IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 July 2023, accepted 14 August 2023, date of publication 24 August 2023, date of current version 8 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3308143

RESEARCH ARTICLE

PRTFNet: HRTF Individualization for Accurate Spectral Cues Using a Compact PRTF

BYEONG-YUN KO, GYEONG-TAE LEE, HYEONUK NAM, (Member, IEEE), AND YONG-HWA PARK, (Member, IEEE)

Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea

Corresponding author: Yong-Hwa Park (yhpark@kaist.ac.kr)

This work was supported by the BK21 FOUR Program of the National Research Foundation Korea (NRF) grant funded by the Ministry of Education (MOE).

ABSTRACT Spatial audio rendering relies on accurate localization perception, which requires individual head-related transfer functions (HRTFs). Previous methods based on deep neural networks (DNNs) for predicting HRTF magnitude spectra from pinna images used HRTF log-magnitude as the network output during the training stage. However, HRTFs encompass the acoustical characteristics of the head and torso, making it challenging to reconstruct the spectral cues necessary for elevation localization. To tackle this issue, we propose PRTFNet to reconstruct the individual spectral cues in HRTFs by mitigating the influence of the head and torso. PRTFNet consists of an end-to-end convolutional neural network (CNN) model and leverages a compact pinna-related transfer function (PRTF) that eliminates the impact of sound reflections from the head and torso in the head-related impulse response (HRIR) as network output. Additionally, we introduce HRTF phase personalization, a technique that utilizes the phase spectra of a selected HRTFs from a database and adjusts the phase by multiplying it by the ratio of the target listener's head width to that of the subject of the selected HRTFs. We evaluated the proposed HRTF individualization methods using the HUTUBS dataset, and the results demonstrate that PRTFNet is highly effective in reconstructing the first and second spectral cues. In terms of log spectral distortion (LSD) and effective LSD (LSD_E), PRTFNet outperforms previous deep learning-based model. Furthermore, multiplying the selected phase by the head width ratio reduces the root mean square error (RMSE) of interaural time difference (ITD) by 0.003 ms.

INDEX TERMS Head-related transfer functions, individualization, pinna-related transfer functions, spectral cues, spatial hearing.

I. INTRODUCTION

Spatial audio rendering is a crucial technique used to replicate the human perception of spatial audio scenes through headphone or loudspeaker systems. With the advent of metaverse technologies like virtual reality (VR) and augmented reality (AR), the demand for spatial audio rendering has increased to provide a natural and immersive auditory experience for users [1]. Furthermore, spatial audio rendering finds applications in various fields, including multimedia healthcare [2], entertainment audio industry [3], and blind assistance [4], emphasizing its significance beyond virtual environments.

To simulate spatial audio scenes, head-related transfer function (HRTF) should be applied to the sound source signal. HRTF represents a frequency response that describes the transmission of sound from an arbitrary direction to the ear [5]. As HRTF varies significantly among individuals, using a non-individualized HRTF for spatial audio synthesis can lead to perception errors such as front-back confusion, the perception of a sound image rising, and localization inside the head [6]. The temporal and spectral characteristics of the HRTF are influenced by interactions between the sound wave and the torso, head, and pinna. These characteristics provide localization cues for sound azimuth and elevation to the human auditory system. For instance, the interaural time difference (ITD) and the interaural level difference (ILD) are

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

important cues for azimuth localization [7]. In terms of elevation localization, the ear pinna generates distinctive spectral patterns in the HRTF, including main peaks and notches, known as spectral cues [8]. The dorsal cochlear nucleus in the auditory brainstem uses these spectral cues to detect the elevation of a sound source [9]. Consequently, individualized HRTFs, incorporating accurate ITD, ILD, and spectral cues, are essential for a precise spatial hearing experience.

Obtaining individual HRTFs poses several challenges. Firstly, the determination of hundreds of frequency bins in the HRTF relies on various acoustical effects such as reflection, refraction, and diffraction from different body parts. Accurate 3D geometry data for curved non-convex objects (e.g., the ear pinna) and substantial computational resources are necessary to identify these acoustical effects. Secondly, spectral cues play a crucial role in perceiving sound elevation, but their distribution in the HRTF is highly sensitive to sound direction. Moreover, spectral cues predominantly exist in the high-frequency range, making it difficult to analyze the frequency response using mechanical models. The characteristics of spectral cues and the complex structures near the ear canal complicate the prediction of spectral cues for arbitrary pinna shapes.

In this context, several methods have been proposed to obtain individual HRTFs. One approach involves acoustical measurement, which has been suggested [10]. However, this method can be costly in terms of equipment and time-consuming when measuring individual HRTFs for all directions. Alternatively, a method has been proposed that involves scanning the head shape and estimating the corresponding individual HRTF through numerical simulation techniques such as the finite element method, boundary element method, and finite difference time domain method [11], [12], [13]. These methods have shown promising results, particularly at low frequencies, where the simulated HRTFs closely resemble the measured ones. However, the complex frequency response of the ear pinna in the higher frequency range has posed challenges for these methods. They tend to introduce alterations in the spectral cues beyond 4 kHz, which is the frequency range crucial for perceiving the vertical position [8]. Moreover, implementing this method requires bulky and expensive 3D scanning equipment like a 3D laser scanner or MRI. Additionally, intensive computation is necessary, especially for accurately capturing the high-frequency range.

To analyze the acoustical effects of body parts and estimate HRTF, researchers have proposed simplified physical models of the pinna [14] and head-and-torso [15], as well as structural models [16]. These models aim to provide insights into the underlying physical mechanisms of HRTF. However, despite the advantages of these simple acoustical models, such as requiring only a small set of anthropometric parameters and low computational load, accurately estimating HRTF in the high-frequency range, particularly capturing spectral cues for diverse pinna shapes, remains challenging. Furthermore, the specific anthropometric parameters needed to characterize the pinna model have not been explicitly defined yet.

On the other hand, researchers have explored various data-driven approaches to obtain individualized HRTFs based on anthropometric features of the ear pinna, head, and torso. These approaches leverage the strong correlation observed between HRTFs and anthropometric features. Techniques such as multiple linear regression [17], support vector regression [18], artificial neural network (ANN) [19], and deep neural network (DNN) [20] have been employed to establish the relationship between anthropometric features and HRTF. However, one limitation of anthropometric-based methods is the requirement for measuring anthropometric data. The process of obtaining individual anthropometric measurements can be inconvenient, time-consuming, and prone to variation, especially in the case of measuring ear pinna features [21]. To address this issue, some studies have proposed alternative methods. For instance, instead of directly measuring anthropometric features, researchers have manually marked landmark points on a single pinna image and calculated the distances between these landmarks to obtain pinna anthropometric features [5], [22]. Furthermore, [23] have designed a U-Net model to automatically extract the positions of pinna landmarks from a pinna image, eliminating the need for manual annotation.

With the advancements in deep learning techniques for image processing, such as pattern recognition and image classification, DNN based HRTF individualization using pinna images have been proposed as practical solutions, instead of relying solely on pinna anthropometric features. In previous work, Lee and Kim [24] utilized both a pinna image and anthropometric features of the head and torso as inputs to predict individual HRTFs. More recently, based on experimental findings highlighting the role of ear pinna in generating spectral cues of HRTFs [25], DNN architectures that generate the magnitude spectra of individual HRTFs using only pinna images have been proposed [26], [27]. These DNNs typically employ an autoencoder structure, known for efficient dimensionality reduction of HRTFs [28]. They consist of three sub-networks that convert pinna images through latent variables to HRTF magnitude. However, one limitation of these methods is that the sub-networks are trained separately, which prevents the simultaneous optimization of the entire process for synthesizing HRTF magnitude from pinna images. Additionally, using full-spherical HRTFs with hundreds of frequency bins in each sound direction as the network output can lead to overfitting issues. As a consequence, these problems can result in a loss of important spectral cues within the HRTF.

This study introduces a novel end-to-end convolutional neural network (CNN) model called *PRTFNet* to predict individual HRTFs from pinna images while ensuring the preservation of accurate spectral cues. Previous research has shown that listeners can successfully localize sound elevation using HRTFs that only consist of the main peaks and notches

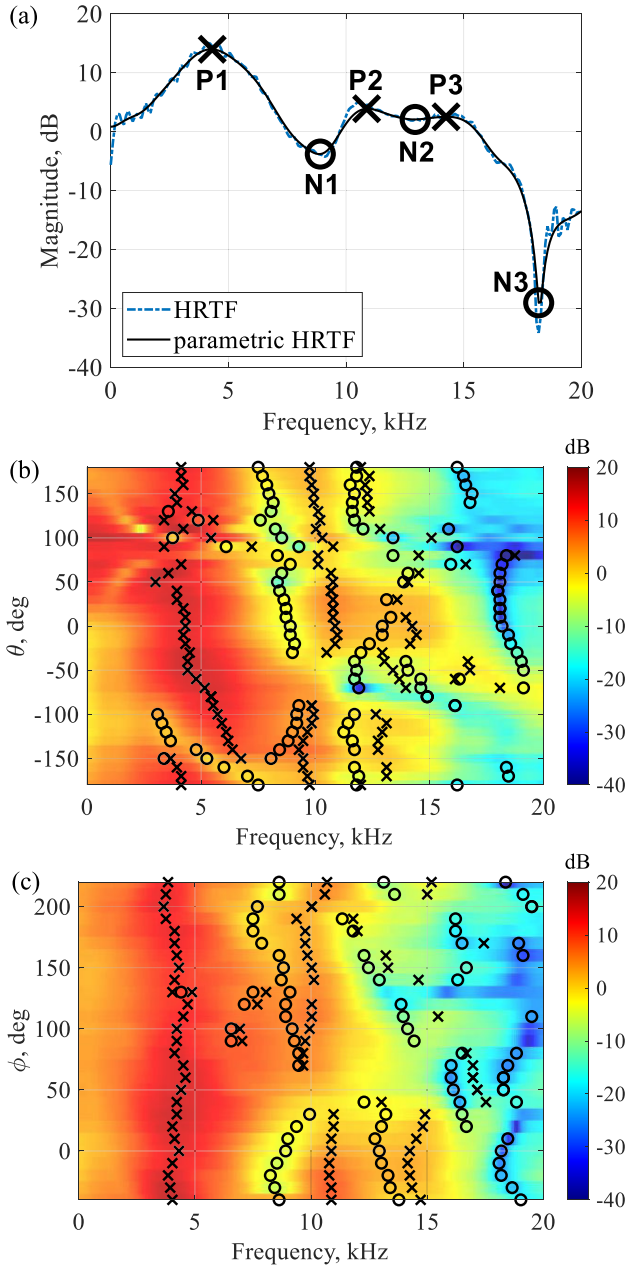


FIGURE 1. Example of the distribution of the 3 lowest-frequency spectral cues in the left ear's Head-Related Transfer Functions (HRTFs) of the B&K Head-and-Torso Simulator Type 4100: (a) in the HRTF for 0° azimuth and 0° elevation; (b) in the horizontal plane HRTFs depending on sound azimuth (−180° ∼180°); (c) in the median plane HRTFs depending on sound elevation (−40° ∼220°). Here, the symbol legend means: X, prominent peak; O, prominent notch.

from measured HRTFs [29]. Hence, to facilitate the learning of spectral cue patterns by the neural network, we eliminate the fine spectral features in the HRTF (network output) caused by sound reflection effects from the head and torso during the training stage. To achieve this, PRTFNet is trained using a compact pinna-related transfer function (PRTF) that is extracted from measured HRTFs, primarily focusing on preserving the spectral cues. This approach allows us to reduce the number of frequency bins in the HRTF without sacrificing crucial spectral cue information. Furthermore, we propose a technique for HRTF phase personalization based solely on the measurement of the head width. This approach involves utilizing the head width of the target listener to select an appropriate HRTF from a pre-existing HRTF database. We then adjust the phase of the selected HRTF by multiplying it with the ratio of the listener's head width to that of the subject of the selected HRTFs. To validate the reconstruction of individual spectral cues and the personalization of HRTF phase, we employ the recently released HUTUBS HRTF dataset (2019) [30], which includes 3D head and ear scans, anthropometry features, and measured HRTFs.

The structure of the remainder of the paper is as follows. In Section II, we provide an explanation of the roles and properties of spectral cues. Sections III–V offer an overview of the HRTF individualization scheme and provides detailed insight into the processes involved in PRTFNet and HRTF phase personalization, respectively. We conduct experimental validation in Sections VI and VII, and present a comprehensive analysis and discussion of the experimental results. Finally, in Section VIII, we draw our conclusions regarding the efficacy and performance of the HRTF individualization methods.

II. SPECTRAL CUES IN HRTF

The magnitude spectrum of HRTF plays a crucial role in determining sound elevation localization. Research has demonstrated that the overall shape of the HRTF magnitude is more significant than the fine spectral details when it comes to localizing sound elevation [31]. Specifically, spectral cues above 5 kHz are responsible for the perception of sound source elevation [32] and the frequency components of HRTF above 16 kHz and below 3.8 kHz do not affect sound elevation localization [8]. In experiments conducted to validate the importance of spectral cues in elevation localization [29], a parametric HRTF was synthesized using only the spectral cues from measured HRTF, as depicted in Fig. 1(a). Subjects of the experiments accurately detected the elevation of the sound source when they listened to the synthesized spatial audio.

The variation of spectral cues in the median plane is relatively more significant compared to those in the horizontal plane across sound source directions [33]. Therefore, listeners can localize sound elevation by the frequency and magnitude of spectral cues. As the sound source moves from the front of the listener to above their head, the frequency of the main notches in the spectral cues shifts higher [32]. In our experiments conducted to measure HRTF [34], we analyzed the left ear's HRTFs of the B&K head-and-torso simulator (HATS) Type 4100 for both the horizontal and median planes, as depicted in Fig. 1(b) and (c) respectively. Our analysis revealed that not only the frequency of notches but also the peaks above 4 kHz undergo changes according to the sound elevation.

The frequency and magnitude of spectral cues display significant variations depending on the shape of the pinna [35]. This indicates that the distribution of spectral cues is unique

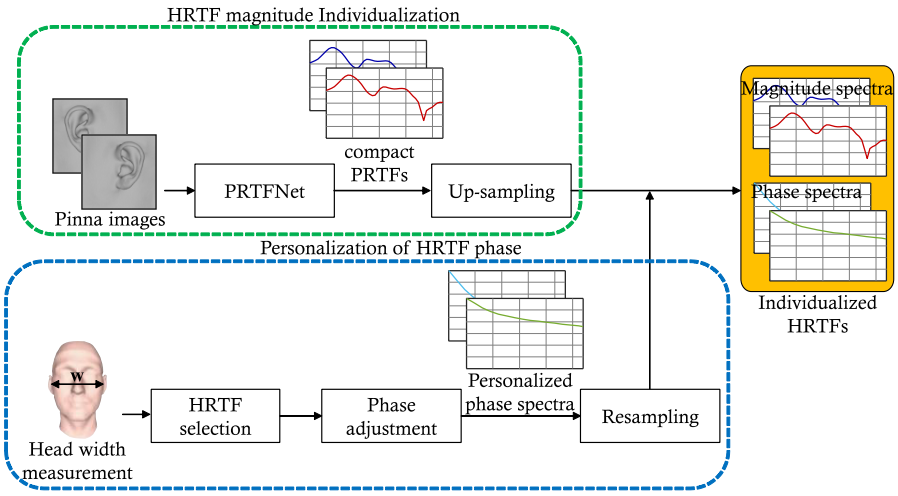


FIGURE 2. Overview of HRTF individualization scheme for synthesizing the magnitude and phase spectra of individualized HRTF, with its inputs, outputs, and constituting elements.

to each individual, resulting in a high level of individual dependence. Given the distinctive characteristics of spectral cues, such as their sensitivity to sound source direction and individual-specific nature, the DNN model to predict individual spectral cues should possess two key characteristics:

- Extraction of the intricate patterns of spectral cues in the HRTF corresponding to the direction of the sound source.
- Capturing the relationship between an individual's pinna shape and the associated spectral cues.

Achieving precise sound elevation localization necessitates the DNN model in HRTF individualization scheme.

III. HRTF INDIVIDUALIZATION SCHEME

The goal of the HRTF individualization scheme is to synthesize the magnitude and phase spectra of individual HRTFs using either DNN or signal processing techniques. The synthesized HRTF magnitude provides the listener with spatial perception of sound elevation, while the ILD and ITD of the HRTF contribute to the perception of sound azimuth. The previous HRTF individualization scheme relied on pinna images for synthesizing HRTF magnitude using DNN and required anthropometric features of the head and shoulders to estimate the phase spectrum of the HRTF through multiple regression analysis. Unfortunately, the spectral cues in the HRTF obtained from the previous individualization scheme were prone to distortion, potentially leading to errors in sound elevation localization. Moreover, measuring the multiple anthropometric features is inconvenient and time-consuming. To address these issues, we have designed a novel HRTF individualization scheme, and an overview of the entire process is presented in Fig. 2. The individualization scheme primarily consists of:

- Our proposed end-to-end CNN, named PRTFNet, which accurately predicts the magnitude spectrum of individual HRTFs with precise spectral cues.
- A signal processing module that generates the phase spectrum of the HRTF solely based on the measurement of head width.

To obtain the individualized HRTFs for both ears of listener, only the pinna images and the measurement of head width are required as inputs for the neural network and the signal processing module, respectively. Initially, PRTFNet takes grayscale image of the listener's pinna as input to synthesize the magnitude spectrum of the HRTF for each ear. The synthesized HRTF magnitude, containing the spectral information of spectral cues primarily in a reduced number of frequency bins, is up-sampled to match the frequency resolution of the sound used for spatial audio rendering. After the up-sampling process, HRTFs that closely correspond to the listener's head width are selected from an HRTF database. The subsequent step involves adjusting the phase spectra of the chosen HRTFs to ensure accurate lateral perception. The phase spectra are resampled to align with the frequency resolution of the sound. Combining the phase spectra with the synthesized HRTF magnitude spectra yields the individualized HRTFs for both ears. These individualized HRTFs are then convolved with the sound to generate spatial audio, creating an immersive auditory experience tailored to the listener's unique hearing characteristics.

IV. PRTFNET

This study presents the PRTFNet, a neural network module used in the HRTF individualization scheme for synthesizing the magnitude spectrum of HRTF with pinna features. The main objective of PRTFNet is to accurately reconstruct spectral cues in HRTF. PRTFNet is composed of an end-to-end CNN structure, and its training procedure consists of three steps. Firstly, head-related impulse responses (HRIRs) are clipped using a window function to eliminate the effects of sound reflections from the head and torso. Secondly, zero-valued samples are removed from the windowed HRIRs, and the HRIRs are transformed into compact PRTFs using Fast Fourier Transform (FFT). Lastly, the end-to-end CNN model

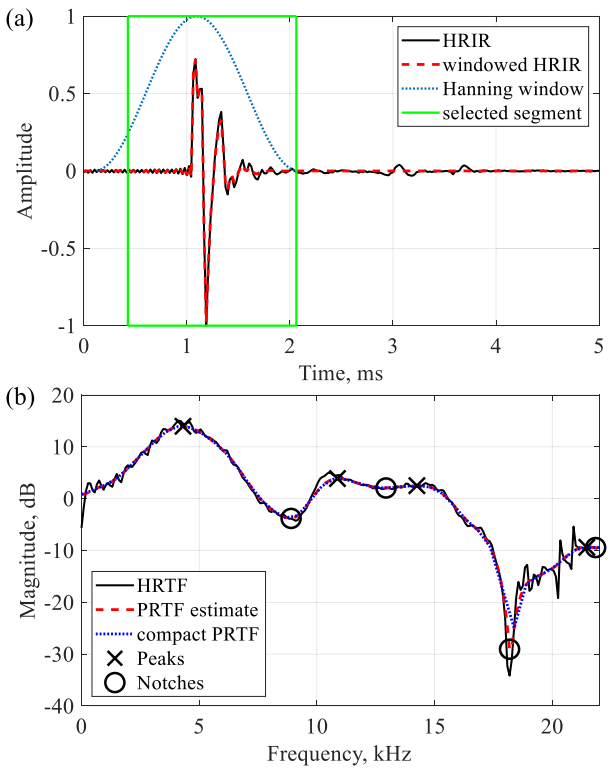


FIGURE 3. Comparison of HRIRs, and HRTFs of artificial head-torso simulator (B&K HATS Type 4100) at azimuth 0° and elevation 0°. (a) HRIRs, (b) HRTFs.

is trained for each specific direction. The training process involves utilizing pinna images and one hot encoding to represent the direction index as inputs to the network, while the corresponding compact PRTF for that direction serves as the network output. For further clarification, a detailed explanation of the PRTFNet procedure is provided below.

A. WINDOWING HRIR

The perception of sound elevation in spatial audio by listeners is closely related to the magnitude spectrum of HRTF, particularly the spectral cues [8]. Research [36] has also demonstrated that the human auditory system perceives HRTF magnitude on a logarithmic scale through localization tests. As a result, HRTF individualization should aim to accurately reconstruct the logarithmic-scale spectral cues of individual HRTF. Previous DNN models used pinna images as network input and log-scale magnitude of HRTF as network output. However, it is important to note that HRTF is influenced not only by the acoustical characteristics of the ear pinna but also by those of the head and torso. Information unavailable from the input can affect the output, thereby compromising the correlation between the individual pinna image input and the estimated individual HRTF output. Consequently, inaccuracies arise in the HRTF estimation process.

To demonstrate the lack of correlation between input and output, we utilized HRIR data obtained from the B&K HATS Type 4100, which was measured using a one-way speaker system in an anechoic chamber [34]. In Fig. 3, it can be observed that when HRIRs are transformed into HRTFs in the frequency domain using FFT, small harmonics (represented by slight fluctuations along the HRTF outline) are present across the entire frequency range. These harmonics stem from reflections occurring in the head and torso [37]. However, training a neural network to predict these spectral components from input data (such as pinna images) that lack any information about the head and torso becomes unfeasible. Moreover, the presence of these harmonics can lead to overfitting of the network to local details, as observed in prior studies that employed HRTF as the output. This overfitting hinders the network's ability to learn the primary pattern of HRTF, namely the spectral cues.

The spectral cues of HRTF, as depicted in Fig. 3(b), exhibit prominent peaks at 4.3 kHz, 10.9 kHz, 14.3 kHz, and 21.5 kHz, along with notches at 8.9 kHz, 12.9 kHz, 18.2 kHz, and 21.8 kHz. By removing the sound reflection effects from the head and torso, it becomes possible to extract these spectral cues. The reflections originating from the head and torso typically arrive more than 1 ms after the direct sound [38]. To eliminate the influence of head and torso effects, a Hanning window with a length of 2 ms is applied, aligning its center with the position of the maximum HRIR amplitude (which corresponds to the arrival time of the direct sound, as shown in Fig. 3(a)). The FFT of the windowed HRIR primarily reflects the acoustical effects of the ear pinna, effectively excluding the impact of reflections stemming from the torso and head. Consequently, it provides an estimation of the PRTF [39]. In Fig. 3(b), the estimated PRTF is represented by a red dashed line. Notably, the small harmonic components caused by the head and torso vanish, while the overall shape of the spectral cues remains closely aligned with that of the HRTF.

B. EXTRACTION OF COMPACT PRTF

The estimated PRTF mentioned above comprises hundreds of frequency bins, representing the number of spectral components should be predicted by individualization of HRTF magnitude. However, datasets used for HRTF individualization, such as CIPIC [40], ITA [41], and HUTUBS [30], contain a relatively small number of samples compared to the number of network parameters [26], [27]. Consequently, if the network model is trained using these datasets' PRTF estimates and pinna images, it may suffer from overfitting due to the vast number of network output points and the limited training samples available [42], [43]. Moreover, the count of frequency bins containing spectral cues, as well as the neighboring bins, is low, while the remaining frequency bins have a significantly higher count. During the training process, the network model aims to minimize the loss function, which computes the average distance between the true and predicted spectra across all frequency bins. Consequently, the trained model is less likely to prioritize accurate estimation of the spectral cues due to the relatively smaller number of adjacent frequency bins.

We propose a method to reduce the number of frequency bins in the PRTF estimate while preserving the spectral cues

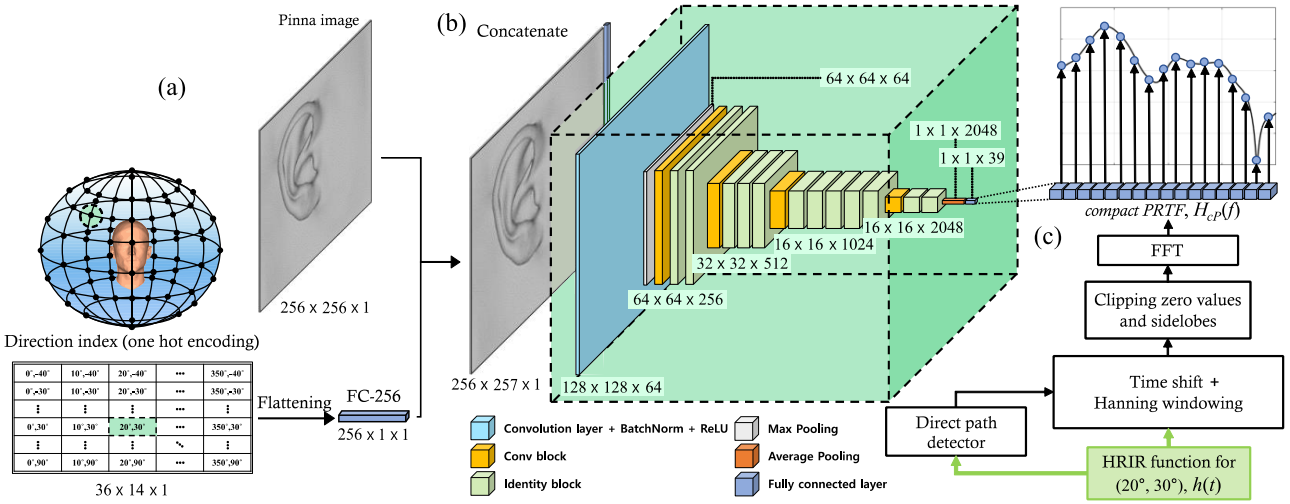


FIGURE 4. Example of training PRTFNet at azimuth 20° and elevation 30°. (a) Inputs of PRTFNet are pinna image and one hot encoding for direction index. (b) PRTFNet is trained to predict compact PRTF from the concatenated input. (c) HRIR function is converted to compact PRTF.

by removing unnecessary time domain samples in the HRIR before applying the FFT. In Fig. 3(a), the windowed HRIR function displays zero-values outside the windowed range. According to the zero-padding theorem [44], zero-padding in the time domain acts as frequency domain data interpolation. By clipping the zero values from the windowed HRIR, we can effectively decrease the number of frequency bins in the resulting PRTF estimate while retaining the dominant spectral pattern (spectral cues). The direct sound in the HRIR, represented by a sinc function, displays sidelobes when broad band noise with a frequency range equal to half of the sampling frequency, f_s is employed as the sound source for HRTF measurement [34]. The passband ripple of the truncated sinc function remains below 1 dB when the length of the sinc function exceeds $64/f_s$ [45]. Thus, we can remove the left sidelobes located more than $32/f_s$ away from the maximum peak, where the pinna effect is not present. The selected segment after these processes is depicted as a solid green line in Fig. 3(a), while the FFT of the windowed HRIR segment is represented by a dotted blue line in Fig. 3(b). We refer to this spectrum as the *compact PRTF*. It is evident that the dominant pattern of the spectrum closely resembles that of the PRTF estimate in section A while effectively reducing the number of frequency bins.

C. NEURAL NETWORK STRUCTURE AND DIRECTION-WISE TRAINING

In previous works [26], [27], individual HRTF synthesis with individual pinna images employed three sub-networks: the variational autoencoder (VAE), fully connected (FC) layers, and conditional VAE (CVAE). While VAE and CVAE models are commonly used for data reconstruction [28], multi-step learning methods can be inefficient and yield suboptimal optimization results since each network is trained separately [46]. To address this limitation, we propose an end-to-end network for unified optimization of HRTF individualization, spanning from the pinna image to HRTF magnitude. For our network model, we utilize a CNN, which has proven effective in various domains such as image recognition, speech processing, and sound event recognition [47], [48]. However, the ear pinna exhibits intricate structures, including the concha, helix, and fossa. Moreover, the resonance modes of the pinna are influenced not only by local structures but also by overall shape factors like the width and depth of the pinna cavity [38]. Thus, the CNN must capture the relationship between the comprehensive structural patterns of the pinna, ranging from local details to the overall shape. To effectively recognize the complex structural patterns of the pinna and account for the acoustical effects of pinna shape on spectral cues, we employ residual blocks inspired by the ResNet architecture [49]. These residual blocks enable the learning of low-to-high level patterns within the pinna image across various network layers. Fig. 4(b) provides a visual depiction of the designed network model, PRTFNet.

Although the data shape of a full-spherical HRTF is determined by the number of frequency bins × the number of azimuths × the number of elevations, the magnitude spectrum of a full-spherical HRTF can be effectively utilized as an output of three sub-networks by employing autoencoder-based dimensionality reduction [27]. However, when employing PRTFNet to synthesize the magnitude spectrum of a full-spherical HRTF, the network model may encounter the issue of overfitting due to the substantial output dimension. This problem can impede PRTFNet from effectively learning the direction-specific characteristics of spectral cues. To address this, we propose direction-wise training for PRTFNet by introducing one hot encoding for the direction index as an additional network input, and employing HRTF magnitude for the corresponding direction as the network output. The choice to utilize one-hot encoding for the direction index is grounded in the insights presented in [50], which indicate the efficacy of one-hot encoding for accurately representing sound source or microphone direction/position within DNN. Fig. 4(a) and (c) illustrate the network inputs and output of

PRTFNet. The 2D one-hot encoding is transformed into a 256-dimensional vector through flattening and an FC layer. The role of the FC layer is to transform the direction index into the embedding vector, capable of seamless integration with the pinna image for input representation. We then concatenate the pinna image with this vector, creating a concatenated input. By adopting direction-wise training, we enable PRTFNet to reduce the dimensionality of the network output. To facilitate direction-wise training, we employ a loss function designed for synthesizing the magnitude spectrum of HRTF, considering the perceptual characteristics of the auditory system. It is well known that the auditory system perceives the direction of sound sources based on the log-scale HRTF magnitude [36]. Previous methods for HRTF individualization have commonly utilized log-spectral distortion (LSD) [51] as the objective metric, which is defined as

$$\mathrm{LSD} = \sqrt{\frac{1}{N_d N_f}\sum_{j=1}^{N_d}\sum_{i=1}^{N_f}\left(20\log\frac{\left|H_{\phi_j,\theta_j}(f_i)\right|}{\left|\tilde{H}_{\phi_j,\theta_j}(f_i)\right|}\right)^2}, \quad (1)$$

where $H_{\phi_j,\theta_j}(f_i)$ represents the true HRTF of the i-th frequency bin at the j-th direction of sound source (ϕ_j for sound elevation and θ_j for sound azimuth). $\tilde{H}_{\phi_j,\theta_j}$ denotes the predicted HRTF. N_f represents the number of frequency bins, and N_d denotes the number of directions. However, in the case of PRTFNet, direction-wise training is conducted separately for each direction, rather than training the model for all directions simultaneously. The loss function chosen for PRTFNet is the LSD for the target direction specified by ϕ and θ. The loss function is defined as

$$\mathrm{Loss} = \sqrt{\frac{1}{N_f}\sum_{i=1}^{N_f}\left(20\log\frac{\left|H_{\phi,\theta}(f_i)\right|}{\left|\tilde{H}_{\phi,\theta}(f_i)\right|}\right)^2}. \quad (2)$$

D. SPATIAL RESOLUTION OF DIRECTION INDEX

In the field of spatial audio, both listeners and sound sources have the freedom to move, which implies that the direction of sound sources can vary in any direction. Consequently, one of the primary objectives in spatial audio rendering is to achieve continuous HRTFs for all directions.

PRTFNet is required to generate continuous HRTFs of the target listener for spatial audio rendering. While enhancing the spatial resolution of continuous HRTFs can alleviate undesirable audio artifacts arising from spatial discretization of HRTF during spatial audio rendering [52], the computation time required to train PRTFNet increases with the spatial resolution of the direction index. Therefore, direction-wise training should be conducted with an appropriate spatial resolution of the direction index, considering computational costs and spatial hearing perception. Research has indicated that the minimum audible angle (MAA) for sound localization perception is 5.4° or higher during source or head movement [5]. Furthermore, HRTF interpolation methods employing manifold learning [53] can reconstruct an HRTF using neighboring HRTFs sampled at intervals below 20°. Taking into consideration the spatial hearing resolution of humans and the HRTF reconstruction performance of interpolation methods, we have chosen a spatial resolution of 10° for both azimuth (with a total of 36 azimuth values) and elevation (with a total of 14 elevation values), as depicted in Fig. 4(a).

V. PERSONALIZATION OF HRTF PHASE

The synthesized HRTF magnitude from PRTFNet is up-sampled to match the frequency resolution of the sound used for spatial audio. However, even after the up-sampling process, the HRTF phase spectrum is still omitted. In other words, the up-sampled HRTF magnitude does not include the information regarding ITD, which is an essential azimuth localization cue. In order to achieve a complete individualized HRTF set, we incorporate a signal processing module in the final step of the HRTF individualization scheme. The signal processing module involves selecting the appropriate HRTFs from a database by the measurement of head width. Additionally, the phase spectra of the selected HRTFs are adjusted to ensure accurate localization cues. By integrating this module into the scheme, we can compensate for the missing ITD information and obtain the individualization of the HRTF set.

A. SIMPLIFIED HRTF SELECTION

The spectral cues present in the HRTF magnitude spectrum predominantly originate from the pinna. However, ITD at lower frequencies, which are particularly significant for azimuth localization compared to ITD at higher frequencies [54], are primarily influenced by the dimensions of the head and torso [15]. Therefore, personalizing the phase spectrum of the HRTF requires additional anthropometric features related to the head and torso. However, measuring numerous anthropometric features can be time-consuming, and the results may vary depending on the person taking the measurements. To address this inconvenience, our objective is to use a minimal set of anthropometric features for synthesizing the HRTF phase spectra. Previous research [55] demonstrated that selecting HRTFs from a database based on the closest head width generally results in lateral perception errors within acceptable thresholds, often not exceeding 1° of localization blur. Furthermore, Algazi et al. [56] established through linear regression analysis that head width is highly correlated with ITD among anthropometric features. Building upon these findings and considering the availability of public HRTF databases, we employ a simplified HRTF selection method to personalize the HRTF phase spectra for both ears. This method relies solely on the measurement of head width.

In order to select the HRTFs, we begin by measuring the target listener's head width. This is done by measuring the distance between the points in front of the tragus on both sides. Subsequently, from a HRTF database containing anthropometric features and measured HRTFs (e.g., CIPIC [40], ITA [41], HUTUBS [30]), we choose the HRTFs of the subject whose head width is most similar to

that of the target listener. Finally, we extract the phase spectra from the selected HRTFs.

B. PHASE ADJUSTMENT

The phase spectra extracted by the simplified HRTF selection method include spectral components below half of the sampling frequency of the HRTF database. However, the high-frequency range of the extracted HRTF phase spectra can lead to distorted lateral perception for the listener, as it highly deviates from the individual phase spectrum. To address this issue and preserve accurate lateral perception, it is necessary to restrict the frequency range of the extracted phase spectra. Research has shown that ITD below 1.5 kHz plays a crucial role in the lateral perception of sound source direction [56]. In experiments where sound stimuli are presented with conflicting ITD and ILD cues, subjects predominantly respond to the direction indicated by the low-frequency ITD [57]. While ITD takes precedence over ILD in lateral perception when wideband stimuli include low frequencies, localization responses align with ILD when the stimuli are high-pass filtered with a cutoff frequency of 2.5 kHz. These findings indicate that spectral components of HRTF phase below 2.5 kHz serve as the primary cue for lateral localization (as ITD is derived from interaural phase difference). Therefore, removing the spectral components of HRTF phase below 2.5 kHz eliminates the primary lateral localization cue. Based on the experimental results verifying the importance of ITD below 2.5 kHz, we truncate the extracted phase spectra above 2.5 kHz. This minimizes the distortion of lateral perception caused by high frequencies without sacrificing the primary lateral localization cue. The removed spectral components of the HRTF phase are replaced with 0 values.

The phase spectra are extracted from the HRTFs selected based on the closest head width, but there can still be a distortion in the low-frequency ITD due to the difference in head width. To address this, we propose adjusting the scale of the extracted HRTF phase spectra to obtain accurate individualized low-frequency ITD. It is observed that the phase spectrum of HRTF is mostly linear [58], and the distribution of ITD depending on elevation and azimuth is similar across human subjects, with differences mainly attributed to the scaling factor influenced by the size of the head [59]. Furthermore, it is noted that listeners are not highly sensitive to the detailed spectral components of HRTF phase in terms of lateral perception [60]. Considering the above findings, we utilize the formula for ITD of a spherical head model [61], which is described as

$$\mathrm{ITD}_{\phi,\theta}(f) = -\frac{\psi_L(f) - \psi_R(f)}{2\pi f} = \frac{a}{c}(\sin\theta + \theta)\cos\phi. \quad (3)$$

c represents the speed of sound, a denotes half the head width, and $\psi_L(f)$ and $\psi_R(f)$ represent the phases of sound pressures at the left and right ears, respectively. Eq. (3) reveals that the interaural phase difference (IPD) in HRTFs is directly proportional to the head width. Thus, we can establish the relationship between the IPD of the target listener and the IPD of the subject, whose HRTF phase spectra for both ears are extracted from the HRTF database, as follows:

$$\psi_{L,l}(f) - \psi_{R,l}(f) = \frac{a_l}{a_s}\left[\psi_{L,s}(f) - \psi_{R,s}(f)\right]. \quad (4)$$

Here, the subscript l and s indicate the listener and subject, respectively. Primary cue for lateral localization is ITD, rather than monaural HRTF phase. To account for the difference in head width measurement between the listener and subject, the scale of the IPD should be adjusted accordingly. Based on the findings, we multiply the extracted HRTF phase spectra of both ears by the ratio of head widths, a_l/a_s, to obtain the accurate individualized ITD for the listener. The personalized HRTF phase spectra is then resampled to align with the frequency resolution of the sound used for spatial audio.

VI. EXPERIMENTAL SETUP

In this study, we conducted validation of the proposed PRTFNet and personalization of HRTF phase spectra using the HUTUBS HRTF database [30]. The database includes measured HRIRs for azimuths ranging from 0° to 350° and elevations ranging from −90° to 90° (sampled at 10° increments in azimuth and elevation). The HRIR measurements were executed within an anechoic chamber, utilizing continuous rotation of the subject to optimize measurement efficiency and minimize unconscious subject movements. Additionally, the database provides a scanned 3D mesh of the head and ears shape, as well as anthropometric measurements of the ears, head, and torso. Notably, the 3D mesh samples contain unadulterated ear shapes, devoid of extraneous elements such as hair, which could potentially divert the focus of the PRTFNet model from the ear's distinctive structure. To generate an individual pinna image, we convert the side view of the corresponding 3D mesh into a 2D grayscale image. Subsequently, we down-sample the grayscale image to a resolution of 256 × 256 pixels, as depicted in Fig. 4(a).

Among the 116 ear samples available in the HUTUBS database, a subset of 90 ear samples (corresponding to 45 subjects) was used for training the PRTFNet model. An additional subset of 14 ear samples (from 7 subjects) was reserved for the test dataset. This selection excluded artificial heads, repeated subjects, and ears with earrings from the datasets. Regarding phase personalization, the 58 subjects were divided into two groups: a HRTF selection pool consisting of 50 subjects, from which a HRTF was chosen for the extraction of phase spectra, and an independent pool of 8 subjects for evaluating the personalized HRTF phase. For the target directions, we defined a total of 36 azimuths (ranging from 0° to 350° with a resolution of 10°) and 14 elevations (ranging from −40° to 90° with a resolution of 10°).

VII. PERFORMANCE EVALUATION

A. LOG SPECTRAL DISTORTION

To objectively evaluate the performance of PRTFNet, we employed LSD as a measure of the synthesis quality of

TABLE 1. Objective performance comparison of different methods for HRTF magnitude individualization in terms of LSD and LSD_E.

Methods	LSD, dB	LSD_E, dB
Baseline [27]	10.4	7.5
Baseline with PRTF estimate	6.9	6.9
Baseline with compact PRTF	6.1	6.3
CNN with full grid HRTF	12.3	14.5
CNN with direction-wise training	8.8	8.5
PRTFNet	**5.0**	**5.1**

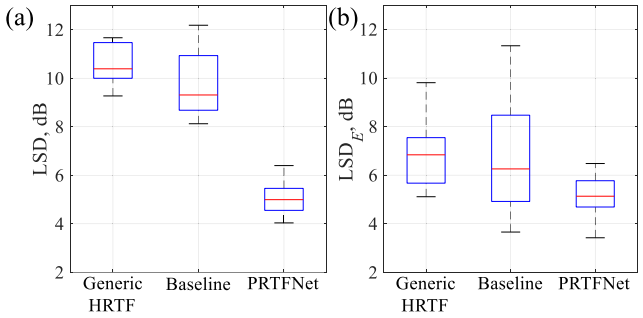


FIGURE 5. Distribution of LSD and LSD_E across test dataset depending on HRTF magnitude individualization methods: central box mark, median; box edges, 25th/75th percentiles; whiskers, 5th/95th percentiles. (a) LSD, (b) LSD_E.

HRTF magnitude. Additionally, we introduced an effective LSD (LSD_E) specifically defined within the frequency range (4-16 kHz) relevant to elevation localization cues [8]. The ground truth for calculating performance metrics was established using the HRTF derived from measured HRIR data. Furthermore, we conducted an ablation study to evaluate the individual contributions of the PRTF estimate, compact PRTF, CNN, and direction-wise training proposed in this study. The average LSD and LSD_E values obtained from the test dataset are presented in Table 1. Compared to the baseline [27], the PRTF estimate and compact PRTF achieved an improvement of more than 3 dB in LSD due to the elimination of sound reflection effects originating from the head and torso. In the case of CNN with full-spherical HRTF, the LSD and LSD_E values were found to degrade due to the high dimensionality of the network output. However, the direction-wise training approach yielded improved LSD and LSD_E values of approximately 4 dB and 6 dB, respectively, while reducing the output dimensionality. The proposed PRTFNet, combining compact PRTF, CNN, and direction-wise training, achieved LSD and LSD_E values of 5.0 dB and 5.1 dB, respectively, surpassing the baseline performance.

The distributions of LSD and LSD_E across the test dataset were compared, and the results are depicted in Fig. 5. Additionally, we included the individualization performance of a generic HRTF, specifically using the HRTFs of the artificial head and torso B&K HATS Type 4100 [34] as a reference. Among the evaluated options, PRTFNet demonstrated the smallest standard deviation of the LSDE distribution (0.9 dB), outperforming the generic HRTF (1.2 dB), the average model of ear shapes, and the baseline (4.3 dB). This outcome clearly illustrates that PRTFNet delivers robust predictions of HRTF magnitude, effectively accommodating individual variations in ear pinna characteristics.

B. COMPARISONS OF NEURAL NETWORK OUTPUTS

Fig. 6 displays the verification of spectral cue reconstruction by plotting the true parametric HRTFs (derived from measured HRIRs), along with the network outputs from PRTFNet and the baseline, in the median plane. This plane is particularly relevant for observing the variations in spectral cues [33], [38]. On the frontal side ($\phi = 0°$, $\theta = 0°$), PRTFNet successfully reconstructed the first and second peaks and first notches within the 4-8 kHz range. On the rear side ($\phi = 180°$, $\theta = 0°$), PRTFNet accurately predicted the steep first notch at approximately 8 kHz, including its center frequency and magnitude. However, due to the dominance of high-order pinna modes at higher frequencies, which are challenging to predict with pinna images, the spectral distortion increased beyond 10 kHz. It is noteworthy that in sound localization tests using parametric HRTFs, the two lowest-frequency notches and peaks provide similar elevation localization performance as measured HRTFs [62]. This implies that the perception of sound elevation is primarily determined by the lowest-frequency spectral cues, and PRTFNet successfully preserves these crucial peaks and notches, enabling accurate elevation localization for listeners. Furthermore, the prediction results from PRTFNet align with previous experimental findings, wherein the frequency of prominent notches shifts higher as the sound source moves from the front of the subject to above their head. This characteristic ensures that localization perception errors, such as the rising of a sound image, can be prevented with PRTFNet by leveraging the elevation angle dependency of the notch frequency.

C. ITD

To validate the personalization of HRTF phase spectra, we compared the IPDs and ITDs for the true (measured) HRTFs of subjects in the independent pool, the selected HRTFs based on closest head width, and the personalized phase spectra. These comparisons are illustrated in Fig. 7(a) and (b). Fig. 7(b) demonstrates that the difference between the true ITD and predicted ITD is reduced with the use of personalized HRTF phase spectra, particularly around 860 Hz. This indicates that the estimation error of ITD can be improved by multiplying the phase spectra of selected HRTFs by the head width ratio. For the objective evaluation of ITD estimation error, we employ the root mean square error (RMSE) of ITD, which is expressed as

$$\varepsilon_{\mathrm{ITD}} = \sqrt{\frac{1}{N_d M_f} \sum_{j=1}^{N_d} \sum_{i=1}^{M_f} \left[\mathrm{ITD}_{\phi_j,\theta_j}(f_i) - \mathrm{ITD}_{\phi_j,\theta_j}(f_i)\right]^2}, \quad (5)$$

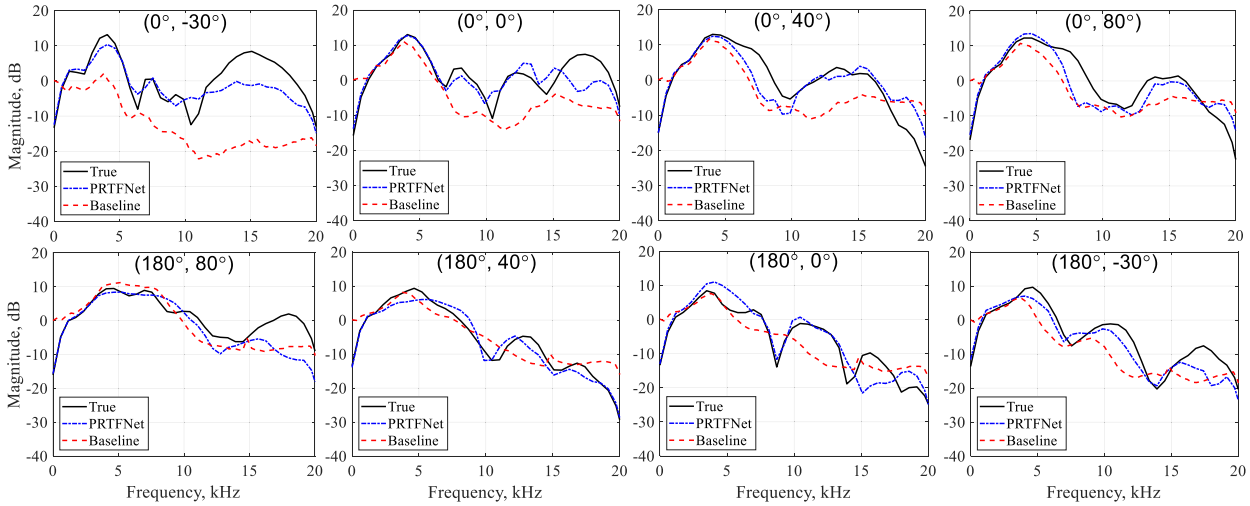


FIGURE 6. Comparison of true parametric HRTFs and network outputs from PRTFNet and baseline. The title (ϕ, θ) denotes azimuth angle of ϕ and elevation angle of θ for DoA of sound source.

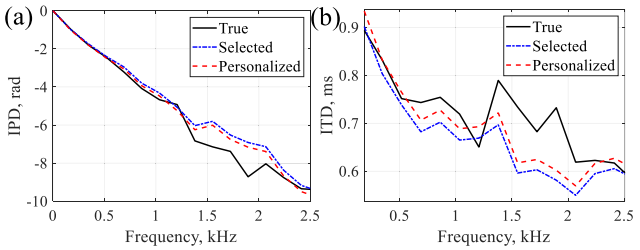


FIGURE 7. Comparison of IPDs and ITDs for true, selected, and personalized phase spectra of both ears at azimuth 70° and elevation 0°. (a) IPDs, (b) ITDs.

TABLE 2. Average RMSE of ITD (ε_{ITD}) on independent pool according to different methods for personalization of HRTF phase.

Methods	ε_{ITD}, ms
Generic HRTF [34]	0.0523
Selected HRTF phase	0.0433
Personalized HRTF phase	**0.0407**

The number of frequency bins below 2.5 kHz is denoted as M_f. The personalization results for the average RMSE on the independent pool are summarized in Table 2. It is noteworthy that the HRTFs selected based solely on head width measurements exhibit more accurate ITDs compared to the HRTFs of the artificial HATS. Furthermore, the personalized HRTF phase for the subject in the independent pool achieved the lowest RMSE of ITD, measuring at 0.0407 ms. This outcome confirms that the proposed phase personalization approach, utilizing the head width ratio, effectively mitigates ITD distortion arising from differences in head width.

VIII. CONCLUSION

This paper presents a novel HRTF individualization scheme that utilizes only pinna image as input for the HRTF magnitude individualization network, and relies on head width measurement to generate personalized HRTF phase spectra. To accurately reconstruct the spectral cues crucial for elevation localization, we propose a deep learning-based model called PRTFNet for HRTF magnitude individualization using pinna images. PRTFNet employs compact PRTF as the output of the network, which effectively eliminates sound reflection effects from the head and torso. This approach ensures a more accurate correlation between the network input and output, while also minimizing overfitting in hundreds of frequency bins. To achieve unified optimization of HRTF magnitude individualization, we employ an end-to-end CNN in the network structure of PRTFNet. The CNN spans from the input pinna image to the output HRTF magnitude. Additionally, we incorporate direction-wise training into the CNN to capture the directional properties of spectral cues and reduce the dimensionality of the network output.

The proposed PRTFNet was validated using the HUTUBS dataset, and its individualization performance was evaluated based on LSD and LSD_E metrics. The results demonstrated that PRTFNet outperformed previous deep learning-based model, achieving significant gains of up to 5 dB and 2 dB in terms of LSD and LSD_E, respectively. Moreover, PRTFNet exhibited the smallest standard deviation, indicating its robustness in predicting HRTF magnitude across various ear pinna variations. Analyzing the HRTF magnitude generated by PRTFNet, we observed the accurate reconstruction of the first and second peaks and first notches below 8 kHz. These spectral features serve as primary localization cues for sound elevation and their faithful reproduction further confirms the efficacy of PRTFNet in providing accurate elevation localization for listeners. Additionally, the elevation angle dependency of the notch frequency in the generated HRTF magnitude aligned with experimental results obtained from measured HRTFs. This alignment serves as compelling evidence that PRTFNet effectively prevents localization perception errors.

We incorporated a phase personalization step in our HRTF individualization scheme to obtain personalized phase spectra for both ears' HRTFs. This phase personalization process involved two key steps: HRTF selection based on head width measurement and phase adjustment, as proposed in this study. The phase adjustment was performed by multiplying the

ratio of the target listener's head width to the head width of the selected subject from the HRTF database. To assess the accuracy of the proposed phase adjustment, we evaluated the RMSE of ITD. By applying the phase adjustment to the selected phase spectra, the RMSE of ITD was reduced by approximately 0.003 ms. These results indicate that the proposed compensation of IPD scale for the head width difference between the target subject and the selected subject effectively alleviates ITD distortion in the selected phase spectra.

As part of our future work, we plan to conduct a subjective test to evaluate the localization perception using the proposed PRTFNet and phase personalization. This subjective test will provide valuable insights into spectral cues in HRTFs and the acceptable range of ITD estimation error for achieving accurate spatial hearing perception. By analyzing the results of the subjective test, we aim to further refine and improve the performance of our methods for HRTF individualization.

REFERENCES

[1] C. Schissler, A. Nicholls, and R. Mehra, "Efficient HRTF-based spatial audio for area and volumetric sources," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 4, pp. 1356–1366, Apr. 2016.

[2] C. A. Dimoulas, "Audiovisual spatial-audio analysis by means of sound localization and imaging: A multimedia healthcare framework in abdominal sound mapping," *IEEE Trans. Multimedia*, vol. 18, no. 10, pp. 1969–1976, Oct. 2016.

[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] X. Hu, A. Song, Z. Wei, and H. Zeng, "StereoPilot: A wearable target location system for blind and visually impaired using spatial audio rendering," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1621–1630, 2022.

[5] 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.

[6] 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, Jul. 1993.

[7] L. Rayleigh, "On our perception of sound direction," *Philos. Mag.*, vol. 13, no. 74, pp. 214–232, 1907.

[8] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *J. Acoust. Soc. Amer.*, vol. 56, no. 6, pp. 1829–1834, Dec. 1974.

[9] L. A. J. Reiss and E. D. Young, "Spectral edge sensitivity in neural circuits of the dorsal cochlear nucleus," *J. Neurosci.*, vol. 25, no. 14, pp. 3680–3691, Apr. 2005.

[10] G. Yu, R. Wu, Y. Liu, and B. Xie, "Near-field head-related transfer-function measurement and database of human subjects," *J. Acoust. Soc. Amer.*, vol. 143, no. 3, pp. EL194–EL198, Mar. 2018.

[11] T. Huttunen, E. T. Seppälä, O. Kirkeby, A. Kärkkäinen, and L. Kärkkäinen, "Simulation of the transfer function for a head-and-torso model over the entire audible frequency range," *J. Comput. Acoust.*, vol. 15, no. 4, pp. 429–448, Dec. 2007.

[12] Y. Kahana and P. A. Nelson, "Numerical modelling of the spatial acoustic response of the human pinna," *J. Sound Vibrat.*, vol. 292, nos. 1–2, pp. 148–178, Apr. 2006.

[13] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato, "Comparison of simulated and measured HRTFs: FDTD simulation using MRI head data," in *Proc. 123rd AES Conv.*, Oct. 2007, p. 7240.

[14] E. A. G. Shaw, "Acoustical features of the human external ear," in *Binaural Spatial Hearing Real Virtual Environments*, R. H. Gilkey T. R. Anderson, Eds. Mahwah, NJ, USA: Lawrence Erlbaum, 1997, pp. 25–47.

[15] V. R. Algazi, R. O. Duda, and D. M. Thompson, "The use of head-and-torso models for improved spatial sound synthesis," in *Proc. 113th Conv. Audio Eng. Soc.*, Los Angeles, CA, USA, Oct. 2002, pp. 1–18.

[16] S. Spagnol, M. Geronazzo, and F. Avanzini, "On the relation between pinna reflection patterns and head-related transfer function features," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 3, pp. 508–519, Mar. 2013.

[17] H. Hu, L. Zhou, J. Zhang, H. Ma, and Z. Wu, "Head related transfer function personalization based on multiple regression analysis," in *Proc. Int. Conf. Comput. Intell. Secur.*, Nov. 2006, pp. 1829–1832.

[18] 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.

[19] 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.

[20] 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.

[21] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato, "Frequency and amplitude estimation of the first peak of head-related transfer functions from individual pinna anthropometry," *J. Acoust. Soc. Amer.*, vol. 137, no. 2, pp. 690–701, Feb. 2015.

[22] J. Lu and X. Qi, "Pre-trained-based individualization model for real-time spatial audio rendering system," *IEEE Access*, vol. 9, pp. 128722–128733, 2021.

[23] B. Zhi, D. N. Zotkin, and R. Duraiswami, "Towards fast and convenient end-to-end HRTF personalization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 441–445.

[24] G. Lee and H. Kim, "Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear," *Appl. Sci.*, vol. 8, no. 11, p. 2180, Nov. 2018.

[25] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson, "Structural composition and decomposition of HRTFs," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust.*, Oct. 2001, pp. 103–106.

[26] 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.

[27] R. Miccini and S. Spagnol, "A hybrid approach to structural modeling of individualized HRTFs," in *Proc. IEEE Conf. Virtual Reality 3D User Interface Abstr. Workshops (VRW)*, Mar. 2021, pp. 80–85.

[28] W. Chen, R. Hu, X. Wang, and D. Li, "HRTF representation with convolutional auto-encoder," in *Proc. Int. Conf. Multimedia Model.*, 2020, pp. 605–616.

[29] K. Iida, M. Itoh, A. Itagaki, and M. Morimoto, "Median plane localization using a parametric model of the head-related transfer function based on spectral cues," *Appl. Acoust.*, vol. 68, no. 8, pp. 835–850, 2007.

[30] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses," *J. Audio Eng. Soc.*, vol. 67, no. 9, pp. 705–718, Sep. 2019.

[31] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization," *J. Acoust. Soc. Amer.*, vol. 88, no. 1, pp. 159–168, 1990.

[32] R. A. Butler and K. Belendiuk, "Spectral cues utilized in the localization of sound in the median sagittal plane," *J. Acoust. Soc. Amer.*, vol. 61, no. 5, pp. 1264–1269, May 1977.

[33] E. A. Lopez-Poveda and R. Meddis, "A physical model of sound diffraction and reflections in the human concha," *J. Acoust. Soc. Amer.*, vol. 100, no. 5, pp. 3248–3259, Nov. 1996.

[34] G.-T. Lee, S.-M. Choi, B.-Y. Ko, and Y.-H. Park, "HRTF measurement for accurate sound localization cues," 2022, *arXiv:2203.03166*.

[35] A. D. Musicant and R. A. Butler, "The influence of pinnae-based spectral cues on sound localization," *J. Acoust. Soc. Amer.*, vol. 75, no. 4, pp. 1195–1200, Apr. 1984.

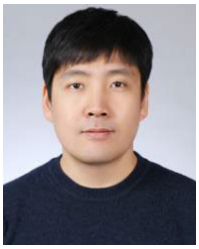
[36] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient real spherical harmonic representation of head-related transfer functions," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 921–930, Aug. 2015.

[37] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *J. Acoust. Soc. Amer.*, vol. 109, no. 3, pp. 1110–1122, Mar. 2001.

[38] K. Iida and M. Oota, "Median plane sound localization using early head-related impulse response," *Appl. Acoust.*, vol. 139, pp. 14–23, Oct. 2018.

[39] S. Spagnol, M. Geronazzo, and F. Avanzini, "Fitting pinna-related transfer functions to anthropometry for binaural sound rendering," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Oct. 2010, pp. 194–199.
[40] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust. (WASSAP)*, Oct. 2001, pp. 99–102.
[41] R. Bomhardt, M. D. L. F. Klein, and J. Fels, "A high-resolution head-related transfer function and three-dimensional ear model database," *Proc. Meetings Acoust.*, vol. 29, no. 1, 2016, Art. no. 050002.
[42] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Beijing, China: O'Reilly Media, 2017.
[43] S. Badillo, "An introduction to machine learning," *Clin. Pharmacol. Ther.*, vol. 107, no. 4, pp. 871–885, 2020.
[44] J. O. Smith, *Mathematics of the Discrete Fourier Transform (DFT): With Audio Applications*. Charleston, SC, USA: BookSurge, 2007. [Online]. Available: https://books.google.com/books?id=fTOxS9huzHoC
[45] M. Viswanathan, *Wireless Communication Systems in MATLAB*, 2nd ed. Chicago, IL, USA: Independent publisher, 2020.
[46] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 6964–6968.
[47] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[48] G.-T. Lee, H. Nam, S.-H. Kim, S.-M. Choi, Y. Kim, and Y.-H. Park, "Deep learning based cough detection camera using enhanced features," *Expert Syst. Appl.*, vol. 206, Nov. 2022, Art. no. 117811.
[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[50] Y. Gong, S. Liu, and X.-L. Zhang, "End-to-end two-dimensional sound source localization with ad-hoc microphone arrays," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2022, pp. 1944–1949.
[51] 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.
[52] J. W. Scarpaci and J. A. White, "A system for real-time virtual auditory space," in *Proc. ICAD*, Jul. 2005, pp. 241–246.
[53] F. Grijalva, L. C. Martini, D. Florencio, and S. Goldenstein, "Interpolation of head-related transfer functions using manifold learning," *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 221–225, Feb. 2017.
[54] H. S. Colburn and N. I. Durlach, "Models of binaural interaction," *Handbook Perception*, vol. 4, pp. 467–518, Jan. 1978.
[55] S. Spagnol, "HRTF selection by anthropometric regression for improving horizontal localization accuracy," *IEEE Signal Process. Lett.*, vol. 27, pp. 590–594, 2020.
[56] 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.
[57] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1648–1661, 1992.
[58] S. Mehrgardt and V. Mellert, "Transformation characteristics of the external human ear," *J. Acoust. Soc. Amer.*, vol. 61, no. 6, pp. 1567–1576, 1977.
[59] I. Tashev, "HRTF phase synthesis via sparse representation of anthropometric features," in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2014, pp. 1–5.
[60] 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.
[61] R. S. Woodworth and H. Schlosberg, *Experimental Psychology*. New York, NY, USA: Holt, Rinehard and Winston, 1962.
[62] K. Iida and Y. Ishii, "Effects of adding a spectral peak generated by the second pinna resonance to a parametric model of head-related transfer functions on upper median plane sound localization," *Appl. Acoust.*, vol. 129, pp. 239–247, Jan. 2018.

BYEONG-YUN KO received the B.S. degree in ship architecture and ocean engineering from Inha University, South Korea, in 2019, and the M.S. degree in mechanical engineering from the Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2021. His research interests include the development of deep learning-based individualization of HRTF and spatial audio rendering technique considering the human auditory perception.



GYEONG-TAE LEE received the B.S. degree in mechanical engineering from Kyonggi University, in 2005, and the M.S. degree in mechanical engineering from Hanyang University, in 2007. He is currently pursuing the Ph.D. degree in mechanical engineering with the Korea Advanced Institute of Science and Technology (KAIST). He was a Senior Engineer with ten years professional experience in acoustics and signal processing, responsible for the research and development in the field of electro-acoustic with Samsung Electronics, South Korea. During this period, he has performed ten projects about the development of slim acoustic solutions and technologies. He developed a number of slim acoustic transducers, speaker response simulators, slim speaker systems, sound measurement and tuning systems, and sound quality evaluation systems. From these, he received five awards from Samsung Electronics. Over the course of his career, he has published 29 papers and 38 patents in his field. His research interests include sound event localization, detection for humanoid robot, AI, audio, electro-acoustics, and signal processing.

HYEONUK NAM (Member, IEEE) received the B.S. and M.S. degrees in mechanical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree in mechanical engineering. His research interests include sound event detection, sound localization, automatic speech recognition, and speech dereverberation.

YONG-HWA PARK (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in mechanical engineering from the Korea Advanced Institute of Science and Technology (KAIST), in 1991, 1993, and 1999, respectively. In 2000, he joined the Aerospace Department, University of Colorado Boulder, as a Research Associate. From 2003 to 2016, he was with the Samsung Electronics in Visual Display Division and Samsung Advanced Institute of Technology (SAIT) as a Research Master in the field of micro-optical systems with applications to imaging and display systems. In 2016, he joined KAIST as an Associate Professor of noise and vibration control plus (NOVIC+) with the Department of Mechanical Engineering devoting to researches on vibration, acoustics, vision sensors, and recognitions for human-machine interactions. His research interests include structural vibration; event/condition recognition from sound and vibration signatures utilizing AI; blood pressure and health monitoring sensors; 3D sensors, and lidar for motion measurements. He has been the Conference Chair of MOEMS and miniaturized systems in SPIE Photonics West, since 2013. He is a Board Member of KSME, KSNVE, KSPE, and SPIE.