,” *Journal of the Acoustical Society of America*, vol. 112, no. 5, pp.2053-2064, 2002.

[3] Hu H, Zhou L, Ma H, and Wu Z, “Head-related transfer function personalization based on partial least square regression,” *Journal of Electronics & Information Technology*, vol. 30, no. 1, pp.154-158 2008.

[4] Hu H, Zhou L, Ma H, and Wu Z, “HRTF personalization based on artificial neural network in individual virtual auditory space” *Journal of Applied Acoustics*, vol. 69, no. 2, pp.163-172, 2008.

[5] Graham G and M. Alex O. V, “A multilinear (tensor) framework for HRTF analysis and synthesis,” in *IEEE Conf. on Acoutics, Speech and Signal Processing,* pp.161-164, 2007.

[6] Rothbucher M, Habigt T, Habgit J, Riedmaier T, and Diepold K, “Measuring anthropometric data for HRTF personalization,” *proceedings of the 6th International Conf. on SITIS*, vol. 112, no. 5, pp.102-106, 2010.

[7] Xu S, Li Z, Zeng L, and G. Salvendy, “A study of morphological influence on head-related transfer functions,” in *IEEE International Conf. on Industrial Engineering and Engineering Management*, pp.472-476, 2007.

[8] L de Lathauwer, B de Moor, and J Vandewalle, “A multilinear sigular value decomposition,” *SIAM Journal of Matrix Analysis and Applications*, vol. 21, no. 4, pp.1253-1278, 2000.

[9] He X, Cai D and Partha N, “Laplacian Score for Feature Selection,” *Proc. of Advances in Neural Information Processing Systems*, pp.507-514, 2006.

[10] Algazi V, Duda R, Thompson D, and Avendano C, “The CIPIC HRTF database,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.99-102, 2001.