IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 31, 2021, accepted September 9, 2021, date of publication September 15, 2021,
date of current version September 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3113133

Pre-Trained-Based Individualization Model for Real-Time Spatial Audio Rendering System

JINYAN LU[1] **AND XIAOKE QI**[2]

[1]School of Electrical and Information Engineering, Henan University of Engineering, Zhengzhou 451191, China
[2]School of Information Management for Law, China University of Political Science and Law, Beijing 102249, China

Corresponding author: Xiaoke Qi (qixiaoke@cupl.edu.cn)

This work was supported in part by the Scientific Research Innovation Project of China University of Political Science and Law under Grant 10821424, in part by the National Natural Science Foundation of China (NSFC) under Grant 61603390, in part by the Fundamental Research Funds for the Central Universities, in part by the Key Research and Development and Promotion Projects of Henan Province under Grant 202102210187, and in part by the Doctoral Fund of Henan University of Engineering under Grant DKj2018003.

ABSTRACT Spatial audio has attracted more and more attention in the fields of virtual reality (VR), blind navigation and so on. The individualized head-related transfer functions (HRTFs) play an important role in generating spatial audio with accurate localization perception. Existing methods only focus on one database, and do not fully utilize the information from multiple databases. In light of this, a pre-trained-based individualization model is proposed to predict HRTFs for any target user in this paper, and a real-time spatial audio rendering system built on a wearable device is implemented to produce an immersive virtual auditory display. The proposed method first builds a pre-trained model based on multiple databases using a DNN-based model combined with an autoencoder-based dimensional reduction method. This model can capture the nonlinear relationship between user-independent HRTFs and position-dependent features. Then, fine tuning is done using a transfer learning technique at a limit number of layers based on the pre-trained model. The key idea behind fine tuning is to transfer the pre-trained user-independent model to the user-dependent one based on anthropometric features. Finally, real-time issues are discussed to guarantee a fluent auditory experience during dynamic scene update, including fine-grained head-related impulse response (HRIR) acquisition, efficient spatial audio reproduction, and parallel synthesis and playback. These techniques ensure that the system is implemented with little computational cost, thus minimizing processing delay. The experimental results show that the proposed model outperforms other methods in terms of subjective and objective metrics. Additionally, our rendering system runs on HTC Vive, with almost unnoticeable delay.

INDEX TERMS Head-related transfer functions, individualization, pre-trained model, real time, spatial hearing.

I. INTRODUCTION

Currently, augmented reality (AR) and virtual reality (VR) technologies are becoming increasingly popular in our lives. Spatial audio plays an important role not only in VR/AR to further improve the naturalness of the vision scene but also in other applications such as navigation for the blind [1], cochlear implants [2] and entertainment [3].

To generate an immersive auditory experience, the precise localization perception in spatial audio is necessary. Spatial audio is obtained by passing the sound source signal through two filters that contain all the localization-related information of two ears, i.e., head-related transfer functions (HRTFs). HRTFs describe the propagation response of sound waves from the sound source to ear drums in the form of refraction, reflection and diffraction in free space, which is highly related to the anthropometric features of a human, such as head width, cavum concha height, and pinna height [4]. Since each subject has different anthropometric values, HRTFs are highly individual-dependent. The use of non-individualized HRTFs is prone to cause up-down inversion, front-back confusion, and localization in the head [5]. Thus, individual HRTFs are required for obtaining more accurate localization perception. On the other hand, since human hearing is continuous in the physical world, HRTFs should seamlessly adapt to changes in the relative position between the sound sources

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Nappi.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

and the user caused by human and/or source movement or head rotation, to create a dynamic virtual environment in three-dimensional (3D) space. Finally, the update of spatial audio should be fast enough to be imperceptible to the user, and a real-time system is essential for a fluent use experience.

With the above observations in mind, there are several challenges for obtaining an immersive spatial hearing experience. First, HRTFs are highly related to anthropometric features, but the complex shapes of the pinna, the head and the torso pose considerable challenges to mathematically modeling the dense reflection, refraction and diffraction. Second, a high spatial resolution measurement is necessary to cover the 3D space for a dynamic virtual environment. Then, HRTFs should be generated in real time as the moving between the user and the sound sources, and the dense convolution is frequently calculated. These bring great computational load, resulting in a nonfluent system.

The most accurate method to obtain individual HRTFs is through direct measurements of the response from the source to the human ears [6]. However, this is an enormously time-consuming, expensive and nonscalable procedure. In light of this, several alternative methods have been proposed. Based on the physical characteristics of wave propagation, theoretical or numerical models attempt to approximate complicated human anatomy, such as the spherical head model [7], the snowman model [8], structural models [9], the boundary element method [10], and the finite-difference time-domain (FDTD) method [11]. These methods solve the wave propagation equation by taking the human as a sphere, a snowman or a set of elements. FDTD generates accurate HRTFs at the cost of expensive data-acquisition hardware and intensive computation. To alleviate this problem, a mesh grading algorithm is proposed in [12], resulting in significant reduction of the computational costs in the HRTF calculation process.

In light of the high correlation between the HRTFs and human's anthropometric features, various anthropometry-based regression methods have been proposed, which can be categorized into several modes. One is to construct individual HRTFs in full space from a small set of measurements [13]. The individual weights outside the database are estimated from a small set of measurements, and then the HRTFs with highly directional resolution are obtained by combining the spatial basis functions and the weights. [14] proposes a parameter-transfer learning method to obtain individualized HRTFs based on a small set of HRTF measurements. However, this requires a priori measured HRTFs, which do not always exist for each subject. Another method is to use a representation of anthropometric features with the assumption that a given HRTF set can be described by the same combination as the anthropometric data [15]. Based on this study, different preprocessing and postprocessing methods are investigated, and the experimental results show that they have a strong influence on HRTF accuracy [16]. However, the hypothesis in these papers that the combination weights of anthropometric features and those of HRTFs for a person are the same is not completely correct. Without the hypothesis, [17] proposes a sparsity-constrained weight mapping method to build a mapping between weights of anthropometries and those of HRTFs. Another promising solution is to model HRTFs in lower-dimensional space [18] and then build a direct correlation between anthropometries and HRTFs with dimensional reductions using an artificial neural network (ANN) [19], [20], support vector regression in conjunction with principal component analysis (PCA), independent component correlation (ICA) [21], autoencoder or deep neural network (DNN) [22], [23]. On the other hand, the processing latency in a practical system must be limited to an acceptable range, such as 250 ms in [24], to produce a fluent hearing experience when the user or sources are moving freely [24]. Constrained by limited computational resources, real-time implementation of the complex processing from frequent HRTF updates and binaural audio generation in a dynamic environment is highly challenging. As a result, though vision-related techniques have been developed and widely used in VR/AR, auditory-related techniques lag relatively behind by comparison, and they have not been widely applied.

Based on the above-mentioned observations, we aim to build a DNN-based individualization model to achieve a nonlinear mapping between anthropometric features and individual HRTFs since DNN-based techniques have been verified to be powerful in many fields. However, the existing methods utilize one HRTF database. Lack of data may lead to overfit or extract insufficient correlation information. Inspired by the idea of the pre-trained model which has been successfully used in natural language processing (NLP) [25], a pre-trained-based individualization model for real-time spatial audio rendering system is proposed in this paper using multiple databases. First, a pre-trained model is built on multiple databases using a DNN-based model combined with an autoencoder-based dimensional reduction method, resulting in a user-independent model. Then, fine tuning is done using a transfer learning technique to introduce the user-dependent information from the anthropometric features, and achieve an individualization model. Additionally, some real-time issues are discussed in detail, including fine-grained head-related impulse response (HRIR) acquisition, efficient spatial audio reproduction and parallel synthesis and playback, ensuring that the system is implemented with little computational cost. Finally, the rendering system is implemented on a HTC Vive device to produce a fluent and immersive hearing experience.

The remainder of the paper is organized as follows. An overview of the proposed real-time rendering system is presented in Section II. Then, the pre-trained-based individualization model is described in detail in Section III and Section IV. Section V discusses real-time issues. Objective and subjective experiments are conducted and the experimental results are analyzed in Section VI. Finally, Section VII presents our discussion and conclusions.

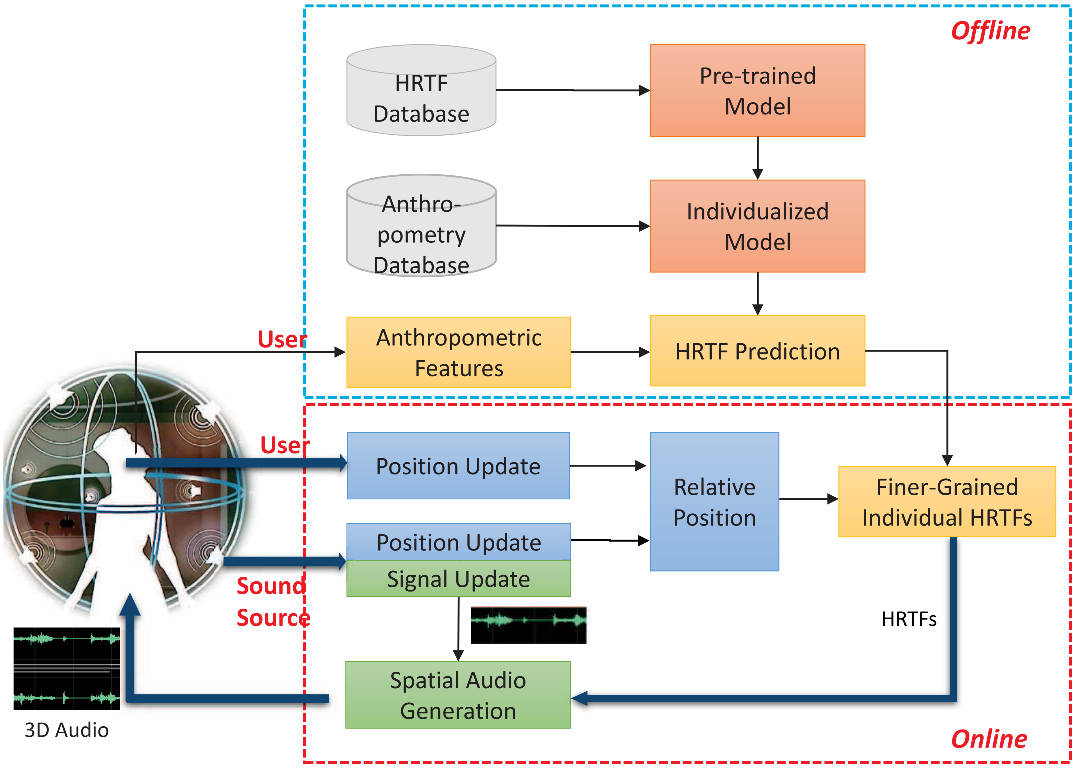


FIGURE 1. The architecture of our proposed individualized real-time spatial audio rendering system based on pre-training and fine-tuning. The parallel operations are marked as bold lines in the figure.

II. SYSTEM OVERVIEW

A conventional spatial audio rendering system involves three steps: acquisition of the relative positions between the user and the sound sources, HRTF selection based on these positions, and spatial audio reproduction. However, these sequential operations are greatly time-consuming in dynamic scene, resulting in a high-latency and nonfluent hearing experience. Furthermore, the lack of individualization is more likely to lead to inside-the-head localization and inaccurate localization perception. To address these problems, we design a real-time spatial audio rendering system on the platform of a VR device, i.e., HTC Vive, as shown in Fig. 1. The basis of the system is to train a pre-trained-based individualization model to achieve individualized HRTFs for accurate localization perception and maximally reduce the computational load during the synthesis and playback of spatial audio.

As shown in Fig. 1, when a user prepares to experience a virtual auditory display, individual HRTFs over full space are required to be convolved with the sound sources to generate spatial audio. To this end, a pre-trained model is built using DNN combined with an autoencoder-based dimensional reduction method, and then the anthropometric features are used to fine tune this model based on transfer learning, generating a HRTF individualization model. After the anthropometric values of a user are measured, individual HRTFs can be obtained using the pre-trained and fine-tuned individualization model.

Furthermore, several strategies are exploited to achieve real-time implementation, by making the maximum number of operations offline and reducing the computational load. First, the proposed individualization models are trained offline. Second, parallel operations are used to set separate threads for the relative position update, synthesis and playback, which produce a fluent auditory experience (indicated using bold lines in Fig. 1). Then, since the user or sound sources can move freely in the virtual scene, the relative positions between the user and sources change continuously. Thus, spatial audio should be synthesized online. Block convolution is used to reduce the computational complexity. These features all focus on minimizing online computational load and achieving a real-time and natural auditory experience.

Specifically, the spatial audio rendering system mainly consists of the following 3 key components:

- a pre-trained model that builds a mapping between the 3D space features and the HRTFs,
- a fine-tuned individualization model that transfers the user-independent pre-trained model to an individualization one, and
- a real-time part that guarantees a fluent and immersive auditory experience.

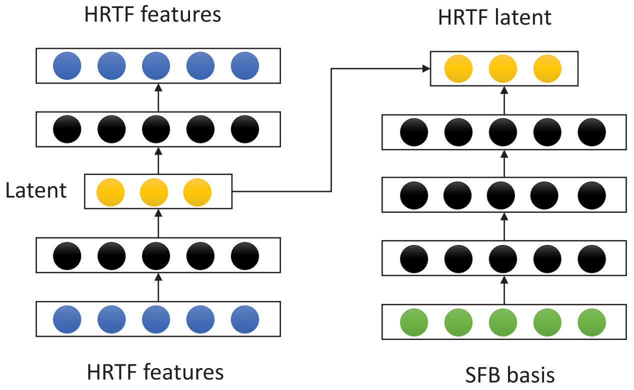


FIGURE 2. The architecture of the proposed pre-trained model.

III. PRE-TRAINED MODEL

Inspired by the idea of pre-training which has been successfully used in NLP, a pre-trained model trained on multiple databases is proposed to generate an individual-independent and position-dependent model over full spatial space. To do this, an autoencoder network reduces the dimension of HRTFs and extracts the latent representation of HRTFs, and then a fully connected DNN builds a mapping between this latent representation and position-relevant features, as shown in Fig. 2.

A. LATENT EXTRACTION OF HRTFs

First, an autoencoder network is used to reduce dimension of HRTFs, since hundreds of points of a HRTF make it prone to overfit when training a DNN-based model.

Before the model training, preprocessing is required to generate HRTF features. In this paper, the log-magnitude HRTFs are chosen to model HRTFs. This choice is based on two observations. First, it was shown in [26] that human is not sensitive to the fine details of the phase spectrum of HRTFs in localization perception. Then, it was experimentally verified in [27] that the logarithmic magnitude of HRTFs is closer to auditory perception of human. To obtain HRTFs on the same scale, we preprocess the log-magnitude HRTFs by calculating the directional transfer function (DTF) followed by feature-scaling normalization as

$$\bar{H}_{m,j,i} = \frac{H_{m,j,i} - \mu_h(i)}{\sigma_h(i)}, \quad i = 1, \ldots, N_b, \tag{1}$$

$$\mu_h(i) = \frac{1}{N_d N_s} \sum_{d=1}^{N_d} \sum_{s=1}^{N_s} H_{s,d,i}, \tag{2}$$

where $H_{m,j,i}$ is the i-th frequency bin of the log-magnitude HRTFs at the j-th direction for the m-th subject. $\mu_h(i)$ and $\sigma_h(i)$ denote the mean and standard variance of log-magnitude HRTFs over all N_d positions and N_s subjects, respectively. $\mu_h(i)$ is also the average DTF of N_s subjects. N_b is the number of frequency bins. Then, when considering HRTFs of left ear and right ear, the preprocessed HRTF matrix can be expressed as

$$\mathbf{H}^p = \begin{bmatrix} \bar{H}^L_{1,1,1} & \cdots & \bar{H}^L_{1,1,N_f} & \bar{H}^R_{1,1,1} & \cdots & \bar{H}^R_{1,1,N_f} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \bar{H}^L_{1,N_d,1} & \cdots & \bar{H}^L_{1,N_d,N_f} & \bar{H}^R_{1,N_d,1} & \cdots & \bar{H}^R_{1,N_d,N_f} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \bar{H}^L_{N_s,1,1} & \cdots & \bar{H}^L_{N_s,1,N_f} & \bar{H}^R_{N_s,1,1} & \cdots & \bar{H}^R_{N_s,1,N_f} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \bar{H}^L_{N_s,N_d,1} & \cdots & \bar{H}^L_{N_s,N_d,N_f} & \bar{H}^R_{N_s,N_d,1} & \cdots & \bar{H}^R_{N_s,N_d,N_f} \end{bmatrix}, \tag{3}$$

where $\mathbf{H}^p \in \mathbb{R}^{N_d N_s \times 2N_f}$. $\bar{H}^L_{m,j,i}$ and $\bar{H}^R_{m,j,i}$ respectively denote the i-th frequency bin of the log-magnitude HRTFs of left ear and right ear at the j-th direction for the m-th subject. Each line of $\mathbf{H}^p$ is an input sample of autoencoder, expressed as $\mathbf{H}^p_n = [\bar{H}^L_{m,j,1}, \cdots, \bar{H}^L_{m,j,N_f}, \bar{H}^R_{m,j,1}, \cdots, \bar{H}^R_{m,j,N_f}] = [H^p_{n,1}, H^p_{n,2}, \cdots, H^p_{n,2N_f}]$ for the n-th input sample, where $n = (m-1)N_s + j$, and thus the number of samples is $N_a = N_d N_s$.

After preprocessing, the samples are fed into the autoencoder to build a pre-trained model, as shown in Fig. 2. An autoencoder network consists of an encoder network connected with a bottleneck layer Enc(·), and a decoder network Dec(·). The bottleneck layer contains a small number of hidden nodes, and outputs a compact representation, called the latent representation of HRTFs, which can be expressed as

$$\mathbf{H}^e = \mathrm{Enc}(\mathbf{H}^p), \tag{4}$$

and the HRTFs can be recovered using a decoder network as

$$\hat{\mathbf{H}}^p = \mathrm{Dec}(\mathbf{H}^e), \tag{5}$$

where $\hat{\mathbf{H}}^p$ denotes the estimate of prepocessed HRTFs $\mathbf{H}^p$, and the n-th line of $\hat{\mathbf{H}}^p$ is expressed as $\hat{\mathbf{H}}^p_n = [\hat{H}^p_{n,1}, \hat{H}^p_{n,2}, \cdots, \hat{H}^p_{n,2N_f}]$.

The autoencoder is trained by minimizing the loss function. We design the loss function by considering human's knowledge of localization perception. Since log-magnitude HRTFs preserve all the perceptually relevant information [28], spectrum distortion (SD) has been widely exploited as the objective metric [15], [19], [29], which is defined as

$$\mathrm{SD} = \sqrt{\frac{1}{2N_a N_f} \sum_{m=1}^{N_s} \sum_{j=1}^{N_d} \sum_{i=1}^{2N_f} \left(H_{m,j,i} - \hat{H}_{m,j,i}\right)^2}, \tag{6}$$

where $\hat{H}_{m,j,i}$ denotes the estimated HRTF of the i-th frequency bin at the j-th position for the m-th subject. Then, the standard variance is used to weight SD of the frequency bins to compensate for the influence of HRTF preprocessing operation. Thus, the loss function is designed as

$$\mathrm{Loss} = \sqrt{\frac{1}{2N_a N_f} \sum_{n=1}^{N_a} \sum_{i=1}^{2N_f} \left[\sigma_h(i) \left(H^p_{n,i} - \hat{H}^p_{n,i}\right)\right]^2}. \tag{7}$$

By relating the loss function to SD, the autoencoder is trained by minimizing the loss function, resulting in minimizing SD. Additionally, the latent representation of HRTFs is obtained as the output of the encoder, which will assist the training of the pre-trained model.

B. PRE-TRAINED MODEL TRAINING

Then, a fully connected DNN builds a mapping between the latent representation of HRTFs $\mathbf{H}^e$ and position-relevant features. The features are generated when considering the propagation of the sound field. The HRTFs are described by the wave equation from the sound source to binaural cochlea, and the solution can be represented by a specific set of orthogonal series related to several factors, such as frequency, distance, azimuth angle, and elevation angle between the source and the user. Based on this, we exploit the spherical Fourier-Bessel (SFB) basis as the position features in 3D space. SFB basis consists of two parts: spherical harmonics (SH) and spherical Bessel functions, which can be used to represent the angular part and radial part of the HRTFs, respectively.

First, considering that the log-magnitude HRTFs are real, real version of SH, a complete set of continuous orthonormal basis functions on a sphere, is used to represent the angular part of the HRTFs. For the elevation angle of ϕ and azimuth angle of θ, SH can be expressed as [27]

$$Y_n^m(\phi,\theta) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}}P_n^{|m|}(\sin\theta)g(|m|\phi), \quad (8)$$

with

$$g(|m|\phi) = \begin{cases} \sin(|m|\phi), & m \le 0 \\ \cos(m\phi), & m > 0, \end{cases} \quad (9)$$

where $n = 0, \ldots, N$ and $|m| \le n$. $P_n^{|m|}(\cdot)$ represents the associated Legendre function with the degree of n and the order of m.

Then, for the radial part of the HRTFs, a normalized spherical Bessel function is used, which is defined on a sphere of radius r as [30]

$$\Phi_{nl}(r) = \frac{1}{\sqrt{N_{nl}}} j_l(k_{nl}r), \quad (10)$$

where $j_l(x)$ is the spherical Bessel function with the order of l, expressed as

$$j_l(x) = \sqrt{\frac{\pi}{2x}} J_{l+1/2}(x), \quad (11)$$

where $J_{l'}(x)$ is the Bessel function with the order of l'. Under the zero-value boundary condition, $k_{nl} = x_{nl}/R_m$ and $N_{nl} = a^3 j_{l+1}^2(x_{nl})/2$. x_{nl} is the nth positive solution to $j_l(x) = 0$ in ascending order, and R_m is the maximum radius.

Therefore, the position features are obtained by concatenating the angular part and radial part for each position, which can be expressed at the position of $\mathbf{d} = (r, \theta, \phi)$ as

$$\mathbf{F}(\mathbf{d}) = [Y_n^m(\theta,\phi), \Phi_{nl}(r)], \quad (12)$$

where $n = 0, \ldots, N$, $|m| \le n$, and $l = 1, \ldots, L$, which contains a total of $N_t = [(N+1)^2 + NL]$ parameters.

Finally, the pre-trained model is trained using the loss function of Eq.(7). It is seen that the model builds a relationship between the position-dependent features and HRTFs, which is individual-independent.

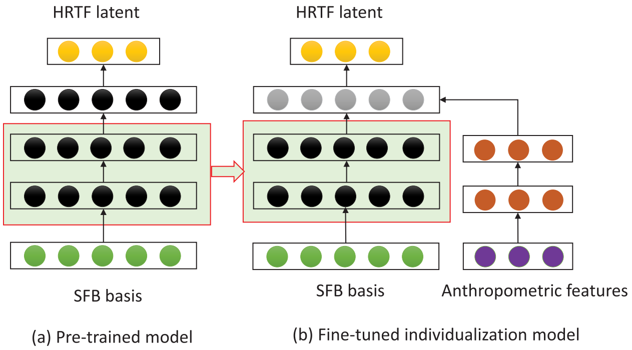


FIGURE 3. The architecture of the proposed fine-tuned individualization model.

IV. FINE-TUNED INDIVIDUALIZATION MODEL

A. MODEL TRAINING

The pre-trained model builds a relationship between position features and HRTFs, and the output is nonindividual. To obtain a HRTF individualization model for a target subject, we introduce anthropometric measurements to tune the pre-trained model to an individualization one, when considering the tight relationship between individualized HRTFs and anthropometries of the subject. The structure of fine-tuned individualization model is shown in Fig. 3. Anthropometric features of subjects are utilized to obtain user-dependent information. Combined with pre-trained model, only a limit number of parameters are tuned to fuse this user-dependent information, while the parameters of other layers are maintained. After fine tuning, the pre-trained model will be adapted to a target user when the user's anthropometric features are input of the individualization model.

First, anthropometric measurements of subjects are required to be preprocessed to limit their values to the same variance. For the k-th anthropometric feature of the m-th subject, denoted as $a_{m,k}$, the normalization is expressed as

$$\bar{a}_{m,k} = \frac{a_{m,k} - \mu_a(k)}{\sigma_a(n)}, \quad (13)$$

$$\mu_a(k) = \frac{1}{N_s}\sum_{m=1}^{N_s} a_{m,k}, \quad (14)$$

where $\mu_a(k)$ and $\sigma_a(n)$ denote the mean and standard variance of the k-th anthropometric feature over all subjects, respectively.

Then, the individualization model is trained to adapt the network parameters to a target. To do this, a transfer learning method is exploited, which do not need to retrain all the parameters. Instead, a potential solution is to only adjust the parameters in a small number of layers, and keep the rest

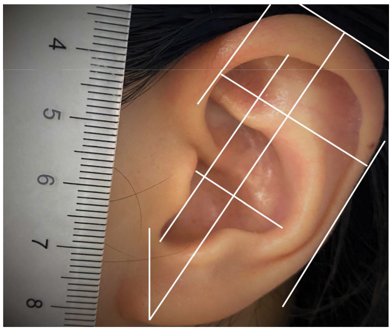


FIGURE 4. Pinna measurements of left ear for a target subject.

parameters unchanged. As shown in Fig. 3, we set the last layer as the fine-tuning layer to adapt with individualized HRTFs of a target user, while the parameters of the rest layers keep the same as the pre-trained model. Furthermore, since it is lack of theoretical study to show which layer is best in fine tuning, the other layers, such as the beginning hidden layer or the middle hidden layer, can also be chosen to fine tune the pre-trained model.

B. MODEL PREDICTION

For generating individualized HRTFs of a target subject, anthropometric parameters should be measured in advance, which can be divided into two parts of body measurements and pinna measurements. For body measurements, we directly measure the target including 17 parameters such as head width, head height, using a ruler. The parameters are the same as [31]. Then, 10 measurements for each pinna are obtained using the method in [32] via a camera, as shown in Fig.4. For each subject, anthropometric measurements will be measured only once before running the spatial audio system.

Given anthropometric measurements of a target subject, the trained individualization model can be used to predict individualized HRTFs based on the relative position between the target and the sound source. First, the model input, anthropometric features and SFB basis at the position $\mathbf{d} = (r, \theta, \phi)$, can be generated and preprocessed, and then forwarded to the fine-tuned individualization model, outputting the estimated latent representation of HRTFs. Passing the latent representation into the decoder network Dec(·), the log-magnitude HRTFs $\hat{\mathbf{H}}^p(r, \theta, \phi)$ can be estimated as the output of the decoder, with the item expressed as $\hat{\hat{H}}_i^p(r, \theta, \phi)$ and $i = 1, 2, \ldots, 2N_f$.

Then, the individualized HRTFs can be recovered by the inverses of Eqs.(1) and (2), which are expressed as

$$\hat{H}_i(r, \theta, \phi) = \sigma_h(i)\hat{\hat{H}}_i^p(r, \theta, \phi) + \mu_h(i), \tag{15}$$

where $\hat{H}_i(r, \theta, \phi)$ is the individualized HRTF of the i-th frequency bin at the position (r, θ, ϕ). $i = 1, 2, \ldots, 2N_f$. Therefore, the HRTFs for the left ear and the right ear can be expressed as $\hat{\mathbf{H}}_{\mathbf{d}}^L = [\hat{H}_1(r, \theta, \phi), \hat{H}_2(r, \theta, \phi), \ldots, \hat{H}_{N_f}(r, \theta, \phi)]$, and $\hat{\mathbf{H}}_{\mathbf{d}}^R = [\hat{H}_{N_f+1}(r, \theta, \phi), \hat{H}_{N_f+2}(r, \theta, \phi), \ldots, \hat{H}_{2N_f}(r, \theta, \phi)]$. Then, the HRIRs for the left ear and the right ear can be obtained using inverse fast Fourier transform (IFFT), i.e., $\hat{\mathbf{h}}_{\mathbf{d}}^L = \text{IFFT}(\hat{\mathbf{H}}_{\mathbf{d}}^L)$ and $\hat{\mathbf{h}}_{\mathbf{d}}^R = \text{IFFT}(\hat{\mathbf{H}}_{\mathbf{d}}^R)$.

V. REAL-TIME IMPLEMENTATION

A. FINE-GRAINED HRIR INTERPOLATION

As the user or sound sources can move freely in the virtual scene, the key task prior to spatial audio reproduction is to calculate the real-time position of the sound sources relative to the user. Then, continuous HRTFs can be achieved by using fine-tuned individualization model, and converted to HRIRs in the time domain for spatial audio generation. To reduce online computation of continuous HRTFs, we build a database including densely individual HRIRs by sampling in full space offline.

Offline: It is shown in [33] that the minimum audible angle (MAA) averages 5.4° or more in the motion conditions. Motivated by this, we choose the points following the criterion of the distance from 20 cm to 1.2 m at a step of 10 cm, θ from 0° to 355° at a step of 5°, and ϕ from $-30°$ to 90° at a step of 5°. For each discrete point, HRTFs are obtained via the fine-tuned individualization model, and HRIRs are obtained by an inverse Hilbert transform and IFFT followed by adding the interaural time difference (ITD) [34]. As a result, a dense HRIR database is built from a total of 19800 positions.

Online: Then, finer-grained HRIR interpolation is implemented online changing with the relative positions in the virtual scene. If the distance between the sound source and the user is no more than 1.2m, the HRIRs from the nearest point of the current relative position in the dense HRIR database are chosen for binaural synthesis. Otherwise, HRTFs are almost irrelevant to the distance and are functions of (θ, ϕ) in far field [34], calculated as

$$\hat{\mathbf{H}}(r, \theta, \phi) = \frac{1.2}{r}\hat{\mathbf{H}}(1.2, \theta^*, \phi^*), \tag{16}$$

where $(1.2, \theta^*, \phi^*)$ denotes the nearest point to (θ, ϕ) on the sphere of $r = 1.2$ m.

Because of the regular positions in the database, the search operation costs little and HRIRs are obtained over the full space with negligible computational load. Thus, the method is highly efficient and suitable for frequent scene updates.

B. SPATIAL AUDIO REPRODUCTION

Spatial audio is generated by respectively convolving the HRIRs of the left ear and the right ear with audio signals, which is expressed as

$$\mathbf{y}^L = \sum_{n=1}^{N_u} \hat{\mathbf{h}}_{\mathbf{d}_n}^L \otimes \mathbf{s}_n, \tag{17}$$

$$\mathbf{y}^R = \sum_{n=1}^{N_u} \hat{\mathbf{h}}_{\mathbf{d}_n}^R \otimes \mathbf{s}_n, \tag{18}$$

where $\mathbf{y}^L$ and $\mathbf{y}^R$ denote the generated binaural audio for the left ear and the right ear, respectively. $\mathbf{s}_n$ is the signal from the n-th sound source. $n = 1, 2, \ldots, N_u$. N_u is the number of audio sources in the virtual scene. $\mathbf{d}_n$ denotes the relative position between the head of the target subject and the n-th audio source. $\hat{\mathbf{h}}^L_{\mathbf{d}_n}$ and $\hat{\mathbf{h}}^R_{\mathbf{d}_n}$ denote the individual HRIRs for the left ear and right ear relative to the n-th sound source, respectively. $\otimes$ denotes the convolution operation.

A dynamic block convolution method using an overlap and add method in the time domain is exploited to produce binaural audio from multiple sources. First, audio signals are divided into several equal-length blocks according to the updating time of the virtual scene, and we assume that the relative position of the user and the sound source are unchanged in a block. Then, an overlap and add method in the time domain is used for block convolution [35] and the audio outputs are available in real time on a block-by-block basis.

C. PARALLEL SYNTHESIS AND PLAYBACK

After the spatial audio signal is prepared, they should be played back via a headphone in real time to provide an immersive auditory experience when the user freely moves in the virtual scene. To this end, we utilize parallel threads and low-level programming. First, we put the tasks of synthesis and playback into separate threads. Therefore, the spatial audio can be synthesized and played back simultaneously. Then, for each task, we use low-level programming for the audio playback. For this task, we predefine the playback parameters, i.e., sampling frequency, number of bits and number of channels. Then, the audio data are directly played back to the sound card rather than saving and decoding an audio file in such formats as wave and mp3, which significantly reduces the processing time.

VI. PERFORMANCE EVALUATION

In this section, the objective and subjective experiments are conducted to evaluate the performance of our proposed system. The spatial audio rendering system is implemented on the HTC Vive platform with the scene shown in Fig. 5. The size of the virtual space is 10 m × 15 m×5 m. Four sound sources are marked in the virtual scene: a bee, a pot, a television and a fountain. The bee flies freely and the other three sources are fixed in the scene. The user wearing the head-mounted display (HMD) is allowed to move freely within the cube in the center of the room. The size of the movement space is related to the physical room, which is set to 2 m × 2.5 m × 2.7 m for our system. The user tracking was implemented using HTC Vive Lighthouse basestations. First, the influence of the update frequency on the virtual scene is analyzed. The update frequency is set to vary from 5 Hz to 50 Hz at a step of 5 Hz. From the subjective experience, it is observed that if the update frequency is too high, such as above 20 Hz, the virtual scene appears highly unstable. When the update frequency is decreased to 10 Hz, the virtual scene updates become smooth. Since this frequency is related to the virtual vision system, we will not discuss it in this paper.

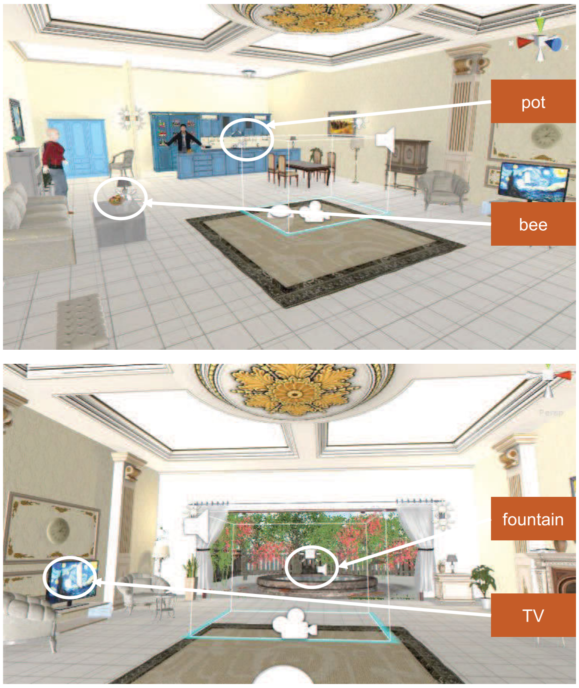


FIGURE 5. The virtual scene on the HTC Vive platform. There are 4 sound sources in the scene: a bee, a pot, a television and a fountain.

A. DATABASE

The databases used for model training are discussed. HRTFs in the PKU&IOA [36] and CIPIC databases [31] are used for pre-trained model training. Anthropometric features of subjects in CIPIC database are used for fine-tuning.

The PKU&IOA database, measured from different distances including near field and far field, allows the extraction of finer information from dense measurements. The database was measured from the KEMAR mannequin at eight distances of 20, 30, 40, 50, 75, 100, 130 and 160 cm, and 793 positions were sampled on each sphere of eight distances. Thus, a total of 6344 HRIR pairs for two ears were obtained. Each HRIR was windowed in approximately 15.625 ms (1024 points) with a sampling frequency of 65.536 kHz.

The CIPIC database contains individual HRTFs from 45 subjects, which allows for individualization modeling. For each subject, HRIRs were measured on a sphere with a radius of 1 m, sampling from 25 azimuth angles and 50 elevation angles. The length of each HRIR is approximately 4.5 ms (200 points) with a sampling frequency of 44.1 kHz. 37 anthropometric features (10 for left pinna, 10 for right pinna, 17 for the head and the torso) were measured for 35 subjects. We choose anthropometric features and the corresponding HRTFs from 30 subjects to train the model, and the remaining data from 5 subjects are used to evaluate the performance of the proposed method as the test database.

B. EXPERIMENTAL SETUP

Prior to modeling, the HRIRs for the two databases are resampled to the same sampling frequency, i.e., 44.1kHz. Then, HRIRs are converted to HRTFs using a 256-point FFT

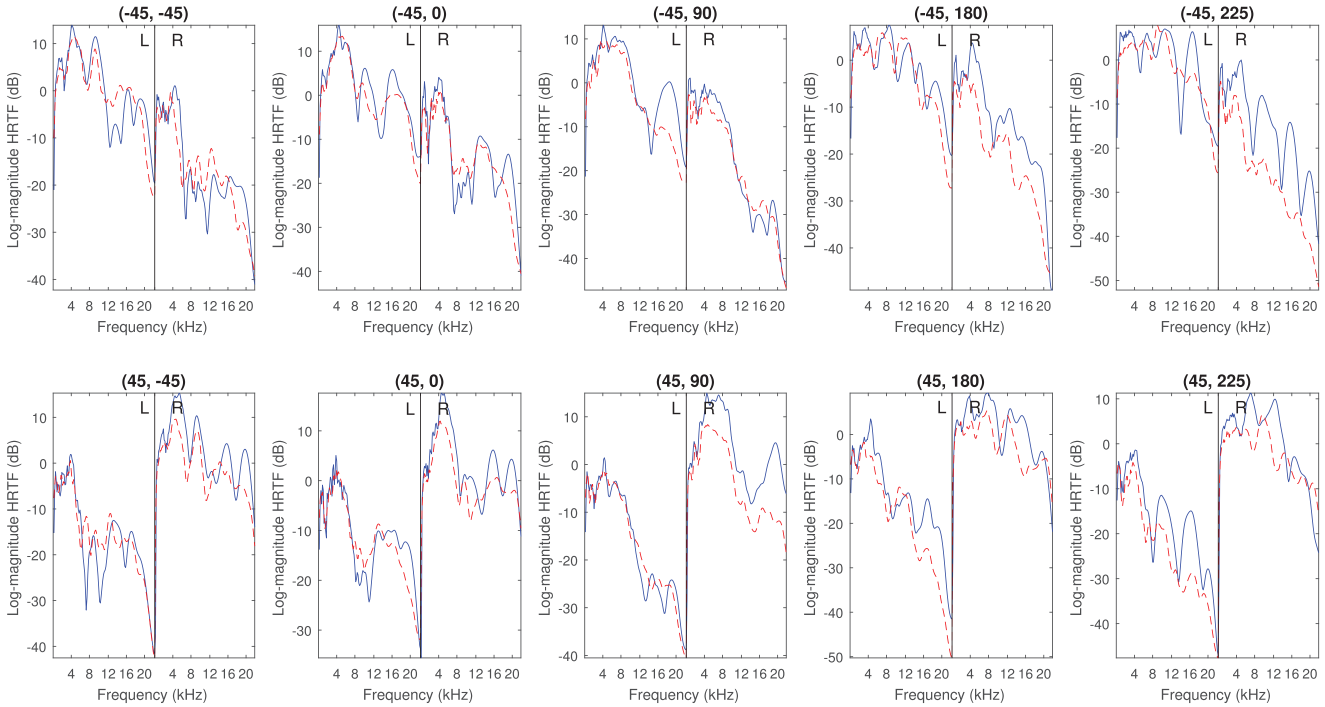


FIGURE 6. The comparison of the ground truth HRTFs (blue solid line) and the predicted HRTFs from our proposed individualization model (red, dash line). The title (x, y) denotes the azimuth angle of x and the elevation angle of y.

followed by constant-Q filtering. We evaluate the frequency bands between 200 Hz and 20 kHz. Together with ITD, there are 233 parameters for each direction.

First, for the autoencoder model, the number of hidden layers is set to 5 with 128 nodes for each layer except the bottleneck layer which has 30 nodes. For the pre-trained model and fine-tuned individualization model, the parameters used to generate SFB basis features are set to $N = 14$, $L = 3$, and $R_m = 220$ cm, resulting in a total of 301 features, and the number of hidden layers is set to 5 with 128 nodes for each layer. We use the *Relu* activation function for the hidden layers, and *linear* activation function for the output layer. The dropout fraction is set to 0.5 and the learning rate is 0.001. Adam is used to optimize the model.

C. OBJECTIVE EVALUATION

In this section, the objective performance of our proposed model is evaluated. First, we analyze the LSD performance of model prediction from the views of position space and frequency space. Then, the experiments are conducted to compare the performance of our method with several existing methods. Three metrics are used for these purposes, including LSD in the frequency domain, the root mean square error (RMSE) and the normalized RMSE (NRMSE) in the time domain.

LSD is defined as log-SD, i.e., LSD $= 20\log_{10}$SD, where SD is calculated as Eq.(6). RMSE describes the root mean square error between the estimated HRIRs $\hat{\mathbf{h}}_\mathbf{d}$ and the measured HRIRs $\mathbf{h}_\mathbf{d}$ in the time domain over all positions $\mathbf{d}$, expressed as

$$\text{RMSE} = \sqrt{\frac{1}{N_d}\sum_{\mathbf{d}} ||\mathbf{h}_\mathbf{d} - \hat{\mathbf{h}}_\mathbf{d}||^2}, \tag{19}$$

where N_d is the number of the positions.

Considering that the measured HRIRs have small values, resulting in no significant difference between MSEs, and the signal energy has little influence on the human's localization perception, we normalize HRIRs, and define NRMSE as the normalized RMSE to improve the efficiency of the evaluation in the time domain. NRMSE can be calculated as

$$\text{NRMSE} = \sqrt{\frac{1}{N_d}\sum_{\mathbf{d}} \left\| \frac{\mathbf{h}_\mathbf{d}}{||\mathbf{h}_\mathbf{d}||} - \frac{\hat{\mathbf{h}}_\mathbf{d}}{||\hat{\mathbf{h}}_\mathbf{d}||} \right\|^2}. \tag{20}$$

First, the comparison results of HRTFs in terms of the ground truth and the predicted HRTFs using our individualization model are analyzed. The results of HRTFs from 10 positions are shown in Fig. 6. There are three observations from the figure as follows. First, it is directed that the predicted HRTFs generally fit to the ground truth, i.e., they have the same trend, which infers the efficiency of the proposed method. Second, in low frequencies, the difference between the ground truth and the predicted HRTFs is small. As the frequency increases, the difference becomes significant. In the high frequencies, there exists pinna notches and peaks, which are hard to match and result in large spectrum distortion. We also illustrate LSD performance under 6 different frequency groups, i.e., 0∼1kHz, 1∼2kHz, 2∼4kHz, 4∼8kHz,

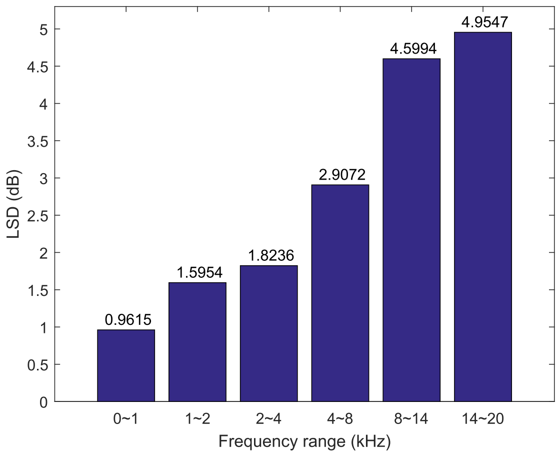
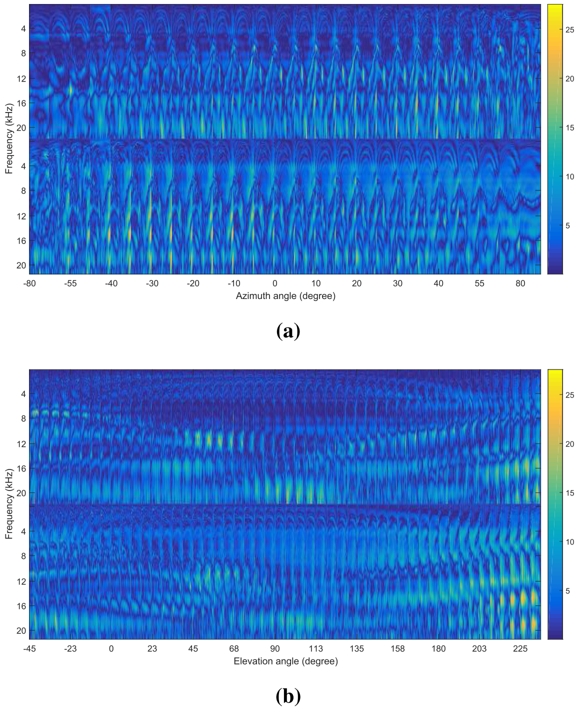


FIGURE 7. The LSD performance under 6 different frequency groups. The LSD values are noted in the figure.



(a)

(b)

FIGURE 8. The LSD performance of HRTFs for different positions. (a) from the view of the azimuth angles. (b) from the view of elevation angles. The upper half and lower half parts denote the LSD of left ear and right ear, respectively.

8∼14kHz and 14∼20kHz, as shown in Fig. 7. It is shown that LSDs in low frequencies, especially under 8kHz, which is the field of human sound, are lower than 3dB, while in high frequencies, LSDs are beyond 4.5dB. Third, from Fig. 6, it is seen that the LSD difference varies from the positions. For example, at the position of (-45, 0), the predicted HRTFs can match the ground truth well, while the difference is significant at the position of (-45, 225). To further analyze the property of LSD in position space, we depict LSDs over all positions in Fig. 8, where the upper half and lower half parts respectively denote the LSDs from left ear and right ear. From Fig. 8a, it is shown that areas with large LSDs mostly locate where the azimuth angle is greater than 0 degree for the left ear, while large LSDs mostly exist where the azimuth angle is smaller than 0 degree for the right ear. That is because when the azimuth angle is greater than 0 degree, the sound source locates on the right of the human, and thus the distance between the source and the left ear is farther. As a result, the source undergoes more paths from reflection, refraction and so on before reaching in the left ear, causing more vibration in HRTFs. The complicated patterns make a hard match and thus large LSDs occur. Furthermore, from Fig. 8b, it is illustrated that large LSDs are mostly at the positions of the elevation angles greater than 210 degree, i.e., at the lower back, where the source is required to bypass the body, shoulder and pinna to reach in the ear, resulting in great attenuation and frequent fluctuation.

Then, we compare the performance of our method with random selection, the methods in [14], [16] and [22] using 5 subjects in the test database. The random selection method randomly chooses a subject's HRTFs from the database as a target subject's HRTFs, where the database is the same as the one used in our model. Therefore, this method generates non-individual HRTFs for the target subject. Meanwhile, other methods output individualized HRTFs. By comparing these methods, the difference of non-individual HRTFs and individualized HRTFs is shown. The overall statistical results are shown in Table 1 in terms of LSD$_L$, LSD, RMSE and NRMSE, where LSD$_L$ denotes LSD in low frequencies from 0 to 8kHz, i.e., human sound frequency band, and LSD means full frequency band. First, it is seen that LSD in low frequency band is much lower than that in full frequency band. It is also observed that the LSD performance of the proposed model outperforms other methods. Especially for random selection method, up to 2dB gains are achieved. Moreover, in terms of RMSE and NRMSE, our method obtains the best performance, inferring that our method can efficiently improve the objective performance of individualization model.

D. SUBJECTIVE EVALUATION

For the subjective evaluation, ten listeners (5 females and 5 males, age 26∼35 years) with normal hearing participated in the experiments. We first conduct experiments to evaluate the subjective performance with only spatial audio involved, and then in the virtual scene with vision and audio involved.

37 anthropometric features, including 10 for each pinna and 17 for the head and the torso, were measured for each listener using the method in [32] via a camera. Combined with the relative position between the user and the sound sources, HRIRs were generated using our proposed individualization model followed by IFFT. Then, the spatial audio signal was produced by convolving the stimulus with the HRIRs, and was playback via a headphone. Prior to the experiments, the subjects underwent procedural training to reduce the influence of procedural factors on the results. The procedural training played binaural audio from 10 different directions with the feedback of the real direction, while in the test phase,

TABLE 1. Objective performance comparison of different methods in terms of LSD$_L$, LSD, RMSE and NRMSE.

Methods	LSD$_L$[dB]	LSD[dB]	RMSE	NRMSE
Random selection	4.21	6.22	0.076	0.84
[16]	3.45	5.61	0.077	0.72
[14]	3.12	4.95	0.072	0.61
[22]	2.94	4.72	0.063	0.51
Proposed model	2.21	4.23	0.060	0.43

TABLE 2. Subjective performance comparison in terms of CR(%), FBR (%), UDR (%) and DR (%).

HRIR data	CR	FBR	UDR	DR
Random selection	40.77	27.12	31.68	14.27
[16]	43.58	25.25	27.37	12.06
[14]	46.67	23.90	26.89	11.31
[22]	48.96	22.81	24.79	10.13
Our model	**51.25**	**21.32**	**22.03**	**9.25**

no feedback was given. During the experiments, replaying was allowed.

First, we sample the positions on the surface of three spheres at the distances of 50, 100, 150cm, with the azimuths from 0° to 350° at a step of 20° and the elevations from −40° to 60° at a step of 20°, and thus, a total of 108 directions are achieved on each sphere. 500 pairs of spatial audio were formed based on these 324 positions. All the testers listened to these spatial audio pair by pair, and recorded the perceptual directions, including the azimuth angle, elevation angle and the relative distance of the paired audio, i.e., which virtual sound source in audio files had a nearer distance or the same distance to themselves. We designed the experiments of distance perception by sorting a pair of distances because of the weak ability of the human to perceive an accurate distance. The overall statistical results are shown in Table 2 in terms of correct ratio (CR), front-back confusion rate (FBR), up-down confusion rate (UDR), and distance confusion rate (DR) which is calculated as the probability of incorrectly relative distance perception. It is seen that our method achieves a total of 10.48% CR gain over random selection method, which significantly improves the perception ability. Besides, the confusion ratios are also reduced to different degrees. Especially, up to 9.65% gains in UDR for our model over other methods indicate that the individualization information, such as pinna shape, can improve the elevation accuracy. Moreover, it is seen that DR is much lower than FBR or UDR, and thus the subjects have the ability to distinguish which source is nearer to them. Furthermore, our proposed method achieves at least 2.29% CR gain and 1.5% reduction of FBR. It is also seen that up-down confusion occurs more frequently, 2%∼4% higher than front-back confusion.

Then, the testers wore HTC Vive and listened to the spatial audio via a headphone in the virtual scene shown in Fig. 5, where multimodal information is involved from vision and auditory. We fixed the update frequency in the scene of HTC Vive to 10 Hz and evaluated the performance of the audio rendering system by allowing the user to freely move in the physical room and listen to the spatial audio from the four sources via headphones. According to the update frequency of the scene, we set the length of the block signal for convolution to 100 ms. As shown in [37], where an assessment methodology is presented to evaluate localization abilities of spatial audio in VR by designing some questionnaires and the performance metrics, we also use the questionnaires and the performance metrics to evaluate the playback fluency and the localization perception accuracy. First, the testers were asked to answer the questions how system fluency and audio latency are using a real number between 0 and 1. A system fluency of 1 indicates that the system updates very smoothly, inferring that playing the spatial audio at the current position does not produce any noticeable latency. A score of 0 indicates the opposite. An audio latency of 1 means that the latency is severe and playing the spatial audio produces great latency, while a score of 0 indicates there is no noticeable latency. Then, the testers were required to write the positions of the spatial audio in the virtual scene where they perceived, including the azimuth angle and elevation angle to evaluate their localization accuracy.

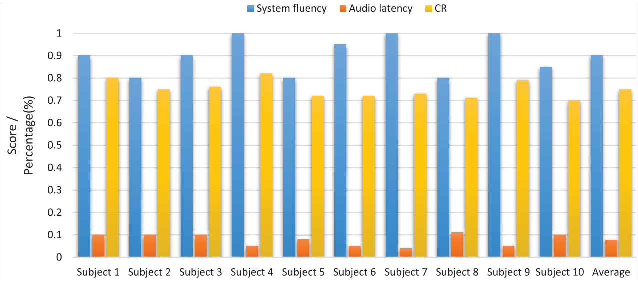


FIGURE 9. The system performance for 10 subjects and the average of all subjects in terms of system fluency, audio latency and localization accuracy.

The results of subjective experiments are shown in Fig. 9. We compare the recorded positions with the ground-truth positions, and calculate the CR of localization perception. From Fig. 9, it is shown that the system update remains smooth after the spatial audio system is embedded in the scene, suggesting that the audio is updated with the relative positions between the user and the sound sources in real time. Occasionally, scene discontinuity occurs due to unreliable data transmission. Then, when the user and the sound sources move freely in the scene, the average latency score is 0.078. The slight latency comes from two aspects. First, the head moves in a small segmented block, resulting in a change in the relative position of the user and the sound sources, and we assume the HRTFs are constant within a block. However, since the block duration is short, the resulting perceptible latency is much lower than the update frequency, and it can be negligible. Second, dislocation can occur in spatial sound perception in dynamic environments due to the finite sound speed, causing the perceptive position to be later than the current position with a delay of $\Delta t = r/c$. In our system, since the scene is limited to a virtual room, the influence on the perception can be neglected. Finally, the localization accuracy is tested by recognizing the positions of the sound sources and tracking the flying bee during the movement of the user and/or sound sources. The results show that an average CR of 75% is achieved, indicating that the rendering system on the HTC Vive can provide a reliable hearing experience. Moreover, the experimental results show that the performance of localization perception is significantly better than that of only spatial audio. The great gain comes from intra-multimodal information enhancement; i.e., vision has a positive influence on the localization perception of the sound source. Therefore, the user receives a more immersive VR experience.

VII. DISCUSSION AND CONCLUSION

In this paper, a real-time spatial audio rendering system on wearable device is designed. Since HRTF individualization is one of key factors in improving the externalization of spatial audio for accurate localization, a pre-trained-based individualization model is proposed to generate HRTFs for a target user. We first build a user-independent pre-trained model between position features and HRTFs using PKU&IOA and CIPIC databases, and then obtain individualization model by fine tuning the pre-trained model using anthropometric features. By utilizing multiple databases, more property of HRTFs can be learned and the model is potential to improve the quality of the spatial audio and achieve better accurate localization perception.

The performance of the proposed method has been evaluated using objective and subjective experiments. First, we analyze the LSD performance in position space and frequency space. The results infer that the better localization perception performance occurs in the field of the front of the user and the low frequency band especially under 8kHz. Further comparison experimental results show that our model outperforms other existing method in terms of LSD, RMSE and NRMSE, where our model achieves up to 2dB gains of LSD, indicating the efficiency of the proposed pre-trained individualization model.

We also designed the subjective experiments to evaluate the accuracy of the localization perception for 10 testers with normal hearing and the results show that our model can achieve better performance in terms of CR, UDR, FBR and DR. Finally, the real-time spatial audio system has been implemented on the platform of HTC Vive, and the subjects report that they can obtain a fluent experience when moving in the virtual scene, and the localization perception is more accurate with an average CR of 75%, implying that the vision in the virtual scene can improve the localization of the audio.

However, there are some limitations in our study. First, the experimental results show that fast attenuation and frequent oscillation appear in high frequency band of HRTFs, resulting in large LSDs, and thus significant LSD differences exist between low frequency band and high frequency band. Then, besides individualized HRTFs, room transfer response, such as reverberation and the Doppler effect from relatively moving between the users and the sound sources, can be beneficial for localization perception [34]. These are not considered in our system.

Planned future work can be carried out from the following aspects. First, we will study reverberation and the Doppler effect to introduce the simulation of the effects of the room response and the speed of motion into the system for more vivid perception. Second, based on perception sensitivity among different frequency bands, optimal weights for different frequency bins should be studied to reduce LSD, especially in high frequency band. Third, we will expand more databases by collecting or measuring to train a more powerful pre-trained model. Finally, more network structures, such as generative adversarial networks (GAN) or graph neural network (GNN) will be investigated to improve the performance of our system.

REFERENCES

[1] J. M. Loomis, R. G. Golledge, and R. L. Klatzky, "Navigation system for the blind: Auditory display modes and guidance," *Presence, Teleoperators Virtual Environ.*, vol. 7, no. 2, pp. 193–203, Apr. 1998.

[2] R. Y. Litovsky, A. Parkinson, and J. Arcaroli, "Spatial hearing and speech intelligibility in bilateral cochlear implant users," *Ear Hearing*, vol. 30, no. 4, p. 419, 2009.

[3] S.-N. Yao, "Headphone-based immersive audio for virtual reality headsets," *IEEE Trans. Consum. Electron.*, vol. 63, no. 3, pp. 300–308, Aug. 2017.

[4] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," *J. Audio Eng. Soc.*, vol. 49, no. 4, pp. 231–249, 2001.

[5] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer-functions," *J. Acoust. Soc. Amer.*, vol. 94, no. 1, pp. 111–123, 1993.

[6] H. Moller, "Fundamentals of binaural technology," *Appl. Acoust.*, vol. 36, nos. 3–4, pp. 171–218, 1992.

[7] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a spherical-head model from anthropometry," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 472–479, 2001.

[8] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *J. Acoust. Soc. Amer.*, vol. 112, no. 5, pp. 2053–2064, Nov. 2002.

[9] M. Geronazzo, S. Spagnol, and F. Avanzini, "Mixed structural modeling of head-related transfer functions for customized binaural audio delivery," in *Proc. 18th Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2013, pp. 1–8.

[10] Y. Kahana and P. A. Nelson, "Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models," *J. Sound Vibrat.*, vol. 300, nos. 3–5, pp. 552–579, Mar. 2007.

[11] H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura, and K. Iida, "Mechanism for generating peaks and notches of head-related transfer functions in the median plane," *J. Acoust. Soc. Amer.*, vol. 132, no. 6, pp. 3832–3841, Dec. 2012.

[12] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "A priori mesh grading for the numerical calculation of the head-related transfer functions," *Appl. Acoust.*, vol. 114, pp. 99–110, Dec. 2016.

[13] B.-S. Xie, "Recovery of individual head-related transfer functions from a small set of measurements," *J. Acoust. Soc. Amer.*, vol. 132, no. 1, pp. 282–294, 2012.

[14] X. Qi and L. Wang, "Parameter-transfer learning for low-resource individualization of head-related transfer functions," in *Proc. Interspeech*, Sep. 2019, pp. 3865–3869.

[15] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4468–4472.

[16] J. He, W.-S. Gan, and E.-L. Tan, "On the preprocessing and postprocessing of HRTF individualization based on sparse representation of anthropometric features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 639–643.

[17] X. Qi and J. Tao, "Sparsity-constrained weight mapping for head-related transfer functions individualization from anthropometric features," in *Proc. Interspeech*, Sep. 2018, pp. 841–845.

[18] S. Shekarchi, J. Christensen-Dalsgaard, and J. Hallam, "A spatial compression technique for head-related transfer function interpolation and complexity estimation," *J. Acoust. Soc. Amer.*, vol. 137, no. 1, pp. 350–361, Jan. 2015.

[19] H. Hu, L. Zhou, H. Ma, and Z. Wu, "HRTF personalization based on artificial neural network in individual virtual auditory space," *Appl. Acoust.*, vol. 69, no. 2, pp. 163–172, Feb. 2008.

[20] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein, "A manifold learning approach for personalizing HRTFs from anthropometric features," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 3, pp. 559–570, Mar. 2016.

[21] Q. H. Huang and Q. L. Zhuang, "HRIR personalisation using support vector regression in independent feature space," *Electron. Lett.*, vol. 45, no. 19, p. 1002, 2009.

[22] R. Miccini and S. Spagnol, "HRTF individualization using deep learning," in *Proc. IEEE Conf. Virtual Reality 3D User Interface Abstr. Workshops (VRW)*, Mar. 2020, pp. 390–395.

[23] T.-Y. Chen, T.-H. Kuo, and T.-S. Chi, "Autoencoding HRTFS for DNN based HRTF personalization using anthropometric features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 271–275.

[24] E. M. Wenzel, "Effect of increasing system latency on localization of virtual sounds," in *Proc. Audio Eng. Soc. Conf., 16th Int. Conf., Spatial Sound Reproduction*. New York, NY, USA: Audio Engineering Society, Mar. 1999, pp. 1–9.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[26] A. Kulkarni, S. K. Isabelle, and H. S. Colburn, "Sensitivity of human subjects to head-related transfer-function phase spectra," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2821–2840, May 1999.

[27] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient real spherical harmonic representation of head-related transfer functions," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 921–930, Aug. 2015.

[28] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1637–1647, 1992.

[29] K. J. Fink and L. Ray, "Tuning principal component weights to individualize HRTFs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 389–392.

[30] Q. Wang, O. Ronneberger, and H. Burkhardt, "Rotational invariance based on Fourier analysis in polar and spherical coordinates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1715–1722, Sep. 2009.

[31] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2001, pp. 99–102.

[32] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Rendering localized spatial audio in a virtual auditory space," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 553–564, Aug. 2004.

[33] W. O. Brimijoin and M. A. Akeroyd, "The moving minimum audible angle is smaller during self motion than during source motion," *Frontiers Neurosci.*, vol. 8, p. 273, Sep. 2014.

[34] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display*. Lauderdale, FL, USA: J. Ross Publishing, 2013.

[35] Y. Song, X. Wang, C. Yang, G. Gao, W. Chen, and W. Tu, "Frame-independent and parallel method for 3D audio real-time rendering on mobile devices," in *Proc. Int. Conf. Multimedia Modeling*, 2017, pp. 221–232.

[36] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, "Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap," *IEEE Trans. Audio Speech Language Process.*, vol. 17, no. 6, pp. 1124–1132, Aug. 2009.

[37] A. N. Moraes, R. Flynn, A. Hines, and N. Murray, "Evaluating the user in a sound localisation task in a virtual reality application," in *Proc. 12th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May 2020, pp. 1–6.

JINYAN LU received the B.S. degree in computer science and technology from Xidian University, Xi'an, China, in 2008, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2016. She is currently with Henan University of Engineering, Zhengzhou, China. Her current research interests include intelligent control, signal processing, and machine learning.

XIAOKE QI received the B.S. degree in communication engineering from Nankai University, Tianjin, China, in 2009, and the Ph.D. degree in signal and information processing from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, in 2014. She is currently with China University of Political Science and Law, Beijing. Her current research interests include multimedia, natural language processing, machine learning, and wireless communication.