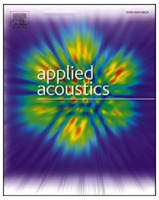
Applied Acoustics 219 (2024) 109938

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust



Efficient prediction of individual head-related transfer functions based on 3D meshes

Jiale Zhao [a,b], Dingding Yao [a], Jianjun Gu [a,b], Junfeng Li [a,b,*]

[a] Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China
[b] University of Chinese Academy of Sciences, Beijing, 100049, China

ARTICLE INFO

Keywords:
Spatial audio
HRTF individualization
3D meshes
Deep learning

ABSTRACT

Individual head-related transfer functions (HRTFs) are critical for binaural spatial audio rendering. In contrast to anthropometric parameters and pinnae images, 3D meshes allow for a more direct and comprehensive representation of the anthropometric structure, which provides highly effective inputs for modeling individualized HRTFs. This paper presents a neural network-based method for predicting individualized HRTFs in full space based on 3D meshes. Unlike many previous methods that estimate HRTF spectra at sampling grids or frequencies separately, the proposed model predicts the HRTF spectra of each vertical plane by considering the spectral correlation and continuity across adjacent sampling grids and frequencies. Evaluation results indicate that the proposed method enhances the prominence of peaks and notches in the obtained HRTF spectra and improves the speed and accuracy of HRTF individualization. The log spectral distortion of the proposed method is lower than that of state-of-the-art methods using anthropometric parameters and pinnae images. Further evaluation confirms that the proposed method requires significantly fewer points in 3D meshes when compared to numerical simulation methods. The evaluation based on localization models demonstrates that the HRTFs predicted by the proposed method are perceptually similar to the measured HRTFs.

1. Introduction

Head-related transfer functions (HRTFs) represent the physical effects on sound waves such as reflection and diffraction caused by the anthropometric structures of individuals [1]. The augmented reality/virtual reality (AR/VR) systems utilize HRTFs to provide immersive binaural audio rendering [2]. The utilization of dummy-head or incompatible HRTFs may result in localization errors of sound sources, including front-back and up-down confusions [3]. The rendering of high-quality spatial audio necessitates the utilization of HRTFs with high spatial resolution [4]. Consequently, there is a significant research interest to obtain individualized HRTFs in full space.

The most direct way for obtaining individualized HRTFs is the measurement in an anechoic chamber [5]. However, the measurements are time-consuming and must be conducted under strict conditions and with complex equipments [6]. Considering the high effort of obtaining individualized HRTFs from direct measurements, some approximation methods were proposed to predict HRTFs based on existing HRTF datasets.

Given that the anthropometric structures of the pinnae, head, and torso are primary influential factors on HRTFs, some methods employ anthropometric parameters to represent the characteristics of these structures and approximate individualized HRTFs [7]. The individualized HRTFs are then predicted by correlating the anthropometric parameters with the dataset [8] or the use of deep neural networks [9]. Due to the reflections, interferences, and other physical processes of sound waves, the high-frequency components of HRTFs are mainly influenced by the details of pinnae [10]. The monaural cues in high frequencies play a crucial role in determining vertical sound localization [11]. However, the anthropometric parameters have limitations in accurately representing the comprehensive characteristics of pinnae, resulting in prediction errors of the high-frequency HRTF magnitude spectra [12]. The utilization of pinnae images has been suggested to offer additional details regarding the structure of pinnae [13]. Nevertheless, affected by conditions such as the capturing positions and distance of cameras, there are a lot of challenges for extracting accurate and robust features from complex images, resulting the limited improvements in predicting HRTFs [14]. Consequently, the HRTF in-

[*] Corresponding author at: Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China.

E-mail address: lijunfeng@hccl.ioa.ac.cn (J. Li).

https://doi.org/10.1016/j.apacoust.2024.109938
Received 22 October 2023; Received in revised form 26 January 2024; Accepted 19 February 2024
Available online 26 February 2024
0003-682X/

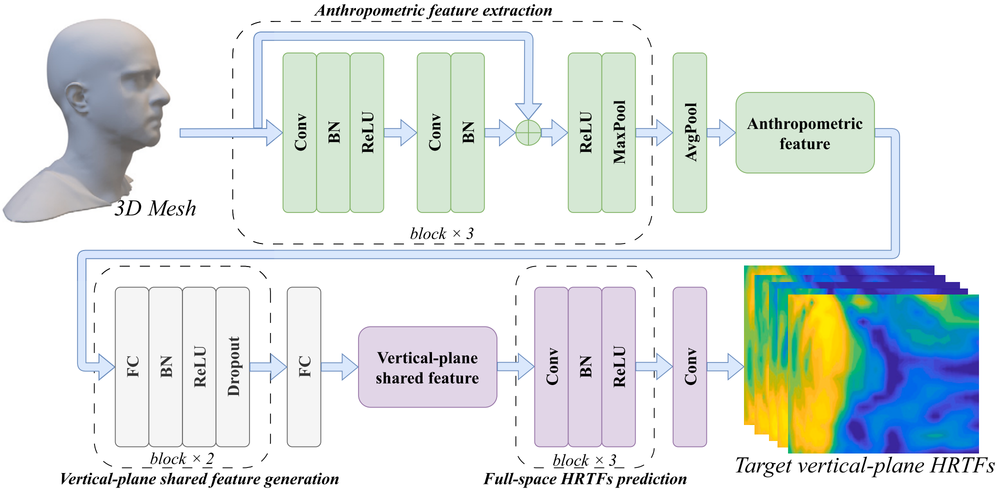

Anthropometric feature extraction
3D Mesh
Conv
BN
ReLU
Conv
BN
ReLU
MaxPool
block × 3
AvgPool
Anthropometric feature
FC
BN
ReLU
Dropout
block × 2
FC
Vertical-plane shared feature
Conv
BN
ReLU
block × 3
Conv
Target vertical-plane HRTFs
Vertical-plane shared feature generation
Full-space HRTFs prediction

Fig. 1. The architecture of the proposed model for predicting vertical-plane HRTF magnitude spectra in full space.

dividualization using 3D meshes is of great interest because the 3D meshes provide the comprehensive representation of individuals.

Various numerical simulation methods have been proposed to simulate individualized HRTFs from 3D meshes, such as the boundary element method (BEM) [15] and the finite-difference time-domain (FDTD) method [16]. The simulation accuracy of HRTFs is susceptible to the accuracy of 3D meshes. The characteristics of hair, skin, and clothes may contribute to deviations in simulated HRTFs compared to the measured HRTFs, primarily due to limitations like rigid boundary conditions [17,18]. Although there are faster methods to accelerate the simulation [19,20], the simulation methods are still complex and time-consuming. Furthermore, accurately capturing the 3D meshes of individuals poses a significant challenge, which is primarily due to the limitations in the scanner resolution and complex postprocessings [21].

Given limitations of the simulation methods, the neural networks were utilized to predict individualized HRTFs based on 3D meshes due to its strong abilities in feature extraction. The model proposed in [22] utilized ear meshes to predict the magnitudes of simulated HRTF spectra, highlighting the advantages of neural networks in predicting speed and complexity. However, this method only utilized ear meshes to characterize the anthropometric structures of subjects, while neglecting the torso and head. In addition, the training data in [22] was generated by the FDTD method. Due to the presence of skin, hair, and other factors, there are significant differences between simulated HRTFs and measured HRTFs [23], leading to the limitations of this method in predicting measured HRTFs. This model can be considered as a model of the FDTD method, rather than the acoustic processes between anthropometric structures and HRTFs. Therefore, there remains a lack of efficient and precise methods for predicting measured HRTFs utilizing 3D meshes.

This paper proposes a neural network-based method for predicting individualized HRTF magnitude spectra using 3D meshes. The proposed method aims to establish the interactions between anthropometric structures and HRTFs by utilizing the comprehensive representation of anthropometric features in 3D meshes, to achieve more accurate HRTF individualization. Given the correlation and continuity of HRTF magnitude spectra across various sampling grids and frequencies, the proposed model concurrently predicts the HRTF spectra for each vertical plane in full space. This strategy exhibits the excellent performance in terms of predicting notch locations, training speed, and prediction efficiency. Evaluation results demonstrate that the prediction error of the proposed method is lower than that of state-of-the-art methods using anthropometric parameters and pinnae images. Further evaluation demonstrates that the proposed method requires much fewer points in 3D meshes in contrast to numerical simulation methods. The evaluation based on localization models illustrates that the HRTFs predicted by the proposed method are perceptually similar to the measured HRTFs.

2. Method

The common models for predicting individualized HRTFs involve setting a single HRTF spectrum as the output [24,25]. Due to the independent prediction of HRTF magnitude spectra, this kind of model has not effectively utilized the correlation and continuity of HRTF magnitude spectra across various sampling grids and frequencies. The variations of peak and notch frequencies of HRTFs in different directions are crucial for elevation perception [26,27]. With the azimuth of the sound source gradually changes, the sound wave is influenced by various anthropometric structures, leading to the corresponding changes in peak and notch frequencies. Considering the continuity in variations of characteristics between adjacent vertical-plane spectra, a vertical-plane shared feature is proposed to represent the features across all vertical planes. Consequently, the vertical-plane HRTF spectra for each azimuth can be generated from this vertical-plane shared feature. Given that predicting the HRTF magnitudes is much more challenging than predicting interaural time differences (ITDs), this paper primarily focuses on estimating the magnitude spectra of HRTFs. The architecture of the proposed model is described in Fig. 1.

2.1. Anthropometric feature extraction

As depicted in Fig. 1, the proposed model takes the 3D mesh, which contains the torso, head and pinnae, as input to generate the anthropometric feature. The anthropometric feature is a one-dimensional vector, which represents the characteristics of the anthropometric structure. The neural networks for extracting anthropometric feature consist of three blocks, each composed of a one-dimensional convolution layer, batch normalization, rectified linear unit (ReLU) activation, and max pooling. To preserve the details of the original mesh to the greatest extent possible, a skip-connection is utilized in each block [22]. The anthropometric feature vector is output by an average pooling operation. The convolution layers were configured with channel sizes of 3, 36, 72, and 144, respectively. The kernel size was set to 3 and the padding size was set to 1. Additionally, both kernel and stride size of pooling operations were set to 8.

2.2. Vertical-plane shared feature generation

The neural networks for generating the vertical-plane shared feature contain two blocks and an output fully connected (FC) layer. Each

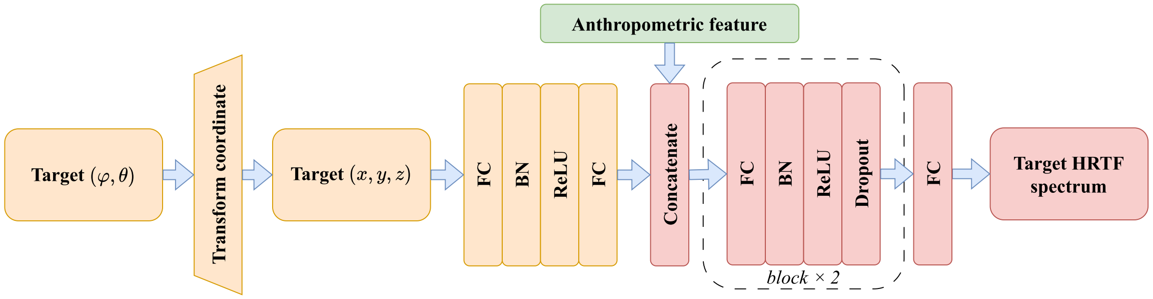


Fig. 2. The architecture of the comparative model for predicting single HRTF magnitude spectrum in the target direction.

block comprises a fully connected layer, followed by the batch normalization, ReLU activation, and dropout. The output of this module is reshaped into a two-dimensional vector of size $N_e \times N_f$ to represent the shared feature across different vertical planes, where N_e represents the number of elevations in each vertical plane and N_f represents the number of frequency bins. The vertical-plane HRTF magnitude spectra in each azimuth can be predicted with this shared feature. As $N_e = 18$ and $N_f = 103$ in this paper, the width of fully connected layers was set to 144, 512, 1024, and 1854, respectively.

2.3. Full-space HRTFs prediction

The neural networks for predicting target HRTF magnitude spectra consist of three blocks and a convolution layer. Each block is constructed using a convolution layer, followed by the batch normalization and ReLU activation. The output of this module is a three-dimensional vector of size $N_a \times N_e \times N_f$, where N_a represents half the number of azimuths. The number of channels in the convolution layers was set to 16, 32, 32, and 36, respectively. The kernel size of the convolution layers was set to 3, and the padding size was set to 1. This module constructs HRTFs in each vertical plane from the shared feature. By utilizing the proposed model depicted in Fig. 1, individualized HRTF magnitude spectra in full space can be predicted at once.

3. Evaluation

3.1. Dataset and preprocessing

The evaluations in this paper were all conducted on the 3D3A dataset [28]. Head-related impulse responses (HRIRs) were measured for each subject at 648 directions on a spherical surface. The azimuth ranges from 0° to 360° with a 5° azimuth interval, where 0° corresponds to front and 90° corresponds to left. For each azimuth, there are eight elevations equally spaced in 15° increments from −30° to 75°, with an additional sampling grid positioned separately at −57° elevation. The elevation of 0° corresponds to the horizontal plane. The HRIRs were resampled at a rate of 44.1 kHz. The 3D3A dataset consists of 30 individuals with valid meshes, excluding the dummy head. The HRTFs were obtained by utilizing a 256-point Fast Fourier Transform (FFT) on HRIRs. The HRTF magnitude spectrum used in experiments is a 103-dimensional vector ($N_f = 103$), preserving the HRTF spectrum of the frequency range between 200 Hz and 18 kHz.

In consideration of the auditory perceptual characteristics, the magnitude spectrum of HRTF was obtained by performing a logarithmic domain transformation [29], which is formulated as

$$H_{\log} = 20 \log_{10} |H| \tag{1}$$

where $H_{\log}$ is the magnitude spectrum of HRTF (H) in the logarithmic domain. The magnitude spectra of HRTFs were smoothed using an equivalent rectangular bandwidth (ERB) filter to eliminate imperceptible microscopic patterns in the high-frequency range, such as spectral burrs [30–32]. The 3D mesh is presented by a group of three-dimensional coordinates in space, which are obtained via the 3D scanner. The head and torso scans were acquired using a PrimeSense Carmine 1.08 sensor, providing a 3.4 mm resolution, while the pinnae scans were obtained using an Artec Space Spider structured-light scanner with a resolution of 0.1 mm [28]. The 3D mesh transforms the anthropometric structures of the subject, including the head, torso, and pinnae, into the numerical form. The 3D meshes were resampled at N_m points and oriented to face the positive x-axis, with heads aligned along the positive z-axis. Initially, N_m was set to 2048. Further discussion regarding N_m is presented later.

3.2. Training settings

Without loss of generality, the evaluations in this paper focus on the prediction performance of the left ear. The dataset of 30 subjects was randomly divided into three groups, with 24, 3, and 3 subjects in the train, validation, and test sets, respectively. To achieve more generalized results, a 10-fold cross validation was conducted. The proposed model utilized the adaptive moment estimation (Adam) optimizer during training [33]. The initial learning rate was set to 0.001 and decayed by a factor of 0.9 every 10 epochs. The batch size was 2048, and the training process comprised 200 epochs. The rates of the dropout layers were all set to 0.5. The mean absolute error (MAE) was utilized as the loss function, given by

$$\mathrm{MAE}(H_{\log}, \hat{H}_{\log}) = \frac{1}{K} \sum_{k=1}^{K} |H_{\log}(k) - \hat{H}_{\log}(k)| \tag{2}$$

where $H_{\log}(k)$ and $\hat{H}_{\log}(k)$ are the measured and predicted HRTF spectra at the kth frequency bin, respectively.

3.3. Performance evaluation of the proposed model

In order to illustrate the effectiveness of the architecture of the proposed model, the widely-used architecture of models for predicting the single HRTF magnitude spectrum was constructed as the comparative model [12,22,24]. The model depicted in Fig. 2 takes the 3D mesh, target azimuth and elevation as the input, and outputs the corresponding target HRTF magnitude spectrum.

The comparative model receives the target direction (φ, θ) and 3D mesh as input, where φ represents the target azimuth and θ represents the target elevation. The anthropometric feature is generated by the same module in Fig. 1. The target direction described using the polar coordinate may lead to misunderstandings for neural networks, particularly in distinguishing between $\varphi = 2\pi$ and $\varphi = 0$. Therefore, the target direction is transformed into (x, y, z), which resides on the unit sphere in Cartesian coordinates. To prevent any dimension mismatch between the target direction and the anthropometric feature [12], the extended feature of the target direction is generated. The width of fully connected layers for extending the feature of target direction was set to 3, 64, and 128, respectively. Consequently, the comparative model combine the extended feature of target direction and the anthropometric feature to predict the HRTF magnitude spectrum. The width of fully connected layers in this part was set to 272, 256, 128, and 103, respectively.

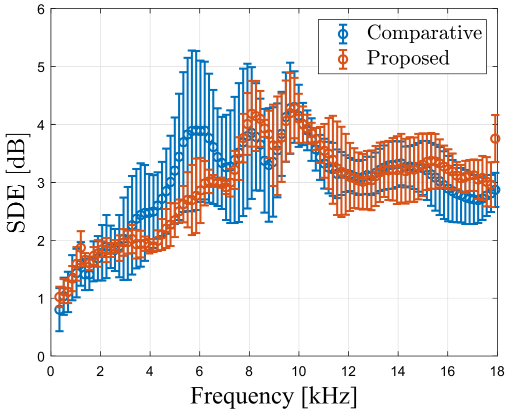


Fig. 3. The distribution of average SDE across all directions and subjects in the test set.

To illustrate the error distribution across different frequency bins, the spectral distance error (SDE) is used to evaluate both models, given by

$$\mathrm{SDE}(\varphi,\theta,s,k) = |H_{\log}(\varphi,\theta,s,k) - \hat{H}_{\log}(\varphi,\theta,s,k)| \tag{3}$$

where s denotes the sth subject in the test set.

The Fig. 3 depicts the distribution of average SDE across all directions and subjects in the test set. There is small difference in the mean SDE between the proposed and comparative model. The comparative model exhibits higher standard variance in the medium and low frequency bins below 10 kHz, while the proposed model exhibits higher standard variance in the high frequency bins above 10 kHz. There is no significant difference between the performance of both models in terms of SDE.

The contour maps in different vertical planes were analyzed further for both models to show the details of the predicted HRTF magnitude spectra. To highlight the differences between each predicted and measured contour map, the correlation coefficients were evaluated and labeled above the subplots. As shown in Fig. 4, the contour maps predicted by the proposed model share more similarities with the measured contour maps. Compared with the comparative model which predicts the single HRTF spectrum, the proposed model exhibits more accurate peak and notch locations. As the elevation and azimuth change, the HRTF spectra predicted by the comparative model manifest discontinuities. For instance, when the elevation is changed from 75° to 105°, noticeably discontinuous amplitude changes are observed in the contour maps of the comparative model. In contrast, the HRTF spectra predicted by the proposed model exhibit superior continuity in both amplitude and notch locations, due to the utilization of the vertical-plane shared feature.

Furthermore, the proposed model significantly outperforms the comparative model in terms of prediction efficiency, because it predicts the HRTF spectra for full space in a single operation. Under the same experimental conditions, where both models were trained and evaluated on a P100 GPU, the training and predicting time for one epoch were evaluated. As shown in Table 1, the results demonstrate the superior training and predicting efficiency of the proposed model. All subsequent evaluations were exclusively carried out on the proposed model.

3.4. Comparison with other methods

The proposed method was compared to state-of-the-art HRTF individualization methods using anthropometric parameters, pinnae images, and 3D meshes. The AE-DNN-VAE method in [12] utilizes an autoencoder (AE) to extract features of anthropometric parameters and HRTF spectra, and a variational autoencoder (VAE) to decode these features into the target HRTF spectrum. The UNet-EAR method introduced in [14] utilizes neural networks to extract features from pinnae images and create individualized HRTFs. Additionally, the DNN-BEM method suggested in [34] employs 3D meshes to simulate HRTFs using BEM and combines anthropometric parameters with the simulated HRTFs to predict measured HRTFs based on neural networks. To compare the prediction error of each method, the log spectral distortion (LSD) metric is utilized to quantify the disparity between the measured HRTFs (considered as the ground-truth) and the predicted HRTFs, given by

$$\mathrm{LSD}(\varphi,\theta,s) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(H_{\log}(\varphi,\theta,s,k) - \hat{H}_{\log}(\varphi,\theta,s,k))^2} \tag{4}$$

The average LSD of the proposed method under all sampling directions and subjects in test set was calculated and compared with the AE-DNN-VAE, UNet-EAR, and DNN-BEM methods.

According to Table 2, the proposed method outperforms the compared methods in terms of mean LSD and its standard deviation. In comparison to the AE-DNN-VAE method that utilizes anthropometric parameters, the proposed method incorporates more personalized information about the anthropometric structure. The UNet-EAR method using ear images is affected by the various factors, such as the image background and camera positions, resulting higher LSD and standard deviation. The DNN-BEM method utilizes the deep neural networks to improve the HRTFs simulated from 3D meshes. However, the predicted HRTFs still remain errors due to differences between the simulated and measured HRTFs. Due to that the number of points in 3D meshes is typically in tens of thousands, robust and precise feature extraction from 3D meshes faces significant challenges. The proposed neural network-based method partially mitigates this issue by leveraging a well-designed model, data augmentation, and other techniques. It is worth noting that both AE-DNN-VAE and DNN-BEM method were evaluated on the HUTUBS dataset [35], while the UNet-EAR method was evaluated on the CIPIC dataset [36]. Considering there is no single database consisting of the anthropometric parameters, pinnae images, and meshes at the same time, the compared methods could not be reimplemented using the same preprocessing. The data in Table 2 are all derived from the corresponding papers, aiming to compare the proposed method with the methods using different anthropometric features.

Table 1
Comparison of the training and predicting time (in seconds) between the proposed model for predicting vertical-plane HRTFs and comparative model for predicting single HRTF.

Target prediction	Training [s]	Predicting [s]
single HRTF	960.54	2.26
vertical-plane HRTFs	**19.30**	**0.95**

Table 2
Comparison of the mean LSD across all directions and subjects in test set and corresponding standard deviation (s.d.) between the proposed method and compared methods.

Method	mean [dB]	s.d. [dB]
AE-DNN-VAE	4.21	1.35
UNet-EAR	5.32	3.15
DNN-BEM	4.80	–
Proposed	**3.78**	**1.18**

3.5. Perceptual evaluation based on localization models

3.5.1. Evaluation on median plane

Considering that vertical sound localization is largely determined by the HRTF magnitude spectra, especially in static rendering, the localization performance of the median sagittal plane was evaluated using

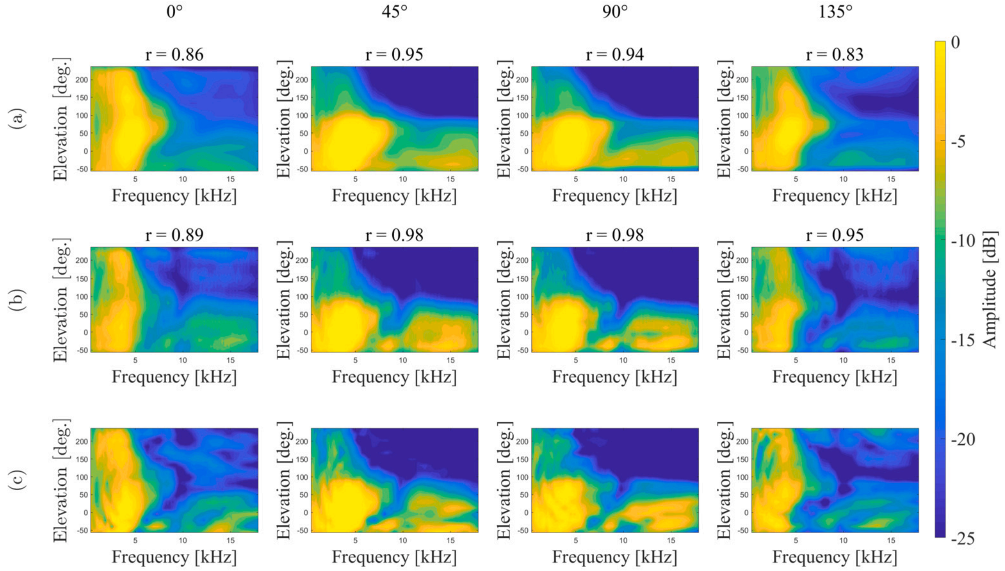


Fig. 4. Comparison of the contour maps in different vertical planes for the (a) comparative model, (b) proposed model, and (c) measured HRTFs. The contour maps in different columns correspond to four different azimuths. All the contour maps depict the HRTFs from the same randomly selected subject in the test set. Spectral amplitude (dB), represented by the heat map color schema, is plotted from 200 Hz to 18 kHz. The elevation ranges from −57° to 75°, and from 105° to 237°. The elevations between 105° and 237° correspond to the contralateral elevations between 75° and −57°. The correlation coefficient, which is denoted as r, indicates the correlation between each predicted and measured contour map.

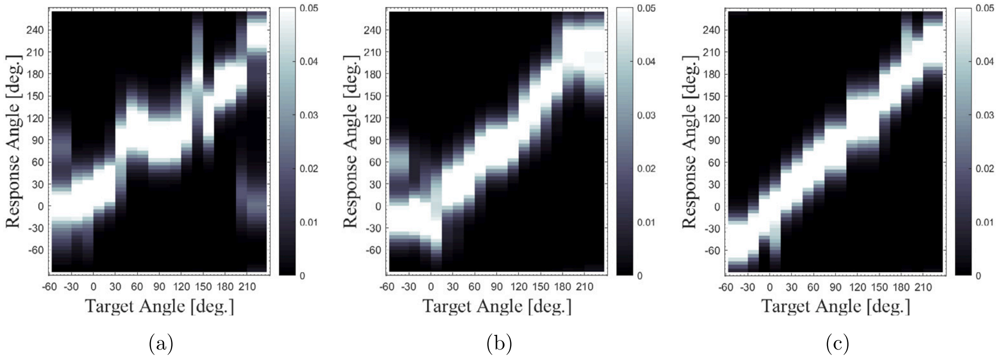


Fig. 5. The localization performance of the (a) generic, (b) proposed, and (c) reference method evaluated by the median-plane localization experiments. The localization performance of the reference method was assessed using the ground-truth data as input. Variation in color within the figures corresponds to the varying probability of localization. Here, a single subject was selected randomly from the test set for evaluating.

the BAUMGARTNER2014 function from the Auditory Modeling Toolbox [37]. This function employs a template-based comparison model that allows for the evaluation of the localization performance of individualized HRTFs. The listener-specific sensitivity threshold was set to 1 and the differential order of the spectral gradient extraction was set to 0. The remaining settings of the function were default. Two runs were performed to minimize irrelevant errors.

The generic method utilizes HRTFs of the dummy head in the 3D3A dataset. In order to avoid the errors caused by irrelevant factors, the reference method using measured HRTFs is also compared with the proposed method. The results of the median-plane localization experiments are illustrated in Fig. 5 and Table 3. The polar root-mean square error (PE) and the angle of error were calculated for assessing the accuracy of localization. When compared to the generic method, it is evident that the performance of the proposed method shares more similarity with the reference method. Fig. 5 provides further evidence that the proposed method performs more similarly to the reference.

Table 3
The results of the proposed and compared methods in median-plane localization experiments.

Method	Confusion rate (%)		Angle of	PE
	Up-down	Front-back	error [deg.]	[deg.]
Generic	26.85	31.48	44.69 ± 4.33	63.64
Proposed	13.89	8.33	27.34 ± 1.31	36.16
Reference	8.33	4.63	16.21 ± 1.22	21.18

3.5.2. Evaluation on horizontal plane

The performance of the horizontal plane was evaluated using the model proposed by [38]. This probabilistic model is constructed using a trained Gaussian mixture model that weighs interaural level differences (ILDs) and ITDs to estimate the azimuthal position of a sound source. It is important to note that ITD individualization is not discussed in this paper. The evaluation in this section pays more attention on the ILDs.

As shown in Table 4, the angle of errors with standard deviations were obtained by comparing the response and target angles. The generic method exhibits a higher localization error on the horizontal plane due to the lack of torso structure in the 3D mesh of the dummy head. The performance of the proposed method is almost the same as the reference method, demonstrating the effectiveness of the proposed method in horizontal localization.

Table 4
The angle of error for the proposed and compared methods in horizontal-plane localization experiments.

Method	Angle of error [deg.]
Generic	3.42 ± 2.45
Proposed	2.75 ± 1.77
Reference	2.70 ± 1.68

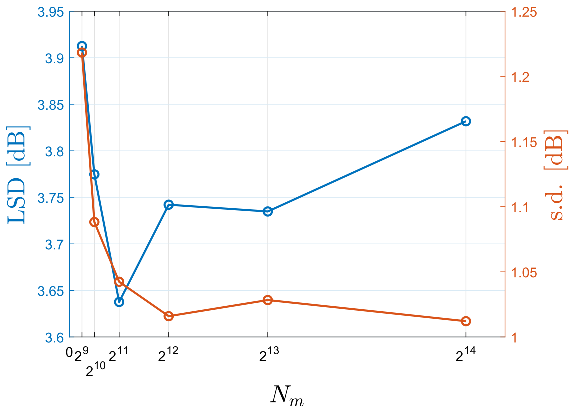


Fig. 6. The mean LSD and corresponding standard deviation (s.d.) for different numbers of points in 3D meshes.

3.6. The effect of 3D mesh density

The number of points in the 3D mesh directly influences the precision of individual physical feature representation, which significantly impacts the accuracy of HRTF estimation outcomes. Regarding numerical simulation methods, 10 or 20 elements per wavelength are suggested for an adequate accuracy [39,16]. Assuming a sound speed of $c = 343$ m/s, the maximum distance between two adjacent elements is $\Delta x = \frac{c}{f_{max} \times 10} = 1.91$ mm, which is challenging to achieve in practical scanning processes. Given that the structure of pinnae requires higher accuracy compared to the structure of the torso and head [40], using the number of points, denoted as N_m, is a more suitable approach to represent the density of the whole 3D mesh, compared to utilizing the distance between adjacent elements. The N_m also represents the size of the input layer in the proposed model.

The number of points in the 3D mesh was set to 2^9, 2^{10}, 2^{11}, 2^{12}, 2^{13}, and 2^{14}, respectively. Fig. 6 illustrates that the prediction error is minimum when N_m is set to 2^{11}. However, as N_m gradually increases, the prediction error does not decrease due to the difficulty in extracting the anthropometric feature from complex 3D meshes and the overfitting issue. Having more points in meshes may represent more features of the anthropometric structure, while the larger database should also be utilized to ensure the robustness of the neural network. Consequently, the optimal number of points in 3D meshes for the proposed method is approximately 2^{11}.

The numerical simulation methods necessitate considering the wavelength of sound waves to determine scanning accuracy, resulting in a large number of points in 3D meshes and subsequently high computational costs, as demonstrated by the utilization of approximately 2^{17} points in [41]. In contrast, the proposed method requires fewer points compared to the numerical simulation methods, thereby resulting in a substantial reduction in computational costs.

4. Discussion

Compared with the anthropometric parameters and pinnae images, 3D meshes can represent more comprehensive personalized characteristics of subjects. The numerical simulation methods calculate individualized HRTFs based on wave equations and corresponding boundary conditions, resulting high demand of both 3D meshes accuracy and computational cost. Owing to the effectiveness of deep neural networks in extracting features, the method introduced in [22] utilized neural networks to predict simulated HRTF magnitude spectra from 3D meshes. However, its modeling capability for measured HRTFs has not been verified, as the training data used is generated by FDTD method. Furthermore, the ear meshes in [22] are voxelized into volume data to facilitate the network construction. The size of 3D tensors is up to 32^3. This kind of preprocessing is not suitable for 3D meshes that include the torso, head, and pinnae. It would result in an excessively large size of the input 3D tensor. Currently, there still remains a scarcity of efficient and accurate methods for predicting measured HRTFs using 3D meshes.

The proposed method utilizes neural networks to extract the anthropometric feature for predicting individualized HRTF magnitude spectra in full space. Considering the continuity of HRTFs across adjacent sampling grids and frequencies, the vertical-plane shared feature is proposed to represent the feature of HRTFs across all vertical planes. Subsequently, the individualized HRTFs of each vertical plane are predicted based on this vertical-plane shared feature. Due to the important spectral features for auditory perception, such as the locations of peaks and notches, consistently vary with sampling grids and frequencies, the architecture of predicting vertical-plane features guides the model to learn more valuable spectral features. Furthermore, the efficiency in training and predicting of the proposed method is significantly improved as it outputs full-space HRTFs at once. The evaluation results demonstrate that the proposed method outperforms the method that predicts each HRTF spectrum separately.

In comparison to the AE-DNN-VAE, UNet-EAR, and DNN-BEM methods, the proposed method provides more accurate HRTF magnitude prediction due to the use of more comprehensive features of the anthropometric structures. Furthermore, the perceptual evaluations demonstrate the outstanding performance of the proposed method. It is worth noting that the torso structure described in this study is the truncated torso. The entire torso, hair, and even clothes may also have effects on HRTFs, contributing to the deviations between measured and simulated HRTFs [42,17,18]. Considering the difficulties in scanning, there is currently no dataset that includes these factors. Therefore, the method proposed in this study has not considered these factors. The localization experiments also indicate instances of the up-down confusions and other inaccurate perceptions. Theoretically, the HRTF modeled by the complete 3D mesh should be nearly indistinguishable from the measured HRTF. With the expansion of databases and advancements in feature extraction techniques, neural network-based models for predicting individualized HRTFs based on 3D meshes will achieve better performance.

Upon keeping the total number of points in meshes constant, utilizing denser grids near the pinnae may capture more precise pinna features, potentially enhancing the performance of the proposed method. This is attributed to the fact that errors in HRTF individualization predominantly arise in high-frequency bins. However, concentrating more points near the pinnae without altering the total number of points may also lead to reduced density in the head and torso, potentially resulting in the loss of critical features. The reduction in prediction error as N_m progressively increases to 2^{11} may be attributed not only to the more comprehensive pinnae features but also to the more comprehensive features of the torso and head. The optimal distribution of each anthropometric structure in the mesh should be explored under the condition of significantly denser meshes to mitigate this uncertainty. Limited by the amount of data, denser 3D meshes may lead to the significant overfitting issue in the proposed neural network, as demonstrated by the results in Fig. 6. In order to preserve the features of different anthropometric structures to the fullest extent, the density distribution of the mesh used in this study remained unchanged relative to the original mesh, as the entire mesh was down-sampled at once. With an expanded database, future research will explore advanced methods for downsampling meshes according to diverse anthropometric structures in order to enhance the performance of the proposed method. In addition, 3D meshes of subjects are typically obtained using 3D scanners in practical applications, while consumer-grade 3D scanners often fail to generate meshes with sufficient accuracy, which further decreases during the postprocessing, such as the noise removal and smoothing. The inaccurate points in meshes may introduce errors in the anthropometric feature extracted by the proposed method. This issue could be discussed in further researches by utilizing the more comprehensive database, which consists of meshes scanned by various consumer-grade 3D scanners.

5. Conclusion

Considering that 3D meshes contain more comprehensive cues for describing the anthropometric structures of subjects compared to anthropometric parameters and pinnae images, a neural network-based model was proposed in this paper to predict more accurate individualized HRTF magnitude spectra based on 3D meshes. The anthropometric feature was first extracted from meshes to represent the individual characteristics of the anthropometric structure, and the vertical-plane share feature was then generated to model the feature of HRTFs across all vertical planes. The individualized HRTF magnitude spectra in full space were finally predicted by the proposed method. Compared with the method which predicts a single HRTF magnitude spectrum at once, the proposed method exhibits the greater efficiency and capability to predict more precise notch locations. In contrast to the state-of-the-art methods using anthropometric parameters, pinnae images, and 3D meshes, the proposed method has lower LSD and greater stability. Moreover, the proposed method requires fewer points in 3D meshes, resulting in the reduction of the computational costs. Further research will be conducted to investigate the improvement of robustness and accuracy using a larger dataset.

CRediT authorship contribution statement

Jiale Zhao: Methodology, Software, Writing – original draft. **Dingding Yao:** Formal analysis, Investigation, Supervision, Writing – review & editing. **Jianjun Gu:** Formal analysis, Supervision. **Junfeng Li:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research is partially supported by STI 2030—Major Projects (No. 2021ZD0201503), the National Natural Science Foundation of China (No. 12104483).

References

[1] Blauert J, Hearing S. The psychophysics of human sound localization. In: Spatial hearing. Cambridge, MA, USA: MIT Press; 1997.

[2] Kleiner M, Dalenbäck B-I, Svensson P. Auralization-an overview. J Audio Eng Soc 1993;41(11):861–75.

[3] Wenzel EM, Arruda M, Kistler DJ, Wightman FL. Localization using nonindividualized head-related transfer functions. J Acoust Soc Am 1993;94:111–23.

[4] Wightman FL, Kistler DJ. Headphone simulation of free-field listening. I: stimulus synthesis. J Acoust Soc Am 1989;85(2):858–67.

[5] Zotkin DN, Duraiswami R, Grassi E, Gumerov NA. Fast head-related transfer function measurement via reciprocity. J Acoust Soc Am 2006;120(4):2202–15.

[6] Majdak P, Balazs P, Laback B. Multiple exponential sweep method for fast measurement of head-related transfer functions. J Audio Eng Soc 2007;55(7/8):623–37.

[7] Algazi VR, Duda RO, Duraiswami R, Gumerov NA, Tang Z. Approximating the head-related transfer function using simple geometric models of the head and torso. J Acoust Soc Am 2002;112(5):2053–64.

[8] Zotkin D, Hwang J, Duraiswaini R, Davis L. HRTF personalization using anthropometric measurements. In: 2003 IEEE workshop on applications of signal processing to audio and acoustics; 2003. p. 157–60.

[9] Chun CJ, Moon JM, Lee GW, Kim NK, Kim HK. Deep neural network based HRTF personalization using anthropometric measurements. In: Audio engineering society convention, vol. 143. 2017.

[10] Shaw EA. Acoustical features of the human external ear. In: Binaural and spatial hearing in real and virtual environments, vol. 25. 1997. p. 47.

[11] Hebrank J, Wright D. Spectral cues used in the localization of sound sources on the median plane. J Acoust Soc Am 1974;56(6):1829–34.

[12] Yao D, Zhao J, Cheng L, Li J, Li X, Guo X, et al. An individualization approach for head-related transfer function in arbitrary directions based on deep learning. JASA Express Lett 2022;2(6):064401.

[13] Lee GW, Kim HK. Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear. Appl Sci 2018;8(11):2180.

[14] Zhi B, Zotkin DN, Duraiswami R. Towards fast and convenient end-to-end HRTF personalization. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing. IEEE; 2022. p. 441–5.

[15] Katz BF. Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. J Acoust Soc Am 2001;110(5):2440–8.

[16] Xiao T, Liu Q Huo. Finite difference computation of head-related transfer function for human hearing. J Acoust Soc Am 2003;113(5):2434–41.

[17] Di Giusto F, Sinev D, Pollack K, van Ophem S, Deckers E, Make F. Analysis of impedance effects on head-related transfer functions of 3D printed pinna and ear canal replicas. In: Proceedings of forum acusticum 2023. European Acoustics Association; 2023.

[18] Brinkmann F, Lindau A, Weinzierl S, Müller-Trapet M, Opdam R, Vorländer M, et al. A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations. J Audio Eng Soc 2017;65(10):841–8.

[19] Ziegelwanger H, Kreuzer W, Majdak P. Mesh2HRTF: open-source software package for the numerical calculation of head-related transfer functions. In: 22nd international congress on sound and vibration; 2015.

[20] Brinkmann F, Kreuzer W, Thomsen J, Dombrovskis S, Pollack K, Weinzierl S, et al. Recent advances in an open software for numerical HRTF calculation. J Audio Eng Soc 2023;71(7/8):502–14.

[21] Gumerov NA, O'Donovan AE, Duraiswami R, Zotkin DN. Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation. J Acoust Soc Am 2010;127:370–86.

[22] Zhou Y, Jiang H, Ithapu VK. On the predictability of HRTFs from ear shapes using deep networks. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing; 2021. p. 441–5.

[23] Prepelita S, Geronazzo M, Avanzini F, Savioja L. Influence of voxelization on finite difference time domain simulations of head-related transfer functions. J Acoust Soc Am 2016;139(5):2489–504.

[24] Chen T-Y, Kuo T-H, Chi T-S. Autoencoding HRTFS for DNN based HRTF personalization using anthropometric features. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing; 2019. p. 271–5.

[25] Miccini R, Spagnol S. A hybrid approach to structural modeling of individualized HRTFs. In: 2021 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops; 2021. p. 80–5.

[26] Iida K, Itoh M, Itagaki A, Morimoto M. Median plane localization using a parametric model of the head-related transfer function based on spectral cues. Appl Acoust 2007;68(8):835–50.

[27] Yao D, Li J, Xia R. A parametric elevation control approach for binaural reproduction. Appl Acoust 2019;148:360–5.

[28] Sridhar R, Tylka JG, Choueiri E. A database of head-related transfer functions and morphological measurements. In: Audio engineering society convention, vol. 143. Audio Engineering Society; 2017.

[29] Smith III JO. Techniques for digital filter design and system identification, with application to the violin. Stanford, California, USA: Stanford University; 1983.

[30] Moore BC. An introduction to the psychology of hearing. Brill; 2012.

[31] Asano F, Suzuki Y, Sone T. Role of spectral cues in median plane localization. J Acoust Soc Am 1990;88:159–68.

[32] Xie B, Zhang T. The audibility of spectral detail of head-related transfer functions at high frequency. Acta Acust Acust 2010;96(2):328–39.

[33] Kingma DP, Ba Adam J. A method for stochastic optimization. In: ICLR (Poster); 2015. p. 1.

[34] Zhang M, Wang J-H, James DL. Personalized HRTF modeling using DNN-augmented bem. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing. IEEE; 2021. p. 451–5.

[35] Brinkmann F, Dinakaran M, Pelzer R, Grosche P, Voss D, Weinzierl S. A cross-evaluated database of measured and simulated HRTFs including 3D head meshes,

anthropometric features, and headphone impulse responses. J Audio Eng Soc 2019;67(9):705–18.

[36] Algazi VR, Duda RO, Thompson DM, Avendano C. The cipic HRTF database. In: Proceedings of the 2001 IEEE workshop on the applications of signal processing to audio and acoustics (cat. no. 01TH8575). IEEE; 2001. p. 99–102.

[37] Majdak Piotr, Hollomey Clara, Baumgartner Robert. AMT 1.x: a toolbox for reproducible research in auditory modeling. Acta Acust 2022;6:19. https://doi.org/10.1051/aacus/2022011.

[38] May T, Van De Par S, Kohlrausch A. A probabilistic model for robust localization based on a binaural auditory front-end. IEEE Trans Audio Speech Lang Process 2010;19:1–13.

[39] Begault DR, Trejo LJ. 3-D sound for virtual reality and multimedia. Tech. Rep. NASA; 2000.

[40] Dinakaran M, Brinkmann F, Harder S, Pelzer R, Grosche P, Paulsen RR, et al. Perceptually motivated analysis of numerically simulated head-related transfer functions generated by various 3D surface scanning systems. In: 2018 IEEE international conference on acoustics, speech and signal processing. IEEE; 2018. p. 551–5.

[41] Ziegelwanger H, Majdak P, Kreuzer W. Numerical calculation of listener-specific head-related transfer functions and sound localization: microphone model and mesh discretization. J Acoust Soc Am 2015;138:208–22.

[42] Lan Y, Yin T, Yu G. Effect of torso reflections from simplified torso models on head-related transfer function simulation and ipsilateral perception of elevation. Appl Acoust 2022;201:109095.