HRTF Individualization Based on Anthropometric Measurements Extracted from 3D Head Meshes

Davide Fantini
Dept. of Computer Science
University of Milan
Milan, Italy
davide.fantini1@studenti.unimi.it

Federico Avanzini
Dept. of Computer Science
University of Milan
Milan, Italy
federico.avanzini@unimi.it

Stavros Ntalampiras
Dept. of Computer Science
University of Milan
Milan, Italy
stavros.ntalampiras@unimi.it

Giorgio Presti
Dept. of Computer Science
University of Milan
Milan, Italy
giorgio.presti@unimi.it

2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA) | 978-1-6654-0998-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/I3DA48870.2021.9610904

***Abstract*—In the field of 3D audio, the use of Head-Related Transfer Functions (HRTFs) compliant to the subject anatomical traits is crucial to guarantee a proper individual experience. This work proposes an HRTF individualization method based on anthropometric features automatically extracted from 3D head meshes. The method aims at a fully automated process able to estimate individual median plane HRTF starting from a 3D mesh of the subject's pinna. The method relies on the HUTUBS dataset including 3D meshes, anthropometry and HRTFs. In the first phase, a set of pinna anthropometric parameters is extracted from the 3D meshes converted to range images. A set of landmarks is fitted on the pinna through the Active Shape Model algorithm to outline its shape. Then, the set of pinna anthropometric parameters defined in HUTUBS is automatically extracted exploiting the landmarks. In the second phase, the relationship between pinna anthropometry and HRTFs is modelled. For each elevation angle considered in HUTUBS, a Generalized Regression Neural Network is trained to predict the corresponding HRTF, given the anthropometry. The method is evaluated in both objective and perceptual metrics showing performances comparable to the state of the art.**

I. Introduction

A. Problem overview

The human ability to localize a sound source in the surrounding space highly depends on individual anatomical traits. In their propagation, the sound waves collide with the listener's body which causes several delay and filtering effects. As result, the listener's brain analyses the influence of these effects on the sound to infer the source position. While every anatomical component shapes the incoming sound waves, this work is focused on the most individual one, the pinna. The transformation applied by the human body to the sound waves can be simulated through a pair of *Head-Related Transfer Function* (HRTF) sets [1], one for each ear. The HRTFs model the effects of human body on sound waves as a Linear and Time-Invariant (LTI) system. A HRTF set for a specific subject (a human or a dummy head) is a collection of multiple transfer functions, one for each spatial position of interest. A HRTF is represented by the Head-Related Impulse Response (HRIR) in time domain. The use of HRTFs along with headphones allows the listener to experience the sensation of a sound source positioned in a 3D virtual auditory space.

HRTFs find applications in several domains (e.g. music, gaming, virtual reality and teleconferencing). In user applications, the employed HRTFs are often a generic set identical for each subject, usually recorded from a dummy head. However, a generic HRTF is not suitable for each subject which has distinctive anatomical traits. As consequence of listening with generic HRTF sets, several inadequacies could arise. Some of the most studied ones are front-back confusion [2], lack of externalization [3], localization accuracy degradation (mainly in elevation perception [4] and to a lesser extent in horizontal perception [5]). Nevertheless, the direct measurement of listener's HRTF, known as individual HRTF, is impractical for user applications (high cost equipment, time-consuming). For these reasons, the literature includes several works proposed to approximate individual HRTFs without direct measurement [6], [7]. This task is known as HRTF individualization.

B. Method overview

An HRTF individualization method based on anthropometric features automatically extracted from 3D head meshes is proposed in this paper. The method aims at a fully automated process able to estimate the individual median plane HRTF starting from a 3D mesh of the subject's head. Such a process is articulated in two different phases: (a) pinna anthropometry extraction and (b) HRTF individualization.

In the *pinna anthropometry extraction* phase, a set of anthropometric parameters related to pinna is measured. First, the 3D head mesh is converted to a range image and cropped to include only the pinna. Then, a set of landmarks is fitted on the pinna with the Active Shape Model algorithm to outline its shape. From the range image and the outlined shape, a predefined set of anthropometric parameters describing the pinna is automatically extracted. These parameters aim to describe the pinna shape and its influence on the incoming sound waves.

In the *HRTF individualization* phase, the obtained anthropometric measurements are employed to train a regression model able to estimate the median plane HRTF. The HRTF is decomposed in the directional and common components represented by the *Directional Transfer Function* (DTF) and the *Common Transfer Function* (CTF). The anthropometric parameters are employed as input to *Generalized Regression Neural Network* (GRNN) models. A model for each elevation angle in the median plane is trained to predict the corresponding DTF, while a single model is trained to predict the CTF. Finally,

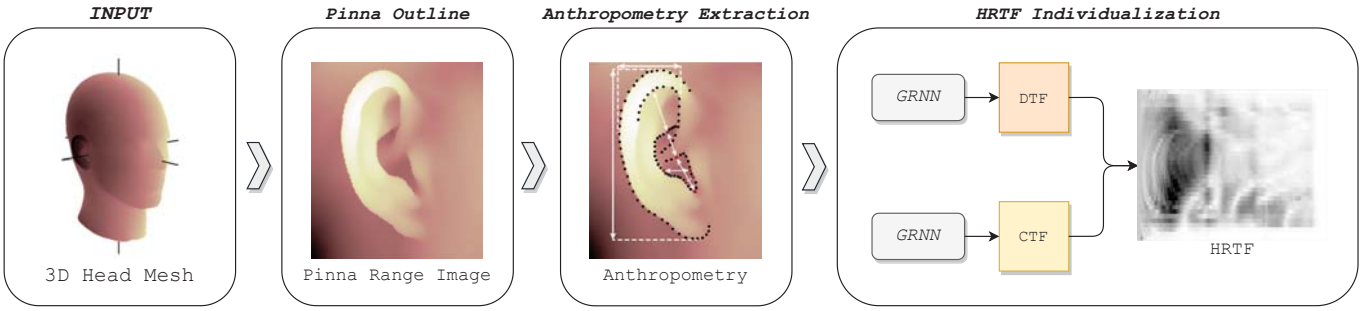

Fig. 1. Block diagram of the proposed HRTF individualization method.

the predicted DTFs and CTF are combined to reconstruct the HRTF. A block diagram of the overall HRTF individualization method is shown in Fig. 1.

C. Paper organization

The remainder of the paper is organized as follows. Section II reviews the HRTF individualization works proposed in literature. In Section III the HUTUBS datasets employed to develop the proposed method is presented. The discussion of the method itself is splitted in two sections: the parts related to pinna anthropometry extraction and to HRTF individualization are described in Section IV and Section V, respectively. Section VI shows the evaluation and the results of the proposed method. Finally, Section VII provides a brief recap of the proposed innovations with respect to the state of the art along with the possible future investigations.

II. Previous Works

Due to the inadequacy of generic HRTF sets and the impracticability of HRTF direct measurements for end-user applications, several HRTF individualization methods have been proposed in the literature. These methods can be organized into three main categories [6], [7]: numerical simulations, subjective selection, and anthropometry-based approaches.

In numerical simulation methods, such as [8]–[10], the propagation of sound waves around the subject is simulated through the numerical resolution of the wave equation, having the body parts as boundary conditions. Despite their potential, these methods are computationally expensive. Subjective selection methods, such as [11]–[13], evaluate the subject perceptual feedback (e.g., localization accuracy) using various HRTFs. Their major drawback is the time-consuming session in which the subject evaluates all the plausible HRTFs.

The anthropometry-based category includes the method proposed in this paper; thus, a deeper analysis of the related literature is presented. Since HRTF describes the way the sound waves interact with the body, several approaches are based on anthropometric parameters extracted from pinnae, head and torso. Although researchers agree on the relationship between HRTF and human anatomical shape and size [14], [15], there are still significant uncertainties on the exact influence of anthropometry on HRTF. Defining a set of parameters that sufficiently describes the HRTF behaviour is a currently open issue. In the literature, several measurement definitions have been suggested. In 2001, Algazi et al. proposed the CIPIC specification [16], which includes 27 anthropometric parameters and is still the most used one.

The anthropometry-based category can be divided in three main approaches: anthropometry matching, adaptation and regression. In anthropometry matching approach [17], [18] a best-match HRTF is selected from a dataset. The best-match HRTF is found by minimizing the distance between the anthropometry of the test subject and the dataset subjects. This method is quite simple but limited in effectiveness since it requires a sufficiently representative database. The adaptation approach [19], [20], instead, takes a non-individual HRTF and adjusts its behaviour processing it according to the subject anthropometry (e.g., frequency scaling proportional to anatomical size). In the regression approach, a regression model is trained to describe the relationship between the anthropometry (input) and the corresponding HRTF (output).

Several methods based on multiple linear regression have been proposed [21]–[23], however non-linear methods seem to be more suitable. Among non-linear methods, the use of deep learning has significantly risen in recent times [24]. Examples of this are the use of PCA [25], HOSVD [26] and Isomap [27] to reduce the HRTF dimensionality and the training of an artificial neural network to estimate the low-dimensional HRTF from anthropometry. However, no perceptual evaluation is reported for these three cited works. In 2018, Lee and Kim [28] proposed a method consisting in three Deep Neural Networks (DNNs) to estimate the individual HRTF. The first one is a feedforward DNN trained using anthropometry, while the second one is a Convolutional Neural Network (CNN) trained with edge-detected pinna images. Finally, a third DNN is trained using the outputs of the previous two networks. The authors showed an improvement in both objective and perceptual metrics compared to other methods. In 2019, Chen et al. [29] employed an autoencoder to reduce the HRTF dimensionality plus a DNN trained to predict the autoencoder's hidden neurons values from anthropometry. The authors showed a decrease of spectral distortion with respect to a pure DNN method, nevertheless no perceptual metric is reported.

Very few works in the literature aim to automatically extract

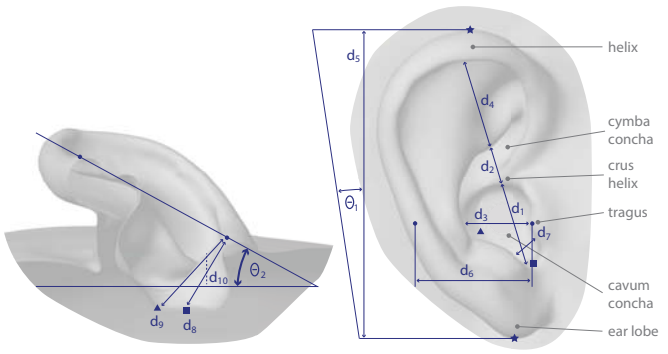

Fig. 2. Definition of HUTUBS pinna anthropometric measurements. Image reproduced from HUTUBS documentation [31].

the anthropometry from subject images. The method proposed in this paper makes use of Active Shape Models (ASM) [30] in order to perform this task. Another method based on pinna anthropometry extraction through ASM is [18]. The authors focused on standard color images, instead of the range images considered in this work. They used pictures of 11 subjects in front view, side view, and pinna alone, and they trained 3 independent ASMs by manually annotating the images with landmarks according to the anatomical shapes. After the fitting phase, they selected fixed landmarks to extract the anthropometry following the specifications proposed in the CIPIC database. Finally, they performed HRTF individualization through a best anthropometry match approach.

III. Dataset

The dataset employed to develop and evaluate the proposed HRTF individualization method is the HUTUBS dataset [31], [32], released in 2019. The dataset contains the HRIRs in SOFA format [33] measured for 96 subjects. Each impulse response (IR) is recorded at 44.1 kHz and it is 256 samples long. Each subject HRIR set is composed by the IRs measured at both ears in 440 positions around the subject with the sound source positioned 1.47 meters away. Using interaural polar coordinates, the elevation angles are equally spaced from $-90°$ to $270°$ by $10°$ intervals. Azimuth angles are spaced from $-90°$ to $90°$ by variable, elevation-dependent intervals.

In addition to the HRIRs, HUTUBS provides a set of anthropometric measurements concerning body, head and pinnae for 93 subjects. In Fig. 2, the definitions of the pinna measurements included in HUTUBS database is shown. This set of measurements is similar to the one reported in CIPIC dataset [16], but it has some changes and additions.

Furthermore, in HUTUBS dataset the 3D head meshes for 58 subjects are reported. All of these three types of data stored in HUTUBS database (HRIR sets, anthropometric measurements and 3D head meshes) are employed in the proposed method of HRTF individualization.

The HUTUBS dataset leads to several improvements with respect to older datasets, such as CIPIC that is the most used in literature. The 3D head meshes included in HUTUBS allow to virtually extract any kind of anthropometric measurement

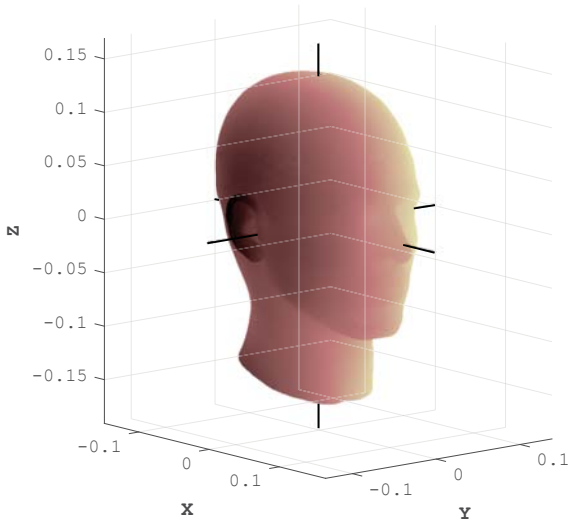

Fig. 3. Example of a 3D head mesh from HUTUBS. The color represents the y coordinate value.

even the ones requiring depth information. Furthermore, the HUTUBS dataset collects the HRTFs of a higher number of subjects with respect to many other HRTF datasets.

IV. Pinna Anthropometry Extraction

In the 3D head meshes, the head shape is described by a set of points in the 3 spatial coordinates x, y and z. All the head meshes are aligned with their centre placed in the axes origin defined as in [32, Sec. 1.1]. In Fig. 3, an example of a 3D head mesh is shown. From the figure, it can be noticed that the ear canals lies on the y axis. Therefore, a fixed area around the ear canal can be easily outlined in order to keep only the mesh points representing the pinna and discard the remaining ones. This operation is applied to both left and right pinnae. Then, the two set of points representing the pinnae are independently converted into two range images Ω_s^l and Ω_s^r, where s is the subject. The selected image resolution is 140×160 pixels.

The pinna anthropometry extraction phase follows. This phase carries out two tasks. In the first one, the pinna shape is outlined through the *Active Shape Model* (ASM) algorithm. In the second task, the obtained shape and the range image are employed to extract the pinna anthropometric measurements.

A. Pinna shape fitting through ASM

ASMs are statistical models that build an initial shape model accordingly to a set of training examples and iteratively transforms the model to fit a new object. The object shape is represented by a set of landmark points defined by their x and y coordinates in the image. The first algorithm step is the annotation phase, where a human operator places the landmarks around the boundaries of the considered shape (the pinna in this case), for each subject in the training set T. In Fig. 4, the scheme of the pinna shape model followed

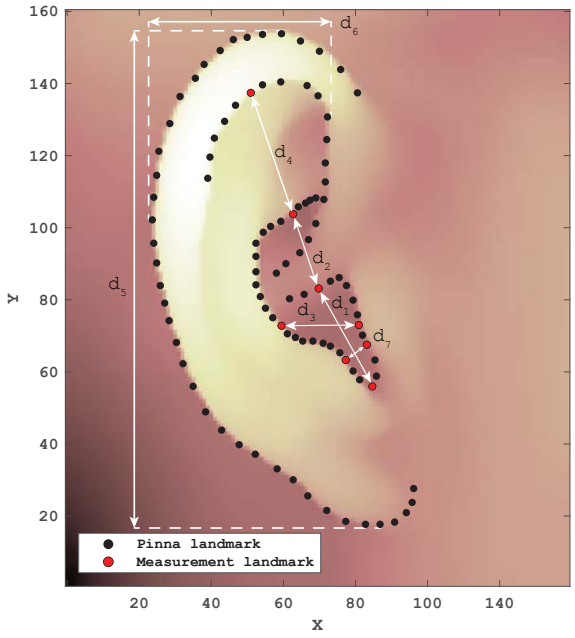

Fig. 4. Example of the schema followed during the pinna shape manual annotation. The landmarks used for anthropometry measurement are highlighted in red. The white arrows show how the parameters are measured exploiting these landmarks.

during the annotation is shown. This is composed by $P = 100$ points that describe the pinna shape and its components. The key guidelines followed in the annotation are the exhaustive description of the pinna shape and the placing of each point in a fixed position corresponding to a specific pinna's part. In this work, all the subject's left pinnae were manually annotated on the pinna range images to form the labelled set. The result of the annotation task is a vector $\mathbf{v}_t$ including the x, y coordinates of the P annotated points for each subject $t \in T$:

$$\mathbf{v}_t = \{x_t^1, y_t^1, \ldots, x_t^P, y_t^P\}.$$

In the ASM fitting phase, these training vectors $\mathbf{v}_t$ are aligned to each other through scale, rotation and translation operations and averaged to obtain the *initial model* $\bar{\mathbf{v}}$. Then, a statistical distribution in a $2P$-dimensional space is assumed over the aligned vectors $\mathbf{v}_t$. PCA is applied to the $\mathbf{v}_t$'s in order to reduce their high dimensional space. Each principal component describes a particular way of how landmarks move together changing the shape. The first γ principal components are selected to approximate the vectors $\mathbf{v}_t$.

The fitting phase of the ASM algorithm follows. The pinna shape vector $\mathbf{v}_s$ for a test subject s is initialized to the initial model $\bar{\mathbf{v}}$. Then, $\mathbf{v}_s$ is iteratively fitted by tuning the weights of the first γ principal components. The weights tuning is performed with the goal of matching the surrounding region for each landmark with the corresponding regions of the training set landmarks. The matching task is achieved by employing *grey profiles*. In the training phase, for each example $t \in T$ and for each of its landmarks (x_t^p, y_t^p), $p \in \{1, \ldots, P\}$, the grey profile $\mathbf{g}_t^p$ is a segment centred in the landmark and perpendicular to the shape line. A Gaussian distribution is assumed over the training set grey profiles derivative, then the mean $\bar{\mathbf{g}}_p$ and the covariance matrix Σ_p are estimated.

In the fitting phase the grey profiles $\mathbf{g}_s^p$ of the new shape $\mathbf{v}_s$ are sampled too. At each iteration of the fitting process, the grey profile $\mathbf{g}_s^p$ for each candidate landmark position is compared through Mahalanobis distance with the corresponding statistical model of the training set with mean $\bar{\mathbf{g}}_p$ and the covariance Σ_p. Then, $\mathbf{v}_s$ is updated with the landmark position minimizing that distance. The algorithm iterates and keeps updating $\mathbf{v}_s$ so that the described shape approaches the new object shape. The algorithm stops when the changes in $\mathbf{v}_s$ between subsequent iterations falls below a given threshold, or when the maximum number of iterations has been reached.

B. Anthropometric measurement

In the pinna anthropometric measurement task, 11 pinna parameters among the ones reported in Fig. 2 are extracted (d_{10} has been ignored in this work due to a lack of a rigorous definition in the documentation):

- d_1: cavum concha height
- d_2: cymba concha height
- d_3: cavum concha width
- d_4: fossa height
- d_5: pinna height
- d_6: pinna width
- d_7: intertragal incisure width
- d_8: cavum concha depth (down)
- d_9: cavum concha depth (back)
- θ_1: pinna rotation angle
- θ_2: pinna flare angle

An automated procedure to measure each of these parameters is proposed. The general approach consists in computing the Euclidean distance between the x and y coordinates of specific landmarks chosen to match the corresponding segments defined in HUTUBS specification. In Fig. 4, the landmarks selected to perform this task are highlighted as red points. This measurement approach is used to obtain values for d_1, d_2, d_3, d_4 and d_7.

For what concern d_5, the pinna's height, a different measurement approach is adopted since defining fixed landmarks would not be a robust method. Therefore, the parameter d_5 for subject s is set to the range of landmarks $\mathbf{v}_s$ along the y coordinate. For parameter d_6, the pinna width, a similar procedure is performed, considering the x coordinate instead of the y one. Further, the landmarks to be considered for the maximum x value (the landmark closest to the face) are only the ones belonging to the internal helix outline.

In order to measure θ_1, the PCA of the vector $\mathbf{v}_s$ is computed to rotate the landmarks according to their variance. The rotation angle applied by PCA corresponds to the complementary angle of θ_1. Thus, by applying the inverse tangent to the PCA coefficients, the pinna rotation angle θ_1, expressed in radians, can be measured. Then, the angle is converted in degrees to match HUTUBS units.

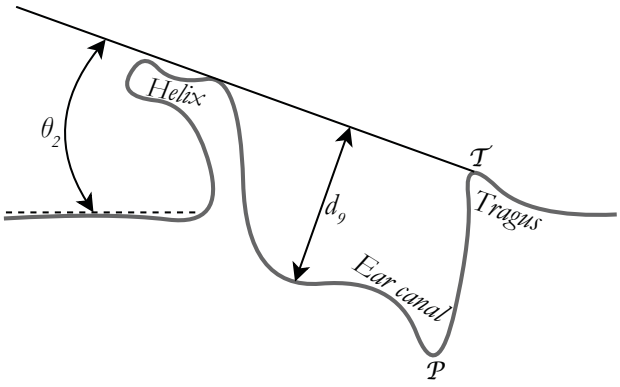

Fig. 5. Pinna horizontal section showing the measurement approaches adopted for d_9 and θ_2.

The θ_2, d_8 and d_9 parameters are measured using the range image, too. Fig. 5 shows θ_2 and d_9 measurement approaches. First, the tragus position $\mathcal{T}$ is found as the peak closest to the ear canal position $\mathcal{P}$ along the x coordinate. Then, a segment in the horizontal plane with one extremity in $\mathcal{T}$ and tangent to the helix is outlined. The θ_2 parameter value is assigned to the angle between the segment and the x axis. The d_9 parameter is set to the maximum distance between the segment joining the tragus and the helix, and the pinna surface.

Finally, in order to automatically measure d_8, a rectangular area R is drawn on the range image. The area R is defined by the vertices $\mathcal{T}$, i.e. the tragus position, and $\mathcal{Z}$. The position $\mathcal{Z}$ corresponds to the position of the landmark used as bottom extremity of d_1 parameter, i.e. the intertragal notch (see Fig. 4). Then, the d_8 parameter is defined as the distance in the 3 dimensions (x, y, z) between $\mathcal{T}$ and the most prominent local minimum of the concha depth found in the area R.

All the measured anthropometric parameters are expressed in pixels, except for θ_1 and θ_2 which are expressed in degrees. Thus, a conversion to some metric unit is required. Since no conversion factors are explicitly provided in HUTUBS, we estimated the centimeters scale ζ, i.e. the value in cm spanned by each pixel. This is set to the value that, when multiplied by the anthropometric measurements in pixels, minimizes the difference with the actual values from HUTUBS.

V. HRTF Individualization

In the *HRTF individualization* phase, a regression model is trained between the anthropometric parameters and the HRTFs. Since this paper is focused on the pinna influence, only median plane HRTFs are considered. This choice was made according to previous evidence showing that the pinna helps the localization task mainly in elevation [4]. HUTUBS reports the HRTFs for $\Phi = 34$ elevation angles in the median plane from $-90°$ to $260°$.

Before training the model, a pre-processing step is undertaken both for anthropometry and HRTFs.

A. Pinna anthropometry pre-processing

The pinna anthropometry pre-processing operation consists in the *z-score* normalization for each parameter. This brings the mean and the standard deviation of each anthropometric parameter to be 0 and 1, respectively.

Let $A = \{d_1, \dots d_9, \theta_1, \theta_2\}$ be the set of the 11 considered anthropometric parameters, and let $a \in A$ be any of those parameters. Then, a is normalized as follows:

$$a \leftarrow \frac{a - \mu_a}{\sigma_a}, \tag{1}$$

where μ_a and σ_a are the mean and standard deviation of a, respectively.

B. HRTF pre-processing

Instead of training the regression model on the HRTF as is, the HRTF is first decomposed in *Directional Transfer Function* (DTF) and *Common Transfer Function* (CTF) [34]. Given an HRTF set H_s^ϕ for a subject s at elevation index $\phi \in 1, \dots, \Phi$, H_s^ϕ can be represented as follows:

$$H_s^\phi(f) = C_s(f) D_s^\phi(f), \tag{2}$$

where C and D are the CTF and DTF, respectively, and f is the frequency. The CTF represents the HRTF components that are common for all source directions. Therefore the CTF includes the ear canal resonance, the transfer functions of the devices used in the HRTF recording (e.g., microphone and speakers) and all the components that remain unchanged as the angle varies. In the DTF, the directional components remain, i.e. the ones specific to the source direction and thus needed for the localization perception.

The CTF and DTF are computed with the method reported in [35], [36]. The CTF magnitude $|C_s|$ is computed as the average of the HRTF magnitude $|H_s|$ for all directions:

$$|C_s|(f) = \exp\left(\frac{1}{\Phi}\sum_{\phi=1}^{\Phi} \log\left|H_s^\phi(f)\right|\right). \tag{3}$$

The CTF phase $\angle C_s$ is reconstructed via minimum-phase approximation of the CTF amplitude spectrum. Then, the complex DTF D_s^ϕ is computed dividing the complex HRTF H_s^ϕ and CTF C_s for each elevation ϕ:

$$D_s^\phi(f) = \frac{H_s^\phi(f)}{C_s(f)}. \tag{4}$$

C. Regression

Two different regression models were tested, namely a Multiple Linear Regression (MLR) model and a *Generalized Regression Neural Network* (GRNN) [37]. Due to better performances of the latter, only the evaluation with GRNN is reported here (see [38] for details). GRNN is a variation of radial basis neural networks. These kind of neural networks use a Radial Basis Function (RBF) as activation function, such as the Gaussian kernel here employed. GRNNs are single-pass neural networks, that is they are not based on training

algorithms such as back-propagation, but their parameters are directly determined from training data. This feature allows GRNNs to be trained quickly and to be able to yield good performances even with limited training data. The only hyper-parameter of GRNN is the spread parameter σ. In this work, σ has been set to 1.3 for all the trained models.

The regression focuses on the magnitude of DTF $|D|$ and CTF $|C|$. Since the DTF varies over Φ elevations, a different GRNN model $\mathfrak{R}_D^\phi$ is trained for each elevation ϕ in order to predict each DTF $\left|D^\phi\right|$. The model $\mathfrak{R}_D^\phi$ takes as input the 11 anthropometric parameters in the set A and maps them into the $F = 128$ frequency bins of the DTF at elevation ϕ. In order to reconstruct the HRTFs from the predicted DTFs, the CTF is needed, too. Therefore, an independent regression model $\mathfrak{R}_C$ is trained between the anthropometry and the CTF magnitude $|C|$. Once both DTF and CTF have been modelled, they can be combined to obtain the predicted HRTF. While the HRTF magnitude is predicted with the described method, the HRTF phase is reconstructed via minimum-phase approximation. This choice follows the common practice of approximating a HRTF by means of a minimum-phase function cascaded with a linear phase or a pure delay, with the latter approximating the interaural time delay [1, Secs. 3.1.3, 5.2.3]. Moreover, in this work we consider median plane HRTFs only, for which interaural time delays are negligible.

VI. Evaluation and Discussion

The method evaluation is performed on the 58 subjects of HUTUBS for which all the required data are available: 3D head meshes, anthropometric measurements, and HRTFs.

A. Pinna anthropometry extraction evaluation

All the left pinnae of the 58 subjects were manually annotated with landmarks. In order to evaluate the ASM, a *Leave-One-Out Cross-Validation* (LOOCV) was employed. Since the manual landmarks are available only for the left pinnae, the LOOCV is performed only on these images.

The performance of the pinna anthropometry extraction approach is evaluated by means of two errors: the landmark fitting error LE, and the anthropometry measurement error AE. The former is the error made by the ASM in placing the landmarks, and is defined as the Euclidean distance between the positions of the annotated and predicted landmarks, for each landmark p and subject s:

$$LE_s^p = \zeta d(\mathbf{v}_s^p, \hat{\mathbf{v}}_s^p), \tag{5}$$

where ζ is the centimeter scale, d is the Euclidean distance, $\mathbf{v}_s^p$ is the manually annotated position of landmark p for subject s, and $\hat{\mathbf{v}}_s^p$ is the corresponding position estimated with ASM. The LOOCV results in a mean landmark fitting error LE, averaged for all subjects and for all landmarks, of 0.24 cm with a standard deviation of 0.14 cm.

Then, the pinna anthropometry estimation error AE is evaluated. This is defined as the difference between the actual anthropometric parameters values reported in HUTUBS and those estimated with the automated procedure proposed here:

$$AE_s^a = a_s - \hat{a}_s, \tag{6}$$

where a_s is the actual value of the anthropometric parameter a for the subject s, while $\hat{a}_s$ is the corresponding estimated value. Moreover, we also consider the absolute error $|AE|$, defined similarly to (6) but computing the absolute difference.

Table I reports the performances of the proposed measurement procedure. For each parameter a, in addition to AE^a and $|AE^a|$ the table reports the anthropometric mean absolute error $|AEM^a|$ between the actual anthropometric parameters in the dataset and those estimated using the manually placed landmarks. Along with each error column, the standard deviation and the relative percentage error are reported. The relative percentage error is the percentage value of the error with respect to the parameter mean.

The mean absolute errors $|AE|$ estimated with the automated measurement procedure range from 0.09 cm for d_2 to 0.5 cm for d_5. However, considering the relative percentage absolute errors, d_5 is one of the parameters with the lowest percentage error along with d_6, while θ_1 has the worst measurement performances with an absolute percentage error of 25.9%. Relative percentage absolute errors range from 6.9% to 25.9%, with a mean of 15.9%.

These errors can be considered acceptable if compared with the ones available in the literature for similar tasks. The relative percentage errors for pinna parameters reported by Torres-Gallegos et al. [18], range from 1.1% to 23.8% with a mean of 10.6%. Dinakaran et al. [39], employing a different measurement approach, report a mean percentage absolute error ranging from 4.15% to 16.17% with a mean of 11.1%. Although the performances of the method here proposed appear to be slightly worse than the ones reported in [18], [39], some additional considerations are in order. The cited studies are based on the CIPIC measurements and they consider only a subset of the pinna parameters: from d_1 to d_7 in [18] and from d_1 to d_6 in [39]. The proposed method has the advantage to estimate all the pinna parameters defined in HUTUBS (except for d_{10}), including the angles (θ_1 and θ_2) and depth (d_8 and d_9) parameters which are usually not measured. Considering only the parameters from d_1 to d_7 the proposed method has a mean percentage absolute error of 13.2%, that is comparable to the state-of-the-art results. Furthermore, the above studies rely on self-created datasets where the anthropometry ground truth was measured by the authors themselves. Instead, in this work the ground truth is extracted by a third-part. As consequence, the error of the automatic measurement procedure includes the interpretation error of the HUTUBS anthropometric measurements definition.

A possible indicator of this interpretation error is represented by the last column of Table I, i.e. the mean absolute error $|AEM|$ between the anthropometric parameters reported in HUTUBS and the ones estimated with the manually annotated pinna landmarks. The $|AEM|$ values are comparable to the $|AE|$ values: for the parameters d_1, d_3, d_4, d_7, d_8 and θ_1, $|AEM|$ is only slightly lower than $|AE|$; on the other hand,

TABLE I
RESULTS OF THE LOOCV FOR THE ANTHROPOMETRIC MEASUREMENT PROCEDURE.

Parameter name a [unit]	Parameter a $\mu_a \pm \sigma_a$	AE $\mu_a \pm \sigma_a$ (%)	$\lvert AE \rvert$ $\mu_a \pm \sigma_a$ (%)	$\lvert AEM \rvert$ $\mu_a \pm \sigma_a$ (%)
d_1 [cm]	1.80 ± 0.16	0.16 ± 0.16 (9.0)	0.18 ± 0.13 (10.0)	0.16 ± 0.11 (9.0)
d_2 [cm]	0.99 ± 0.12	0.01 ± 0.12 (1.0)	0.09 ± 0.07 (9.4)	0.14 ± 0.09 (13.9)
d_3 [cm]	1.75 ± 0.20	0.37 ± 0.18 (21.0)	0.37 ± 0.18 (21.0)	0.35 ± 0.17 (20.1)
d_4 [cm]	2.07 ± 0.24	0.31 ± 0.20 (14.8)	0.32 ± 0.18 (15.4)	0.29 ± 0.25 (14.0)
d_5 [cm]	6.09 ± 0.38	-0.48 ± 0.24 (-7.8)	0.50 ± 0.19 (8.2)	0.57 ± 0.22 (9.3)
d_6 [cm]	2.95 ± 0.25	-0.06 ± 0.24 (-2.2)	0.20 ± 0.14 (6.9)	0.24 ± 0.17 (8.0)
d_7 [cm]	0.62 ± 0.14	0.10 ± 0.14 (16.3)	0.13 ± 0.11 (21.3)	0.12 ± 0.10 (19.8)
d_8 [cm]	1.15 ± 0.14	-0.16 ± 0.22 (14.0)	0.24 ± 0.12 (20.8)	0.23 ± 0.13 (20.2)
d_9 [cm]	1.19 ± 0.12	-0.16 ± 0.15 (-13.4)	0.18 ± 0.12 (15.3)	0.18 ± 0.12 (15.3)
θ_1 [°]	11.49 ± 5.18	-2.79 ± 2.64 (-25.3)	2.97 ± 2.42 (25.9)	2.82 ± 2.26 (24.5)
θ_2 [°]	25.25 ± 7.83	-5.15 ± 4.21 (-20.4)	5.19 ± 4.15 (20.6)	5.19 ± 4.15 (20.6)

for the parameters d_2, d_5 and d_6 the error $|AEM|$ is even higher than $|AE|$. The remaining parameters d_9 and θ_2 are not based on the pinna landmarks, hence $|AE|$ and $|AEM|$ are equivalent. The comparison between $|AE|$ and $|AEM|$ suggests that further improvements should be done in the manual annotation phase and in the automated measurement procedure, in order to match the HUTUBS measurement definitions as closely as possible.

In Fig. 6 the distributions of the relative percentages anthropometric errors AE^a for each parameter a are shown using a violin plot. This plot, along with the AE column in Table I, helps to figure out how the measurement procedure performs across the subjects for each parameter. Fig. 6 confirms that the best performances are achieved for the height parameter d_5 where the error is distributed close to 0% ($AE^{d_5} = -7.8\%$) and with short tails. This consideration may be extended to the d_1, d_2 and d_6 parameters, where $AE^{d_1} = 9\%$, $AE^{d_2} = 1\%$ and $AE^{d_6} = -2.2\%$. Then, there are some parameters with short tails but with a mean error significantly different from 0%, such as d_3, d_4 and d_9 where $AE^{d_3} = 21\%$, $AE^{d_4} = 14.8\%$ and $AE^{d_9} = -13.4\%$. For the remaining parameters the tails are longer, especially for θ_1 where the error distribution reaches values close to -100%. In conclusion, the AE distributions show that some parameters are underestimated, such as d_3, d_4 and d_7, while other parameters are overestimated, such as d_8, θ_1 and θ_2. In case of a distribution with an offset and a limited tail, the measurement may be still considered acceptable if the offset can be shifted closer to 0% (e.g., by improving the manual annotation phase). Instead, when the tail is long, even with small offsets, the error is more likely to be attributed to inherent limitations of the proposed approach.

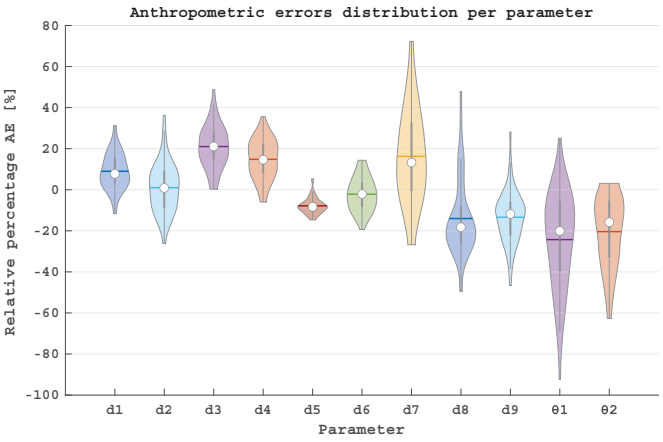

Fig. 6. Violin plot of the relative percentage anthropometric error AE distributions for each parameter. The white points represent the medians. The horizontal lines colored with the same hue as the violin plot area but darker represent the means. The vertical bold grey lines represent the interquartile range.

B. HRTF individualization evaluation

In the evaluation of HRTF individualization both left and right pinnae are considered, and each pinna is treated as a single instance. Thus $2S = 116$ examples are available. All the regression models were evaluated via k-fold cross-validation with $k = 4$. For the evaluation, the anthropometry preprocessing described in Sec. V-A was performed separately for each training set before applying the regression model. The HRTFs were estimated from the predicted DTFs and CTFs. The error between the actual HRTF and the estimated one is measured in decibels (dB) with *Spectral Distortion* (SD). The SD between the HRTFs H_1 and H_2 is defined as follows:

$$SD^{\phi}(H_1, H_2) = \sqrt{\frac{1}{f_h - f_l} \sum_{f=f_l}^{f_h} \left(20 \log \frac{\left| H_1^{\phi}(f) \right|}{\left| H_2^{\phi}(f) \right|} \right)^2}, \quad (7)$$

where ϕ is the elevation angle, while f_l and f_h are the lower and higher frequency bounds, respectively.

The SD is a widely used metric in the HRTF individualization field and it provides a measure of how much the magnitude spectrum of H_2 differs from the one of H_1 in dB.

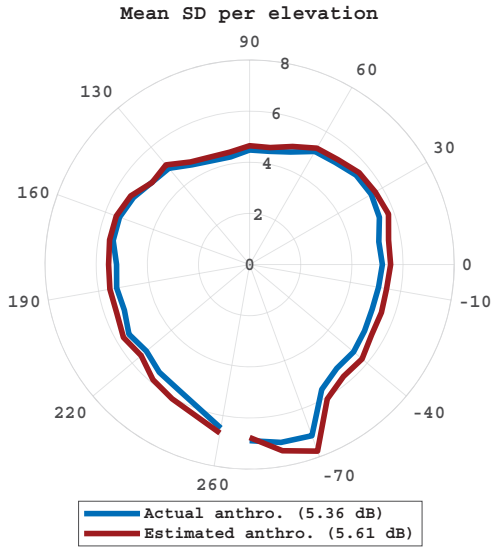

Fig. 7. Plot in polar coordinate system of SD^{ϕ}_{act} in blue and SD^{ϕ}_{est} in red. The radius represents the SD value and the polar angle represents the elevation angle ϕ.

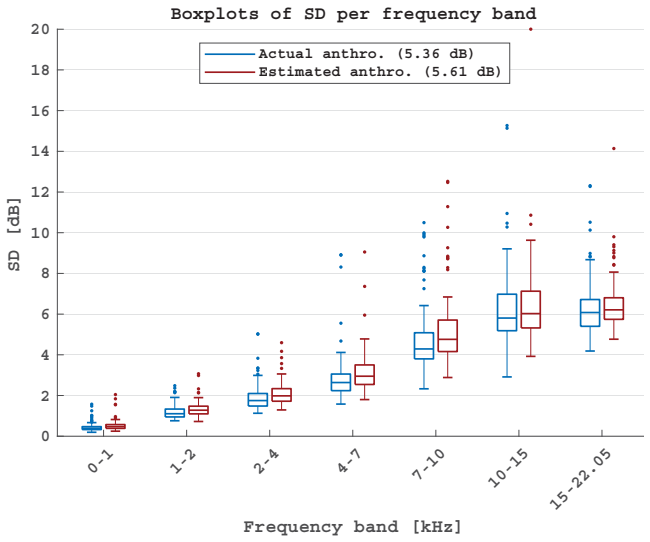

Fig. 8. Box plots of spectral distortion in dB for each frequency band b compared for the HRTFs predicted from the actual HUTUBS anthropometry in blue (SD_{act}) and the ones predicted from estimated anthropometry in red (SD_{est}).

The proposed HRTF individualization method is evaluated by computing the SD between the actual HRTFs available in HUTUBS and two different HRTF estimations: the first one is the HRTFs predicted from the actual anthropometric parameters reported in HUTUBS; the second estimation is based on the anthropometry extracted with the automated procedure described in Section IV. The corresponding spectral deviations are denoted as SD_{act} and SD_{est}, respectively.

The SD is averaged across subjects and elevation angles to obtain a mean SD value. The mean value of SD_{act} is 5.36 dB while the mean value of SD_{est} is 5.61 dB. Figure 7 shows instead the values SD^{ϕ}_{act} and SD^{ϕ}_{est} averaged across subjects only for each elevation angle ϕ, in a polar plot. The figure shows that the performances reach their best in the top area of the median plane, while as we move down in the median plane the SDs increase. This behaviour is shared by SD^{ϕ}_{act} and SD^{ϕ}_{est} but, it can be noticed that in the bottom area of the median plane SD^{ϕ}_{est} have higher values than SD^{ϕ}_{act}. However, a t-test, conducted on SD^{ϕ}_{act} and SD^{ϕ}_{est} distributions, showed that their difference is not statistically significant ($p = 0.14$).

Additional insights can be gained by looking at how SD varies in frequency. Fig. 8 shows the box plots of the distributions across the subjects of SD^{b}_{act} and SD^{b}_{est} for each frequency band b. The selected frequency bands are 0–1, 1–2, 2–4, 4–7, 7–10, 10–15 and 15–22.05 kHz. A positive correlation between the SDs and the frequency band can be noticed. This confirms that the estimation of HRTFs in the high-frequency region is a harder task than for lower frequencies. In Table II, the mean values of these distributions are shown.

These results are comparable with those reported in previous literature. In [18], using a best anthropometry match approach, the authors reported SD values of 3.3, 4.6, 6.0 and 6.2 dB for the frequency bands starting at 0 Hz and with upper limits of 3.4, 8, 17 and 22-0.5 kHz, respectively; this method is based on a subset of CIPIC anthropometric parameters describing pinna as well as head and torso. The method proposed in [27] reports the SD values of 1.3, 1.8, 2.2, 3.5 and 5.8 dB for the frequency bands 0.2–1, 1–2, 2–4, 4–8 and 8–15 kHz.

TABLE II
SPECTRAL DISTORTION IN DB FOR EACH FREQUENCY BAND COMPARED FOR THE HRTFS PREDICTED FROM THE ACTUAL HUTUBS ANTHROPOMETRY (SD_{ACT}) AND THE ONES PREDICTED FROM ESTIMATED ANTHROPOMETRY (SD_{EST}).

	Frequency band [kHz]						
	0–1	1–2	2–4	4–7	7–10	10–15	15–22.05
SD_{act} [dB]	0.44	1.19	1.92	2.85	4.74	6.21	6.29
SD_{est} [dB]	0.52	1.33	2.11	3.14	5.23	6.47	6.47

Finally, Fig. 9 shows an example of the HRTFs predicted with actual and estimated anthropometry compared with the individual HRTFs for a subject in four selected angles.

C. Auditory model evaluation

Although SD is one of the most employed metrics in HRTF individualization studies, it remains a poor indicator of the subject perceptual performances. Therefore, a virtual localization experiment through an auditory model is performed. The selected auditory model was developed by Baumgartner et al. in 2013 [40]; it is a template-based paradigm where the internal representation of the incoming sound (e.g., the individualized HRTF) is compared with a reference template (the individual HRTF). The auditory model works in two stages. The first stage models the effect of human physiology (from the body to the inner ear) while focusing on directional cues

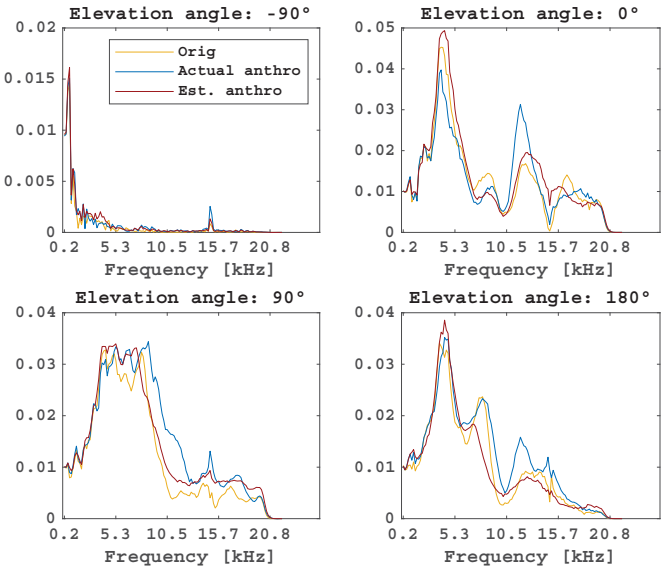

Fig. 9. Individual and estimated HRTFs for HUTUBS subject 5.

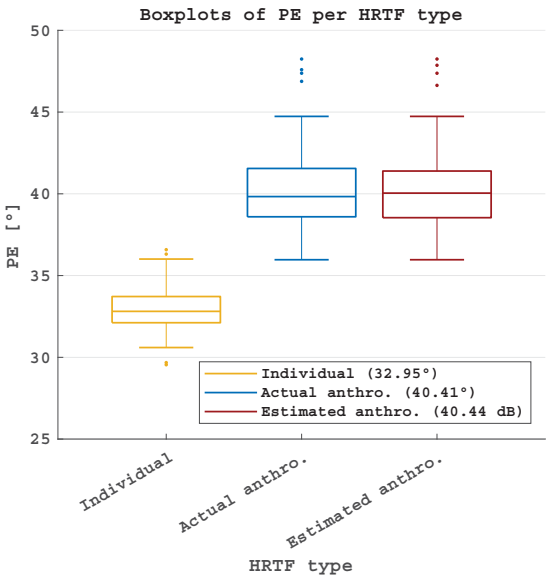

Fig. 10. Box plots of PE for each type of HRTF.

in order to create an internal representation of the incoming sound. Then, the comparison stage computes the *Internal-Spectral Differences* (ISDs), i.e. the differences between the internal representations of the sound and the template, for each template angle and for each frequency. The ISDs are mapped to polar-response probabilities, denoted as *Similarity Indices* (SIs). SIs represent the probability that the virtual subject points to a specific target angle.

One of the metrics on the localization performance returned by the auditory model is the *local polar RMS error* PE_j defined for each elevation response j as [41]:

$$PE_j = \sqrt{\frac{\sum_{i \in L} (\phi_i - \varphi_j)^2 p_j[\phi_i]}{\sum_{i \in L} p_j[\phi_i]}}, \tag{8}$$

where $L = \{i \in N : 1 \leq i \leq N_\phi, |\phi_i - \varphi_j| \mod 180° < 90°\}$ defines the local elevation responses with respect to the local response ϕ_i and the target position φ_j. The probability mass vector $p_j[\phi_i]$ is the subject probability to respond with angle ϕ_i. The PE averaged for each j represents an estimation of the subject error in degrees in the localization task.

A virtual localization experiment on the median plane between $-40°$ and $220°$ was conducted. In the experiment, the PE was measured for each subject using three types of HRTF: PE_{ind} is the polar error for the individual HRTFs, while PE_{act} and PE_{est} are the errors for the HRTFs individualized using actual and estimated anthropometry, respectively. In Fig. 10, the box plots of the PE distributions for each type of HRTF are shown. As expected, the errors for the individual HRTFs have the lowest mean, that is $32.95°$. The distributions of PE_{act} and PE_{est} have means that are $7.46°$ and $7.49°$ higher than the one of PE_{ind}. However, the distributions of PE_{act} and PE_{est} are very similar, and a t-test confirmed that their difference is not statistically significant ($p = 0.33$).

VII. Conclusion

A method for HRTF individualization based on pinna anthropometry has been proposed. Rather than starting from available anthropometry, a widely used approach but inadequate for end-user applications, the method provides an automated procedure to extract a set of relevant pinna measures from a 3D mesh. Specifically, all the pinna parameters defined in HUTUBS (except from d_{10}) are estimated, while only a subset of parameters based on length is usually considered in the literature. The estimated anthropometry is used to train a regression model that predicts the HRTF. Although the measurement procedure introduces some errors, the evaluation showed that the errors of the HRTFs predicted from actual and estimated anthropometry do not have a statistically significant difference. The proposed method achieves performances comparable with the state of the art, both for pinna anthropometry estimation and HRTF individualization.

The analysis on landmark fitting performed by the ASM algorithm may be expanded, particularly to assess whether the estimation is worse for some measurement landmarks than others (see Fig. 4). Additionally, future work should include anthropometric parameters related to head and torso for the estimation of HRTFs even outside the median plane. Moreover, a key future investigation will be the evaluation of the proposed method through listening tests comparing localization accuracy of human subjects using their individual HRTFs compared with the HRTF predicted by the proposed method and, possibly, with an HRTF recorded from a dummy head. Results from such tests will complement objective metrics, such as the SD and the auditory model employed here.

Acknowledgment

This work is part of SONICOM, a project that has received funding from the European Union's Horizon 2020 research and

innovation programme under grant agreement No 101017743.

REFERENCES

[1] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display*. J. Ross Publishing, 2013.

[2] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.

[3] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *Journal of the Audio Engineering Society*, vol. 44, no. 6, pp. 451–469, 1996.

[4] M. B. Gardner and R. S. Gardner, "Problem of localization in the median plane: effect of pinnae cavity occlusion," *The Journal of the Acoustical Society of America*, vol. 53, no. 2, pp. 400–408, 1973.

[5] B. G. Shinn-Cunningham, N. I. Durlach, and R. M. Held, "Adapting to supernormal auditory localization cues. i. bias and resolution," *The Journal of the Acoustical Society of America*, vol. 103, no. 6, pp. 3656–3666, 1998.

[6] R. Bomhardt, *Anthropometric Individualization of Head-Related Transfer Functions Analysis and Modeling*, vol. 28. Logos Verlag Berlin GmbH, 2017.

[7] C. Guezenoc and R. Seguier, "HRTF individualization: A survey," in *Proc. AES Convention*, (New York), Oct. 2018.

[8] B. F. Katz, "Boundary element method calculation of individual head-related transfer function. i. rigid model calculation," *The Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 2440–2448, 2001.

[9] Y. Kahana and P. A. Nelson, "Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models," *Journal of Sound and Vibration*, vol. 300, no. 3-5, pp. 552–579, 2007.

[10] S. Prepeliță, M. Geronazzo, F. Avanzini, and L. Savioja, "Influence of voxelization on finite difference time domain simulations of head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2489–2504, 2016.

[11] S. Hwang, Y. Park, and Y.-s. Park, "Modeling and customization of head-related impulse responses based on general basis functions in time domain," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 965–980, 2008.

[12] B. F. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. EL99–EL105, 2012.

[13] K. Yamamoto and T. Igarashi, "Fully perceptual-based 3d spatial sound individualization with an adaptive variational autoencoder," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, pp. 1–13, 2017.

[14] J. Fels and M. Vorländer, "Anthropometric parameters influencing head-related transfer functions," *Acta Acustica united with Acustica*, vol. 95, no. 2, pp. 331–342, 2009.

[15] M. Zhang, R. Kennedy, T. Abhayapala, and W. Zhang, "Statistical method to identify key anthropometric parameters in hrtf individualization," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, pp. 213–218, IEEE, 2011.

[16] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pp. 99–102, IEEE, 2001.

[17] D. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "Hrtf personalization using anthropometric measurements," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*, pp. 157–160, Ieee, 2003.

[18] E. A. Torres-Gallegos, F. Orduna-Bustamante, and F. Arámbula-Cosío, "Personalization of head-related transfer functions (hrtf) based on automatic photo-anthropometry and inference from a database," *Applied Acoustics*, vol. 97, pp. 84–95, 2015.

[19] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1480–1492, 1999.

[20] T. Nishino, N. Inoue, K. Takeda, and F. Itakura, "Estimation of hrtfs on the horizontal plane using physical features," *Applied Acoustics*, vol. 68, no. 8, pp. 897–908, 2007.

[21] H. Hu, L. Zhou, J. Zhang, H. Ma, and Z. Wu, "Head related transfer function personalization based on multiple regression analysis," in *2006 International Conference on Computational Intelligence and Security*, vol. 2, pp. 1829–1832, IEEE, 2006.

[22] G. Grindlay and M. A. O. Vasilescu, "A multilinear (tensor) framework for hrtf analysis and synthesis," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 1, pp. I–161, IEEE, 2007.

[23] X.-Y. Zeng, S.-G. Wang, and L.-P. Gao, "A hybrid algorithm for selecting head-related transfer function based on similarity of anthropometric structures," *Journal of Sound and Vibration*, vol. 329, no. 19, pp. 4093–4106, 2010.

[24] R. Miccini and S. Spagnol, "Hrtf individualization using deep learning," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 390–395, IEEE, 2020.

[25] H. Hu, L. Zhou, H. Ma, and Z. Wu, "Hrtf personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, vol. 69, no. 2, pp. 163–172, 2008.

[26] L. Li and Q. Huang, "Hrtf personalization modeling based on rbf neural network," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3707–3710, IEEE, 2013.

[27] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein, "A manifold learning approach for personalizing hrtfs from anthropometric features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 559–570, 2016.

[28] G. W. Lee and H. K. Kim, "Personalized hrtf modeling based on deep neural network using anthropometric measurements and images of the ear," *Applied Sciences*, vol. 8, no. 11, p. 2180, 2018.

[29] T.-Y. Chen, T.-H. Kuo, and T.-S. Chi, "Autoencoding hrtfs for dnn based hrtf personalization using anthropometric features," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 271–275, IEEE, 2019.

[30] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[31] B. Fabian, D. Manoj, P. Robert, W. Jan Joschka, S. Fabian, V. Daniel, G. Peter, and W. Stefan, *The HUTUBS head-related transfer function (HRTF) database*, 2019.

[32] 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," *Journal of the Audio Engineering Society*, vol. 67, no. 9, pp. 705–718, 2019.

[33] P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, H. Ziegelwanger, *et al.*, "Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions," in *Audio Engineering Society Convention 134*, Audio Engineering Society, 2013.

[34] J. C. Middlebrooks and D. M. Green, "Directional dependence of interaural envelope delays," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2149–2162, 1990.

[35] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *The Journal of the Acoustical Society of America*, vol. 136, no. 2, pp. 791–802, 2014.

[36] P. Majdak, M. J. Goupell, and B. Laback, "3-d localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Attention, Perception, & Psychophysics*, vol. 72, no. 2, pp. 454–469, 2010.

[37] D. F. Specht *et al.*, "A general regression neural network," *IEEE Transactions on Neural Networks*, vol. 2, no. 6, pp. 568–576, 1991.

[38] D. Fantini, "Individualized binaural rendering through pinna anthropometry extraction from 3d images," Master's thesis, University of Milan, 2019.

[39] M. Dinakaran, P. Grosche, F. Brinkmann, and S. Weinzierl, "Extraction of anthropometric measures from 3d-meshes for the individualization of head-related transfer functions," in *Audio Engineering Society Convention 140*, Audio Engineering Society, 2016.

[40] R. Baumgartner, P. Majdak, and B. Laback, "Assessment of sagittal-plane sound localization performance in spatial-audio applications," in *The technology of binaural listening*, pp. 93–119, Springer, 2013.

[41] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini, "Improving elevation perception with a tool for image-guided head-related transfer function selection," in *Proc. of the 20th Int. Conference on Digital Audio Effects (DAFx-17)*, pp. 397–404, 2017.