IEEE Signal Processing Society
IEEE Open Journal of
Signal Processing

Received 5 August 2024; revised 11 November 2024; accepted 4 December 2024. Date of publication 10 January 2025;
date of current version 4 February 2025. The review of this article was arranged by Associate Editor Romain Serizel.

Digital Object Identifier 10.1109/OJSP.2025.3528330

A Survey on Machine Learning Techniques for Head-Related Transfer Function Individualization

DAVIDE FANTINI, MICHELE GERONAZZO (Senior Member, IEEE), FEDERICO AVANZINI, AND STAVROS NTALAMPIRAS

[1]Laboratory of Music Informatics (LIM), Department of Computer Science, University of Milan, 20133 Milan, Italy
[2]Department of Engineering and Management, University of Padua, 35122 Padova, Italy
[3]Dyson School of Design Engineering, Imperial College London, SW7 2AZ London, U.K.

CORRESPONDING AUTHOR: DAVIDE FANTINI (email: davide.fantini@unimi.it).

This work was supported by SONICOM, through European Union's Horizon 2020 Research and Innovation Programme under Grant 101017743.
This article has supplementary downloadable material available at https://doi.org/10.1109/OJSP.2025.3528330, provided by the authors.

ABSTRACT Machine learning (ML) has become pervasive in various research fields, including binaural synthesis personalization, which is crucial for sound in immersive virtual environments. Researchers have mainly addressed this topic by estimating the individual head-related transfer function (HRTF). HRTFs are utilized to render audio signals at specific spatial positions, thereby simulating real-world sound wave interactions with the human body. As such, an HRTF that is compliant with individual characteristics enhances the realism of the binaural simulation. This survey systematically examines the HRTF individualization works based on ML proposed in the literature. The analyzed works are organized according to the processing steps involved in the ML workflow, including the employed dataset, input and output types, data preprocessing operations, ML models, and model evaluation. In addition to categorizing the works of the existing literature, this survey discusses their achievements, identifies their limitations, and outlines aspects that require further investigation at the crossroads of research communities in acoustics, audio signal processing, and machine learning.

INDEX TERMS HRTF individualization, machine learning, spatial audio, binaural synthesis.

I. INTRODUCTION

Machine learning (ML) can be defined as the learning of algorithms to solve a specific problem based on information extracted from previous experiences or events, rather than explicitly programming the algorithm [1]. ML has become pervasive in several aspects of society over the past few years, with both industrial and scientific applications. The field of spatial audio is no exception. Spatial audio techniques find several applications, including video gaming, teleconferencing, art, flight simulation [2], devices for blind people [3], and audio production [4]. An appropriate spatial audio simulation involves the simulation of the spatial cues used by humans to localize sound sources in space. These spatial cues originate from the interactions between the human body and the sound waves, which result in position-dependent sound alterations. Head-related transfer functions (HRTFs) describe these spatial cues as a linear time-invariant (LTI) system for each sound source position of interest and for each ear. The use of an HRTF of a specific position to spatialize an audio signal spatialized with an HRTF of a specific position through headphones artificially creates the sensation of a sound source in that position. HRTFs are individual due to their close relationship with anatomical traits. Therefore, the use of an HRTF non-compliant with the individual anatomy, i.e., a non-individual HRTF, results in an improper spatial audio experience [5], [6], [7], [8], [9], [10], [11], [12]. Non-individual HRTFs are prevalent in end-user applications due to the practical limitations of accessing individual HRTFs.

Consequently, several methods for HRTF individualization, or personalization, have been proposed in the literature to obtain an estimation of the individual HRTF.

Despite the significant relevance of HRTF individualization for spatial audio technologies, this area remains poorly standardized in research. Different researchers have proposed their methods without a common validation procedure, which would include a reliable dataset, robust objective metrics, rigorous perceptual tests, and so on. This makes it difficult to compare different approaches. In this survey, we provide a demonstration of the aforementioned lack of standardization, and despite that, we organized the related literature in view of future standardization actions. Although there are various methods for HRTF individualization, this survey focuses on ML-based approaches. ML approaches can potentially overcome the limitation of traditional HRTF individualization methods, which can be time-consuming, limited in accuracy, and require input data far from being user-friendly. When properly trained, ML models are able to extrapolate patterns between input and output data, thereby enabling their generalization to unseen data. In addition, an estimated HRTF can be generated in a relatively short time once the model is trained. Recent advancements in ML can also facilitate the HRTF prediction from input data that are readily accessible to end-users, such as pictures. However, ML-based methods require careful training and validation to achieve good and unbiased performances.

Some HRTF individualization surveys have already been published. Several publications have been dedicated to the broad field of HRTF individualization, including articles [13], [14], [15], book chapters [16] [17, Ch. 7], and a Ph.D. thesis [18]. However, none of these works specifically addresses ML. Other publications have focused on the measurement and the numerical simulation of individual HRTFs [19], [20], the role of ML for spatial audio capture, processing, and reproduction [21], and for HRTF dimensionality reduction, categorization, interpolation in addition to HRTF individualization [22]. This survey fills the existing gaps by focusing specifically on ML-based approaches along with the rigorous and formal characterization of data-driven approaches in the processing and evaluation stages.

This survey is organized as follows. After the present introduction (Section I), Section II provides an overview of the basic concepts related to the HRTF individualization field. This includes the definition of HRTF and its related representations, an introduction to the individualization problem, an overview of the main datasets used in this field, and the outline of the ML workflow for HRTF individualization. Section III describes the research methodology conducted to obtain a comprehensive overview of the existing HRTF individualization publications based on ML. Then, these publications are categorized according to the steps of the ML workflow, which includes the input data (Section IV) and their preprocessing (Section V), the output data (Section VI) and their preprocessing (Section VII), the ML models (Section VIII) along with their training and validation approaches (Section IX), and their evaluation metrics (Section X). Section XI presents a discussion of the trends observed in the analyzed publications, emphasizing their limitations, and proposing avenues for future research. Section XII concludes the survey.

II. CONCEPTS AND CATEGORIZATION

A. HRTF DEFINITION

The spatial cues encoded by the HRTF are primarily influenced by three body parts, namely the torso, head and pinnae, each causing different effects. The shape and size of the head affect the time and intensity differences between the sounds received by the two ears, which represent essential spatial cues for binaural hearing [23]. These differences are known as interaural time difference (ITD) and interaural level, or intensity, difference (ILD, or IID), and are crucial in determining the azimuth angle of a sound source. Conversely, the monaural spectral modifications caused by the elevation-dependent filtering effects of the body components are prevalent for elevation perception. The torso, head, and pinnae influence the HRTF spectral structure in different frequency ranges. With regard to azimuth localization, the head influences ILD above 1.5 kHz and ITD below this frequency [17, Sec. 1.4]. Torso reflections affect elevation localization below 3 kHz, especially for low elevation angles, whereas the pinna influence is prevalent between 3–4 and 14–15 kHz circa [23], [24], [25], [26], [27] [17, Sec. 3.4]

HRTFs describe both binaural (ITD, ILD) and monaural (spectral modifications) spatial cues as an LTI system. The HRTF set for a subject is a collection of transfer functions (or impulse responses), one for each sound source position of interest and for each ear. The HRTFs H_L and H_R for the left and right ears describe the sound modifications caused by the human body according to the source position (distance r, azimuth θ, elevation ϕ), the frequency f, and the anatomical characteristics a of the subject [17, Sec. 1.5]:

$$H_{\{L,R\}}(r,\theta,\phi,f,a) = \frac{P_{\{L,R\}}(r,\theta,\phi,f,a)}{P_0(r,f)}, \tag{1}$$

where P_R and P_L are the sound pressure at the left and right ears, whereas P_0 is the free sound pressure in the head center without the head. HRTFs can be equivalently represented by head-related impulse responses (HRIRs) in the time domain. Fig. 1 shows an example of an HRTF and the corresponding HRIR in the median and horizontal planes. However, further representations exist to account for part of the information encoded by the HRTF. For instance, the directional transfer function (DTF) can be extracted from HRTF to isolate the directional components, whereas the direction-independent components (e.g., ear canal resonance, equipment responses, etc.) are represented by the common transfer function (CTF) [28]. The HRTF H at source direction $s = \{\theta, \phi\}$, and frequency f can be decomposed into DTF and CTF as follows [17, Sec. 7.3.2]:

$$H(s,f) = CTF(f)DTF(s,f), \tag{2}$$

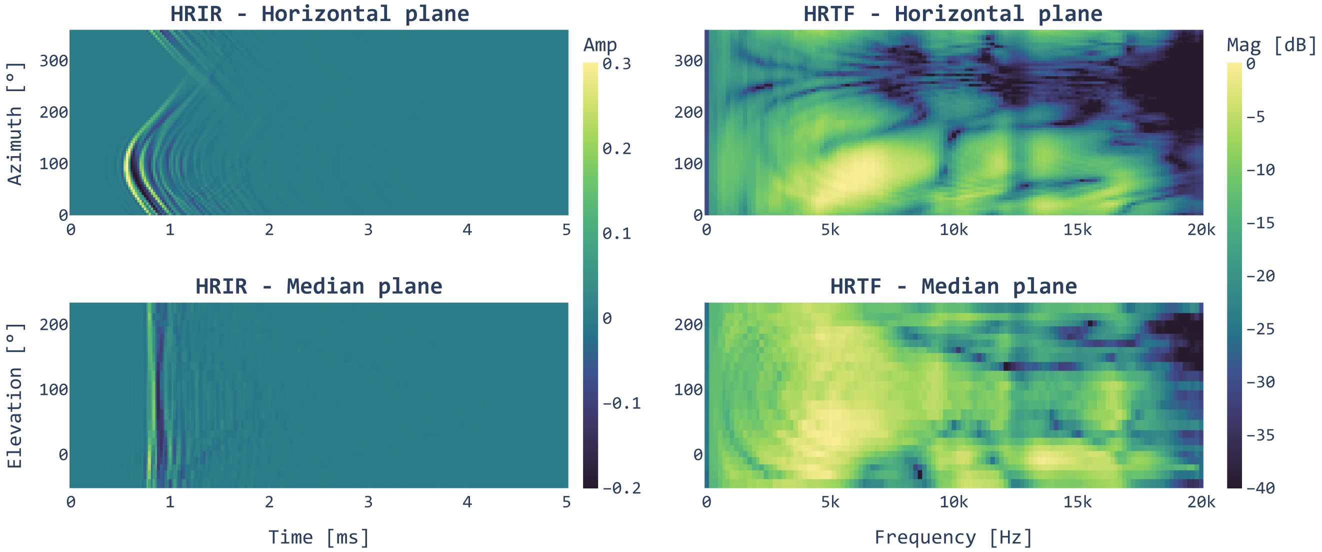

FIGURE 1. Example of HRIRs (left) and HRTFs (right) in the horizontal (top) and median plane (bottom). The example presented here corresponds to the acoustic measurement of a KEMAR dummy head, as reported in the SONICOM dataset [29].

where the CTF magnitude is defined as the root mean square of H averaged across the S source directions:

$$|CTF(f)| = \sqrt{\frac{1}{S}\sum\nolimits_{s=0}^{S-1} |H(s,f)|^2}. \tag{3}$$

Another representation that can be extracted from HRTF is the pinna-related transfer function (PRTF), which encodes the sole influence of the pinna. A PRTF can be directly obtained by isolating the pinna in the acoustic measurement [30] or in the numerical simulation [31]. Alternatively, the PRTF can be partially extracted from HRIR at ipsilateral source positions with a time window that eliminates the influence of the torso and shoulders [26], [32].

B. THE INDIVIDUALIZATION PROBLEM

The considerable influence of anatomy on HRTF denotes its individuality, which is a crucial factor. Individual HRTFs are seldom employed in end-user applications due to the impracticality of their acoustic measurements. These require expansive equipment, time-consuming recording sessions, and experienced personnel. HRTFs are acoustically measured by placing a microphone inside each ear canal. An excitation signal, ideally an impulse, is then reproduced through sound sources placed in the positions of interest, and the microphones capture the impulse response. HRTFs are recorded around the subject using a spherical grid with a radius of typically 1 to 2 meters. The grid's spatial resolution varies and can differ for azimuth and elevation. The azimuth plane is typically fully covered with a resolution between 2.5° and 10°, whereas the lowest elevation angles are neglected because measuring the HRTF underneath the subject presents practical difficulties. Due to the impracticality of measuring HRTFs, a non-individual, or generic, HRTF is often employed, disregarding the subject's individual anatomy. Generic HRTFs are typically recorded using dummy heads that represent the average anatomical characteristics of a certain population [40], [41]. However, utilizing generic HRTFs can result in several drawbacks including front-back and up-down confusions, degradation of accuracy in elevation perception, and lack of externalization [5], [6], [7], [8], [9], [10]. In addition, perceptual aspects other than simple localization may be affected [11], [12]. Due to the difficulties of measuring individual HRTFs and the limitations of non-individual HRTFs, several studies have focused on HRTF individualization, or personalization. An HRTF individualization method estimates the individual HRTF without direct measurements but by retrieving and exploiting other subject-specific information that is correlated with the acoustic characteristics of the personal HRTF. Examples of this kind of information include anthropometric measurements, 3D head scans, and subjective auditory feedback. In contrast, the outcome of HRTF individualization methods is an HRTF set that has been either retrieved from a dataset or generated from scratch.

Traditional methods for HRTF individualization can be broadly grouped into numerical simulation, selection-based, and adaptation approaches [15]. Numerical simulation approaches provide an approximate solution to the wave equation with boundary conditions determined by head, torso, and pinnae represented by 3D scans [20]. They represent an accurate approach as they provide HRTFs having similar spectra [42] and localization performances [43] to acoustically measured ones. Nevertheless, some perceptual differences exist [37, Sec. 3.3] [43], [44]. Numerical simulation has further drawbacks, including the high accuracy required for the 3D scans, the need for scan postprocessing, and the intensive computational load. Selection-based approaches provide a best-match HRTF by finding the subject with the most similar characteristics in a dataset. These characteristics are

TABLE 1. Some of the Publicly Available HRTF Datasets Along With Their Characteristics

Name	N. subjects	Numerically simulated	N. directions	N_θ	N_φ	Anthropometry	Pictures	3D meshes	HpTF
Itakura Lab. Dataset [33]	111	No	72	72	1	80 (KEMAR)	No	No	No
CIPIC [34]	45	No	1250	50	25	43 (CIPIC)	No	No	No
LISTEN [35]	51	No	187	≤ 24	10	50 (CIPIC)	No	No	No
Chinese pilots [36]	58	No	723	≤ 73	13	Yes (CIPIC, GJB 4856-2003)	No	No	No
HUTUBS [37, 38]	96	Also	440	≤ 72	19	93 (CIPIC, Self-defined)	No	58 (head)	96 (HD800S) 64 (HD650)
CHEDAR [39]	1253	Yes	≤ 2522	72	≤ 36	Yes (CIPIC, Self-defined)	No	Yes (head, shoulders)	No
WiDESPREaD [31]	1005	Yes	≤ 2562	72	≤ 36	No	No	Yes (pinna)	No
SONICOM [29]	200 (ongoing)	No	793	72	12	No	Yes (RGB, Depth)	Yes (head, shoulders)	Yes (HD650)

This Table is a reduced version of Table A.1in the supplementary materials.

typically represented by anthropometric parameters [45], [46] or subjective feedback obtained with a listening test [47], [48]. Selection-based methods are quite simple, but limited in effectiveness since they require a sufficiently representative database as they are unable to generalize the relationship between the input and the HRTF. Thus, the selected HRTF is always an approximation. Adaptation approaches adjust a non-individual HRTF according to the characteristics of the test subject. Similarly to selection-based approaches, the subject's characteristics can be represented by anthropometric parameters [28] or subjective feedback [49], [50]. Adaptation approaches have been less investigated in the literature and found limited applicability due to their limitations. For instance, several adaptation approaches assume that only the anatomical size varies across subjects and disregard the highly individual anthropometric characteristics. In addition, methods based on subjective feedback, either selection-based or adaptation approaches, require time-consuming sessions in which the subject is asked to engage with several HRTFs.

C. HRTF DATASETS

HRTF datasets are collections of HRTF sets of human subjects and/or dummy heads. These datasets represent the ground truth to train and evaluate ML models used for HRTF individualization. The datasets used for ML should be of adequate size, of high quality, and representative of real-world data [51], [52]. Table 1 presents some of the HRTF datasets collected to date along with their characteristics. A more comprehensive version of this table is represented by Table A.1in the supplementary materials. Throughout the survey and in the table, we employ vertical-polar coordinates, where the azimuth θ is defined between 0° and 360°, whereas the elevation φ is defined between -90° (bottom) and 90° (top). HRTF datasets typically consist of acoustically measured HRTFs, although some are composed of numerically simulated HRTFs computed from 3D head or pinna meshes. In addition to HRTFs, these datasets may also include anthropometry, pictures, 3D meshes, which can be employed as input for HRTF individualization, and headphone transfer functions (HpTFs).

One of the earliest HRTF datasets is CIPIC [34], which remains the most frequently used dataset for HRTF individualization, despite the existence of larger and more recent ones. Also, CIPIC includes an anthropometric specification that has been largely adopted during the design of subsequent datasets. As shown in Table 1, HRTF datasets typically include less than 100 subjects, as HRTF measurement is impractical. The limited size of the datasets represents a challenge for the training of complex ML models, such as deep neural networks. However, datasets of numerically simulated HRTFs represent an exception, as the acoustic measurement is not needed. For instance, CHEDAR [39] and WiDESPREaD [31] include more than 1000 subjects.

D. ML WORKFLOW FOR HRTF INDIVIDUALIZATION

An HRTF individualization task is defined as the estimation of the individual HRTF based on input data that provides meaningful information on the corresponding subject. Therefore, HRTF individualization is categorized in the ML paradigm of supervised learning, where a model is trained to predict the desired output given the input data. In particular, it can be considered a regression task, as the output, i.e., the HRTF, assumes continuous values. In the scope of this survey, we consider an HRTF individualization method as ML-based if the core of the method is the data-driven training of a supervised learning model with the goal of generalizing the relationship between input and output. The input encompasses

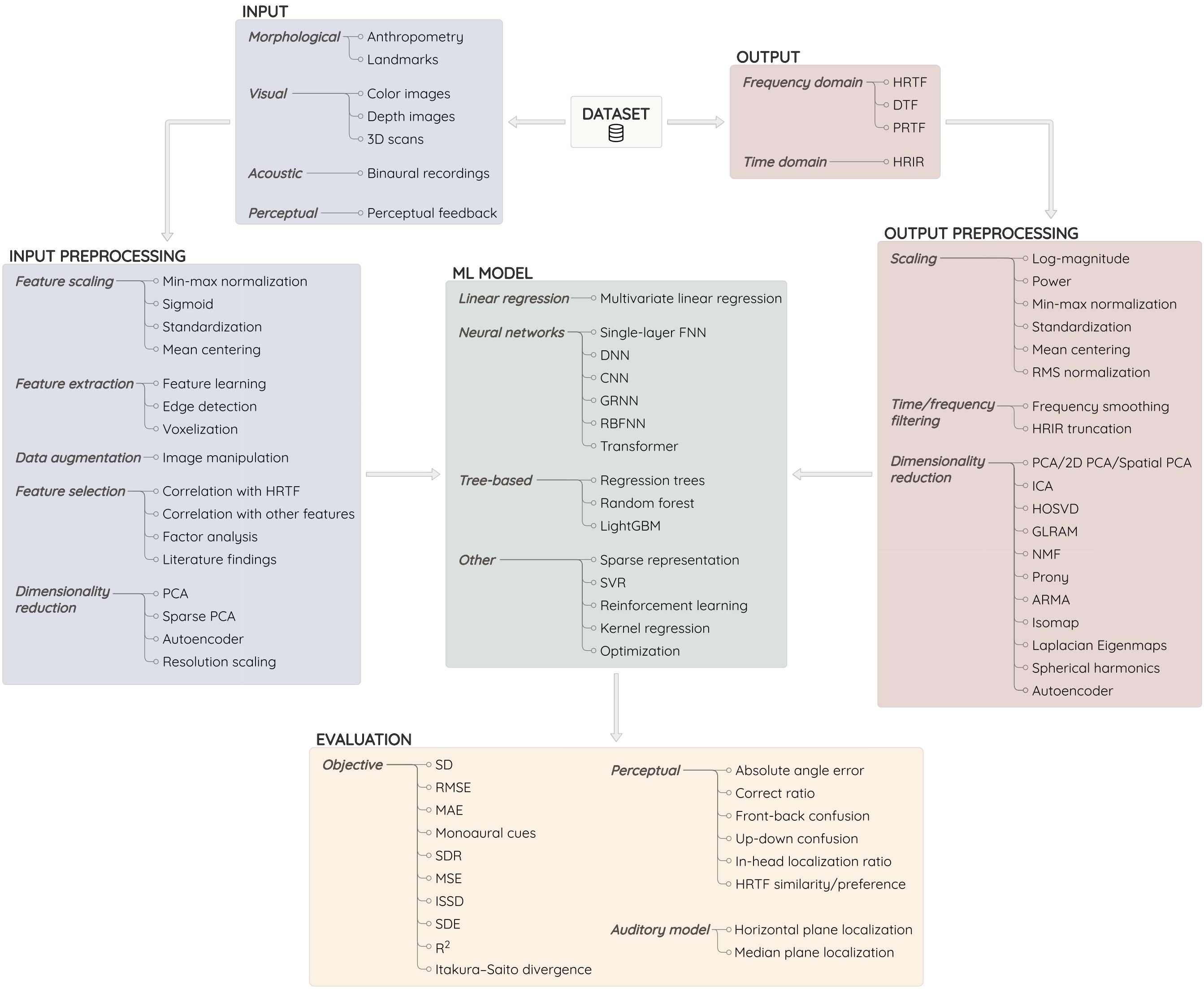

FIGURE 2. ML workflow for HRTF individualization with the main options adopted by the analyzed studies.

any information that exhibits a correlation with the HRTF. The output may be the HRTF response in either the time or frequency domains, or a low-dimensional representation thereof. In this survey, we disregarded HRTF individualization methods that use ML to predict a different type of output or to perform different tasks.

Fig. 2 shows the typical ML workflow followed by HRTF individualization methods, along with the characterizations encompassed in each step by the publications analyzed in this survey. In the following, we present a conceptual overview of these steps, while the remainder of the survey delves into the approaches employed by the analyzed literature studies for each step.

D. INPUT

A variety of input data types can be utilized, which can be broadly grouped into four categories: morphological, visual, perceptual, and binaural. Morphological data, such as anthropometry and landmarks, and visual data, such as images and 3D scans, capture the individual anatomical traits that influence HRTF. Methods relying on perceptual feedback are designed to directly optimize the subjective auditory experience. Finally, binaural recordings encode individual spatial cues, although they are measured in uncontrolled environments, in contrast to HRTFs.

D. OUTPUT

The output of ML models for HRTF individualization is the personalized HRTF. However, one can consider different HRTF representations, such as HRIR, DTF, or PRTF. Methods working in the frequency domain usually focus on the sole magnitude, with the phase information being disregarded. This choice is justified by the possibility of approximating the phase by means of a minimum-phase function cascaded with a pure delay simulating the ITD [53].

TABLE 2. Keywords Used in the Research Queries to Identify the Publications in the Online Databases

Category	Keywords
Field	HRTF, HRTFs, "Head Related Transfer Function", "Head-Related Transfer Function", "Head Related Transfer Functions", "Head-Related Transfer Functions", HRIR, HRIRs, "Head Related Impulse Response", "Head-Related Impulse Response", "Head Related Impulse Responses", "Head-Related Impulse Responses"
Task	individual*, personal*, estimat*, model*, predict*, custom*, recommed*, learn*
Method	"deep learning", "reinforcement learning", "neural network", "neural networks", "deep network", "deep networks", NN, NNs, "NN-based", DNN, DNNs, "DNN-based", regression, *linear, "sparse representation"

The asterisk (*) wildcard refers to zero or more unknown characters. The exact queries are provided in the supplementary materials.

D. INPUT/OUTPUT DATA PREPROCESSING

This is a step to ensure effective training of ML models. A typical preprocessing operation for data used in HRTF individualization is feature scaling, which involves mapping different features within consistent value ranges. Further, feature selection and dimensionality reduction can be used to retain only the relevant information.

D. ML MODEL

Once the data have been pre-processed, an ML algorithm is selected to train one or more regression models. The number of ML models depends on how the multidimensional structure of HRTF data is handled as ML algorithms do not natively support such a complex structure. In the training of ML models, an important choice is the strategy employed to split the dataset into training set, test set, and possibly, validation set. The latter is typically used to evaluate the trained models while tuning the model's hyperparameters.

D. EVALUATION

After training the ML model, the following step is the evaluation of the model's performance. In the context of HRTF individualization, the evaluation of trained models can be objective, perceptual, or based on auditory models. The objective evaluation of an estimated HRTF is typically quantified by spectral distortion (SD), also known as log-spectral distortion (LSD). SD measures the deviation in decibels (dB) between the magnitudes of the ground truth HRTF H and its estimation $\hat{H}$. The SD at azimuth θ and elevation φ averaged for the F frequency bins is computed as follows:

$$SD(\theta, \varphi) = \sqrt{\frac{1}{F} \sum_{f=1}^{F} \left(20 \log \frac{|H(\theta, \varphi, f)|}{|\hat{H}(\theta, \varphi, f)|}\right)^2} \text{[dB]}. \quad (4)$$

Further common objective metrics include the root mean square error (RMSE), which can also be computed for HRIRs in the time domain, and the signal-to-distortion ratio (SDR).

In perceptual experiments conducted to evaluate HRTF individualization methods, the localization performances of subjects using the estimated HRTF are typically analyzed. Alternatively to perceptual experiments, computational auditory models can be employed to predict the localization responses of a simulated subject with a given HRTF.

III. RESEARCH METHODOLOGY

The research and screening of literature studies for this survey were conducted in accordance with the PRISMA methodology [55]. The publications were identified by defining a research query in the following online databases: ACM Digital Library[1], Elsevier Scopus[2], and IEEE Xplore[3]. The query was constructed using the keywords reported in Table 2, which were grouped according to the acoustics field of interest (HRTF), the performed task (individualization), and the method (ML). The keywords were connected with the OR operator within each group and the AND operator between groups. However, the keywords related to the task and method groups were connected with the OR operator for the research in the title. This was necessary because titles contain fewer words than abstracts. The exact queries used for each online database are provided in the supplementary materials. The identification through the queries was conducted to satisfy the following inclusion criteria:

I1. Publications included in journal papers, conference proceedings, or magazines (books, book chapters, Ph.D. theses, and extended abstracts were excluded)
I2. English-language publications
I3. Publications up to November 2024 (query execution date)
I4. Publications proposing an HRTF individualization method employing ML techniques for the prediction of HRTF-related information, including HRTF magnitude, HRIR, PRTF, and DTF

Fig. 3 shows the PRISMA flow diagram for the literature research conducted for this survey. The identification process yielded 858 records, of which 257 duplicates were subsequently removed. The remaining 601 records were screened according to the following exclusion criteria:

E1. Unavailable publications
E2. Non-peer reviewed publications

[1] https://dl.acm.org/search/advanced
[2] https://www.scopus.com/search/form.uri?display=advanced
[3] https://ieeexplore.ieee.org/search/advanced/command

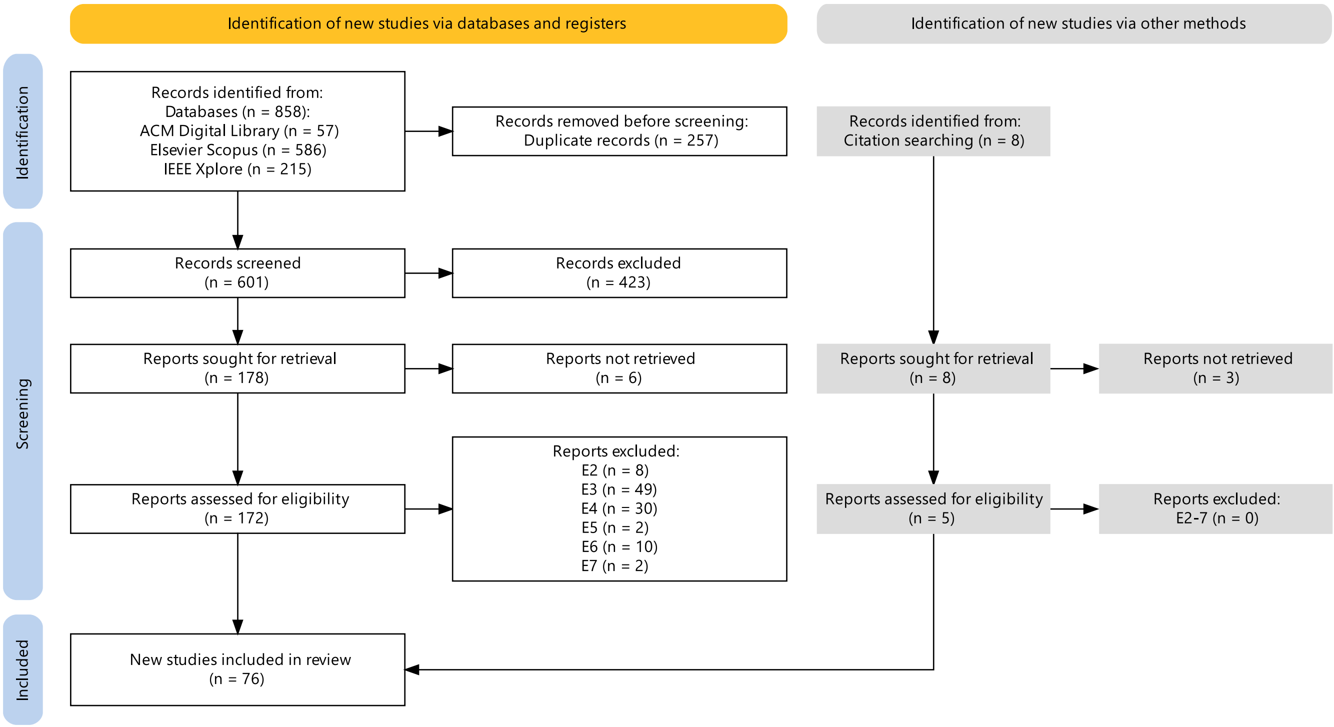
Identification of new studies via databases and registers
Identification of new studies via other methods
Identification
Screening
Included
Records identified from: Databases (n = 858): ACM Digital Library (n = 57) Elsevier Scopus (n = 586) IEEE Xplore (n = 215)
Records removed before screening: Duplicate records (n = 257)
Records identified from: Citation searching (n = 8)
Records screened (n = 601)
Records excluded (n = 423)
Reports sought for retrieval (n = 178)
Reports not retrieved (n = 6)
Reports sought for retrieval (n = 8)
Reports not retrieved (n = 3)
Reports assessed for eligibility (n = 172)
Reports excluded: E2 (n = 8) E3 (n = 49) E4 (n = 30) E5 (n = 2) E6 (n = 10) E7 (n = 2)
Reports assessed for eligibility (n = 5)
Reports excluded: E2-7 (n = 0)
New studies included in review (n = 76)

FIGURE 3. PRISMA flow diagram of the research and screening of literature studies on HRTF individualization based on machine learning [54].

E3. Methods for tasks related but different from HRTF individualization such as (*a*) HRTF upsampling, (*b*) HRTF dimensionality reduction, (*c*) HRTF clustering, (*d*) HRTF filter modeling, (*e*) anthropometry automatic measurement, (*f*) pinna mesh modeling followed by numerical simulation to compute HRTF, (*g*) sound localization automatic prediction, (*h*) HRTF calibration for auditory localization improvement and (*i*) perceptual studies on HRTFs

E4. HRTF individualization methods not based on ML

E5. Methods to estimate ITD and ILD

E6. Methods for HRTF selection relying on ML to analyze the HRTF spectrum's peaks and notches, the anthropometry, and the perceptual outcomes of given HRTFs

E7. Surveys, reviews, datasets, and similar publications on HRTF individualization

In accordance with the PRISMA flow diagram, a preliminary screening was conducted with the titles and abstracts of the publications being analyzed. This screening resulted in the exclusion of 423 records. In addition, other 8 records were identified through a citation search among the references of the identified publications. Then, 9 publications were excluded since unavailable. The resulting 177 records were then assessed for eligibility by full-text screening. Applying the exclusion criteria E2-7, we finally included 76 publications to be examined in this survey. Fig. 4 shows the temporal distribution of these publications. Furthermore, Fig. 5 depicts the occurrences in these publications of the different options of the ML workflow for (a) the HRTF datasets (b) the input types, (c) the output dimensionality reduction techniques, and (d) the ML models.

IV. INPUT

A variety of input data types can be employed to estimate the individual HRTF. Fig. 5 shows the distribution of the input types for the analyzed publications. These data can be broadly grouped into four categories: morphological, visual, perceptual, and binaural. The following sections provide a detailed examination of each of these types of input data.

A. MORPHOLOGICAL DATA

1) ANTHROPOMETRY

The majority of HRTF individualization methods rely on anthropometric parameters that are measured for head, torso, and pinnae. These parameters are typically defined as distances between specific points on the body, although angle measurements are also employed. Anthropometry can be manually measured (e.g., with rulers) or digitally extracted from images or 3D scans. Alternatively, there are methods for automatic measurement [56], [57], [58], [59], [60], [61].

Various anthropometric specifications have been proposed to describe the relationship between anatomy and HRTF [17, Ch. 7]. Anthropometric specifications are body measurements that are typically depicted in two-dimensional sketches. Currently, there is still no complete and inter-independent set of anthropometric parameters that fully describe the HRTF. Despite several studies have investigated the influence of various body components on HRTF [25], [62], [63], [64], the exact influence of anthropometry on HRTF remains a topic of discussion. The KEMAR mannequin's design included an

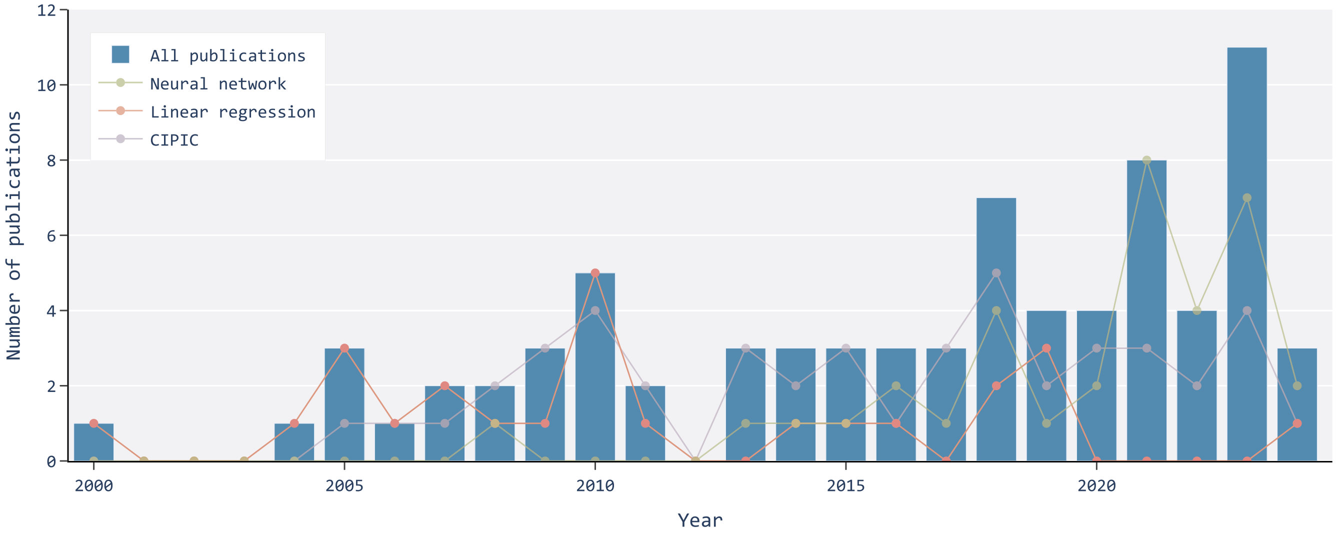

FIGURE 4. Temporal distribution of the publications on ML-based HRTF individualization analyzed in this survey. The number of publications using CIPIC, neural networks, and linear regression are also shown.

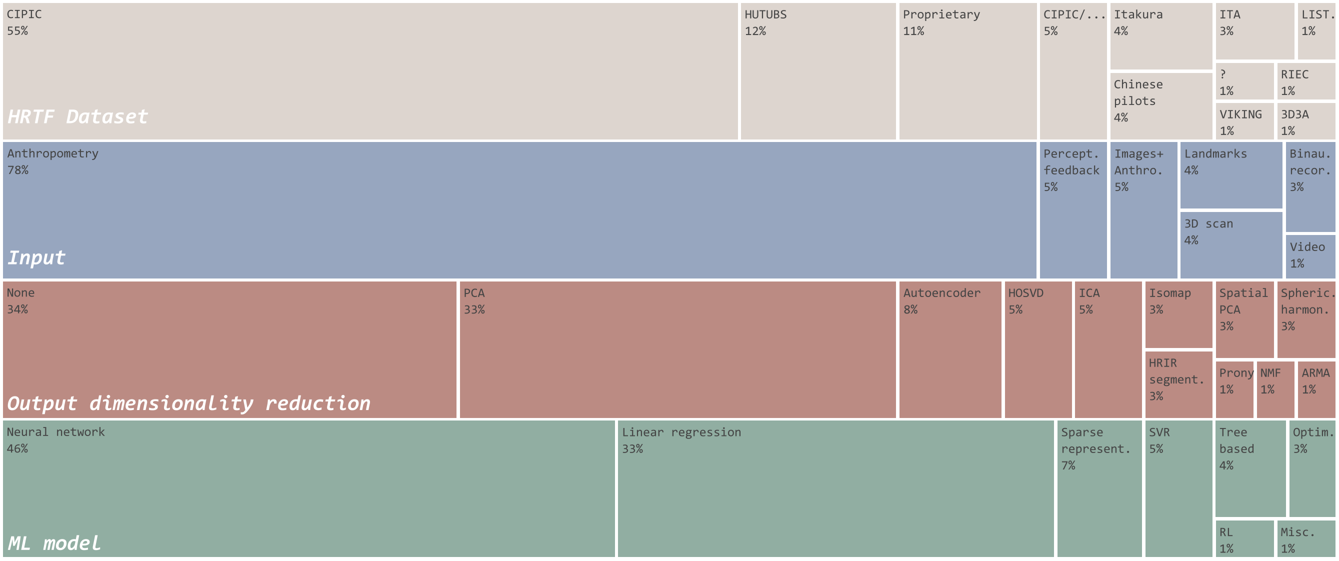

FIGURE 5. Distribution of HRTF datasets, input types, output dimensionality reduction techniques, and ML models for the publications on ML-based HRTF individualization analyzed in this survey.

early proposal for anthropometric specifications, which consisted of ten parameters for the head and torso, and 13 for the pinna [41]. Another specification proposed by Middlebrooks [28] included six pinna parameters, whereas Iida et al. [65] measured ten distances between the tragus and other points on the pinna. Some HRTF datasets included anthropometric parameters defined in national standards, which were not designed for HRTFs. For instance, the HRTF dataset recorded by Xie et al. [66] included anthropometry in accordance to the standard GB/T 2428-1998 [67], whereas the Chinese pilots dataset [36] followed the standard GJB 4856-2003 [68].

In 2001, Algazi et al. [34] proposed an anthropometric specification for the CIPIC dataset, which included 17 head and torso parameters, and ten pinna parameters (see Fig. 6). This specification remains the prevalent one in HRTF datasets and HRTF individualization tasks. Modifications to the CIPIC anthropometric specification have been proposed. Two additional pinna parameters were proposed within the HUTUBS dataset [37], [38], whereas the CHEDAR dataset [39] included five anthropometric parameters derived from the CIPIC ones and two new parameters. Additionally, a subset of the pinna control points proposed by Stitt and Katz [69] is based on CIPIC's specification set. In addition to classical anthropometry, other types of morphological features include area-related parameters. Teng and Zhong [70] proposed five areas of the pinna in addition to a subset of the HUTUBS anthropometry

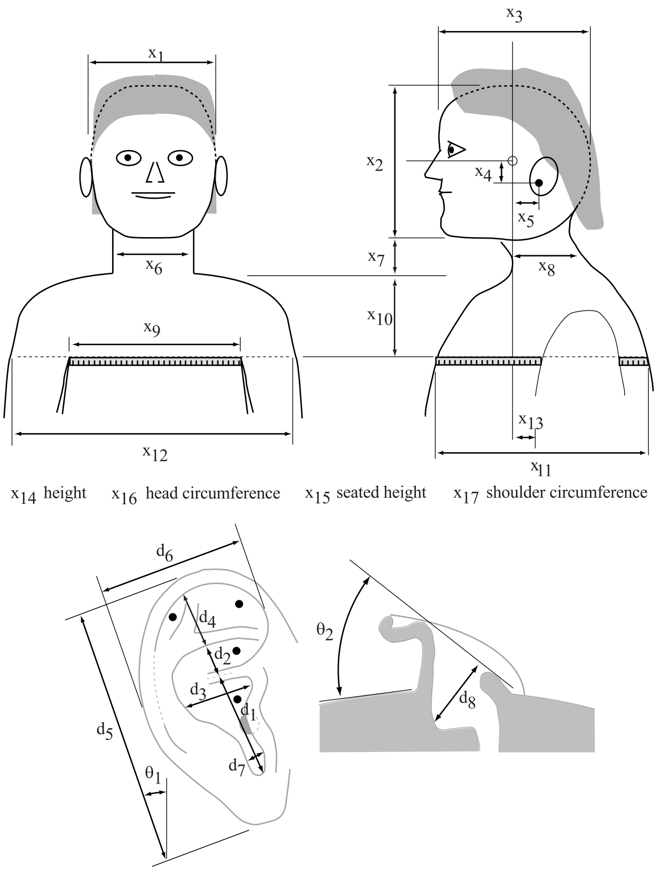

FIGURE 6. CIPIC anthropometric specification adapted from [34].

as input to predict the HRTF magnitude using a random forest. However, the study does not provide any detail regarding the measurement of the area-related parameters. Recently, a wider set of pinna anthropometric parameters, including distances, angles, areas, volumes and depths, was proposed and evaluated in an HRTF selection task showing improved performances over the sole HUTUBS parameters [71].

2) LANDMARKS

Landmarks represent a type of input data that is similar to anthropometry but not strictly defined by specifications. Landmarks are defined as the 2D or 3D coordinates of points located on specific parts of the human body. Jin et al. [72] applied principal component analysis (PCA) on both DTF and a set of 20 3D landmarks positioned on the torso, head, and pinnae. The authors performed a stepwise multivariate linear regression (MLR) to map the principal component (PC) weights of the landmarks to the DTF ones. Nevertheless, the performances of the method were evaluated only on the training set and with a preliminary localization test on a single subject. Lu et al. used 3D landmarks to predict the HRTF using sparse representation [73] and the HRIR using neural networks (NNs) [74], [75]. Landmarks have also been placed on pinna images to automatically measure anthropometric parameters [59], [60], [61], [71].

B. VISUAL DATA

Similarly to anthropometry, visual data, such as 2D pictures or 3D head and pinna scans, encode morphological information. However, they exhibit a higher dimensionality and require increased effort to extract useful information.

1) IMAGES

The recent development of deep learning techniques for images has also contributed to the HRTF individualization field. Several hybrid methods using pinna images and anthropometry have been proposed. Lee and Kim [76] trained two sub-networks with anthropometry and pinna images as input, respectively. A third sub-network was trained to predict the HRIR from the former two sub-networks. Zhao et al. [77] extracted low-dimensional features from pinna images with transfer learning from the VGG19 network using the AWE dataset of ear pictures [78]. They trained a convolutional neural network (CNN) to predict the spherical harmonics (SH) coefficients of the HRTF using low-dimensional features from VGG19in addition to head and torso anthropometry. Both Lee and Kim [76] and Zhao et al. [77] reported lower mean SD values for the HRTFs estimated using pinna anthropometry—3.69 and 5.31 dB, respectively—than those obtained using pinna images—4.47 and 5.4 dB. However, despite worse results, the effort of manually extracting pinna anthropometry is not required when using pinna images. Miccini and Spagnol [79] extracted latent representations of pinna images and HRTFs using a variational autoencoder (VAE) and a conditional VAE (CVAE), respectively. Subsequently, they trained a deep neural network (DNN) to map these two latent spaces. This approach considered only the pinna modeling, with other body parts modeled independently according to the *mixed structural model* paradigm [80]. The authors reported inconclusive results for both SD and evaluation with a computational auditory model [81] in comparison to a generic HRTF.

2) 3D SCANS

In addition to 2D images, researchers have also investigated the use of 3D scans for HRTF individualization. Ko et al. [82] used depth images derived from the HUTUBS 3D head meshes to predict the PRTF magnitude with a CNN. They reported a mean SD of 5 dB, improved over an existing method [79] and a generic HRTF. Zhou et al. [83] predicted the HRTF from voxelized 3D pinna meshes comparing CNN and UNet architectures. The two architectures yielded similar results on a numerically simulated HRTF dataset. In addition, they reported lower errors compared to anthropometry-based methods [84], [85], although the use of simulated HRTFs could have affected the results. Zhao et al. [86] proposed a neural network-based approach to predict HRTF from 3D head meshes. A neural network was trained to learn a feature vector describing the anthropometric structure of the 3D meshes. The feature vector was employed to predict the HRTFs for each vertical plane at once by considering the spectral correlation and continuity across adjacent sampling grids and frequencies. They reported a mean SD of 3.78 dB and improved results over a generic HRTF according to localization metrics computed with an auditory model [81].

C. PERCEPTUAL FEEDBACK

Perceptual user feedback represents an alternative input type to morphological data. Methods based on such input directly optimize the subjective auditory experience, as opposed to learning the relationship between anatomy and HRTF. Perceptual-based methods using ML often depend on extracting low-dimensional HRTF features. This is typically achieved using autoencoders. Luo et al. [87] simulated a virtual user—represented by a Gaussian process regression model—localizing sounds given a query HRTF generated from the low-dimensional space. Then, a recommendation system estimated the best generated HRTF as the one that minimizes the error between the target spatial positions and those provided by the virtual user. They reported improved SDR using an autoencoder over PCA for HRTF dimensionality reduction. Yamamoto and Igarashi [88] collected subjects' ratings about sound localization of pairs of HRTFs. The ratings were used to obtain optimized personal weights between a subject and each HRTF. These weights were used to generate individualized HRTFs through an autoencoder along with latent variables and the desired space position. In a perceptual test, the majority of the 20 recruited participants preferred the estimated HRTF over the best-match HRTF in the dataset. A limitation of the proposed approach is the duration of the session to collect the user's feedback, which could last up to 30 minutes. Hwang et al. [89] extracted 12 PCs from the median plane HRIRs of CIPIC. Subjects were then asked to tune the first three PCs, which were used as input to predict the remaining ones using MLR.

D. BINAURAL RECORDINGS

Some researchers estimated the individual HRTF based on binaural recordings obtained by placing a pair of in-ear microphones in the subject's ear canals. This can be achieved using earbuds with integrated microphones or similar devices. However, the relative position between the microphone and the sound source must be known. These recordings are carried out in-the-wild, i.e., in uncontrolled environments with arbitrary sound sources. The uncontrolled nature of these recordings distinguishes them from HRTF acoustic measurement in anechoic chambers. They also differ from HRTF upsampling, where HRTFs with few sound source positions are interpolated to obtain a higher spatial resolution. One advantage of binaural recordings is that they encode the individual characteristics of the HRTF. However, these recordings are not properly HRTFs since spatial cues are mixed with the sound source content and the room influence. Therefore, the considered HRTF individualization methods train ML models to estimate the individual HRTF from the binaural recordings.

Zandi et al. [90] trained a CVAE to learn a latent representation of HRTFs using the ITA dataset. Then, they asked participants to hold a smartphone emitting a sine sweep for several positions around them. The binaural recordings were used to fine-tune the decoder of the CVAE and generate individualized HRTFs. In a similar approach, Jayaram et al. [91] used a modified version of UNet to predict the individual HRTF from the short-time Fourier transform of binaural recordings. Their method required participants to move their heads in the presence of a stationary arbitrary sound source. They trained UNet with both real data measured from two subjects, and synthesized data obtained by spatializing sound sources with the HRTFs from the RIEC dataset [92]. Both methods yielded acceptable results for both SD, which was between 4 and 5 dB, and in localization experiments. However, these approaches exhibit some limitations, including the difficulty of conducting the recordings in everyday environments, where the conditions may diverge from those tested by the authors. Additionally, the measurement procedure required to the user is prone to errors.

V. INPUT PREPROCESSING

Data preprocessing is a crucial step to guarantee effective training of ML models. Raw data may be noisy, incomplete, heterogeneous, inconsistent, or contain irrelevant information [52, Ch. 3]. This section is devoted to the anthropometry preprocessing, which is the most prevalent input data type for HRTF individualization. Anthropometry preprocessing mostly involves feature scaling and feature selection techniques. Preprocessing approaches for images include resolution scaling [76], [82], [83], data augmentation [76], [77] and edge detection [76], [79]. Conversely, 3D scans preprocessing include voxelization [83], [93] and resolution scaling [86].

A. FEATURE SCALING

Feature scaling entails rescaling the feature values to a different range. This operation is beneficial for several ML algorithms, as it enables the comparison of features distributed in different ranges. A frequently used technique is min-max normalization, which entails rescaling the values to a fixed range, typically between 0 and 1 [94], [95], [96], [97], [98], [99], [100], [101], [102]:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{5}$$

where x and x' are the original and the scaled feature, respectively. In the HRTF individualization context, another commonly employed method to bound values within [0, 1] is based on the sigmoid function [74], [84], [103], [104], [105], [106], [107]:

$$x' = \left(1 + e^{-\frac{x-\mu}{\sigma}}\right)^{-1}, \tag{6}$$

where μ and σ are the mean and the standard deviation of x, respectively. Moreover, standardization is another widespread technique, which transforms the feature vector to have zero mean and unit variance: $x' = (x - \mu)/\sigma$ [60], [85], [108]. A less common method is mean centering [73].

B. FEATURE SELECTION

Several HRTF individualization studies adopted different approaches to remove irrelevant information from the input features and avoid the curse of dimensionality. Whereas anthropometric feature selection is prevalent in this context, some researchers have also explored dimensionality reduction techniques, such as PCA [72], [109], sparse PCA [73], [110], autoencoders [104], and feature learning [77], [86]. Despite the multitude of studies dedicated to anthropometry selection, there is currently no consensus on the optimal set of parameters for HRTF individualization [27] [17, Sec. 7.1.2]. This topic has been investigated from various perspectives. A particular focus of analysis has been the influence of the pinna and its cavities on HRTF. Although the precise relationship between pinna morphology and HRTF remains uncertain, there is a general consensus that the concha and the fossa triangularis are linked to the pinna spectral notches observed in the HRTF [26], [64], [111]. In addition, these cavities play an important role in elevation localization [112] and their anthropometry significantly affects HRTF [27], [69], [113], [114]. A number of studies have focused on other body parts, such as the head shape and size, which are directly related to ITD [28], [34] [17, Sec. 7.1.2] and ILD [115], [116].

The HRTF individualization studies adopted different approaches to select anthropometry. Most of them relied on the CIPIC anthropometric specification, which allows us to summarize their results. Specifically, we examined 17 studies proposing a feature selection approach for CIPIC anthropometry. Table A.2 of supplementary materials provides a complete list of these studies. We also included studies that were not identified through the PRISMA research, as they employed feature selection in conjunction with selection-based methods for HRTF individualization [117], [118]. Then, we excluded some of the studies analyzed in this survey since they reused the results of existing approaches [107], [119], selected parameters without an objective approach [70], [120], or used parameters that did not belong to the CIPIC specification [102], [110], [121], [122]. It should be noted that other anthropometry selection approaches not included in this analysis do exist [44], [71], [123].

Despite the variety of feature selection methods, two common steps are often considered. The first step entails selecting parameters significantly related to the HRTF [100], [117], [124], dimensionality reduced versions of HRTF obtained with PCA [85], [113], [118], [125] or HOSVD [126], [127], or other HRTF features such as ITD, ILD, and pinna notches [128]. This relationship can be quantified through a parameter-based correlation analysis [100], [124], [125], [126], [127], [128], [129], or by computing the parameters' importance with regression models such as MLR [85], [113], [117], [118] and SVM [130]. The second step aims to reduce redundancy by removing parameters that are highly correlated to each other. Anthropometry selection is usually performed independently of the spatial position. Xu et al. [125] compared global—one selection for all directions—and local—one selection for each direction—approaches to anthropometry selection. For both approaches, they used a weighted correlation between the anthropometry and the PC weights of the HRTF for selection. The authors predicted the PC weights with MLR obtaining a non-significant difference in SD values between the two approaches, which were around 5 dB.

Fig. 7 shows the frequency of parameter selection across the 17 studies, revealing that certain parameters are more frequently selected than others. Head width x_1 and head depth x_3 are the most frequently selected parameters among those pertaining head and torso. This could be explained by the influence of head morphology on ITD and ILD. However, pinna parameters are more frequently selected than head and torso parameters on average. The most frequently selected parameter is the cavum concha width d_3, confirming the importance of concha morphology for HRTF. Other frequently selected parameters include pinna height d_5, cavum concha height d_1, pinna width d_6, and fossa height d_4. For a related analysis of pinna anthropometry selection, refer to Ghorbal et al. [27].

VI. OUTPUT

The HRTF individualization methods yield an estimated personalized HRTF as output, although alternative representations of it may be considered. The majority of HRTF individualization methods focused on the HRTF magnitude, whereas phase and ITD are rarely considered [82], [85], [107], [131], [132]. Other methods predicted the HRIR so that the phase modeling was not necessary [74], [75], [76], [89], [98], [99], [109], [120], [129], [133], [134], [135], [136], [137]. Some methods focused on the DTF magnitude to conceal irrelevant information for the ML model [60], [72], [97], [119], [124], [128], [130], [138], [139], [140]. A limited number of methods concentrated on the PRTF. Rodríguez and Ramírez [141] used MLR to predict the PC weights of PRTF from pinna anthropometry. Then, they adjusted the pinna notch frequencies of the estimated PRTF. Ko et al. [82] trained an end-to-end CNN, called PRTFNet, to predict the magnitude of compact PRTF representation from pinna range images.

Besides the use of alternative HRTF representations, the output of ML models can be constrained to specific spatial coordinates and frequency ranges to simplify the problem. For example, some studies focused on the median plane to model the monaural spectral cues affecting elevation perception [60], [89], [126], [133], [139], [141], [142], [143], [144], [145], [146]. Other works focused on the horizontal plane, whereas the majority considered the full sphere around the subject. With regard to frequency ranges, it is common to neglect the low frequencies below:

- 200 Hz [84], [86], [97], [103], [104], [105], [108], [147], [148],
- 500 Hz [79], [132], [135], [149],
- 1 kHz [70], [83], [121],
- 3 kHz [77], [150],
- 4 kHz [93].

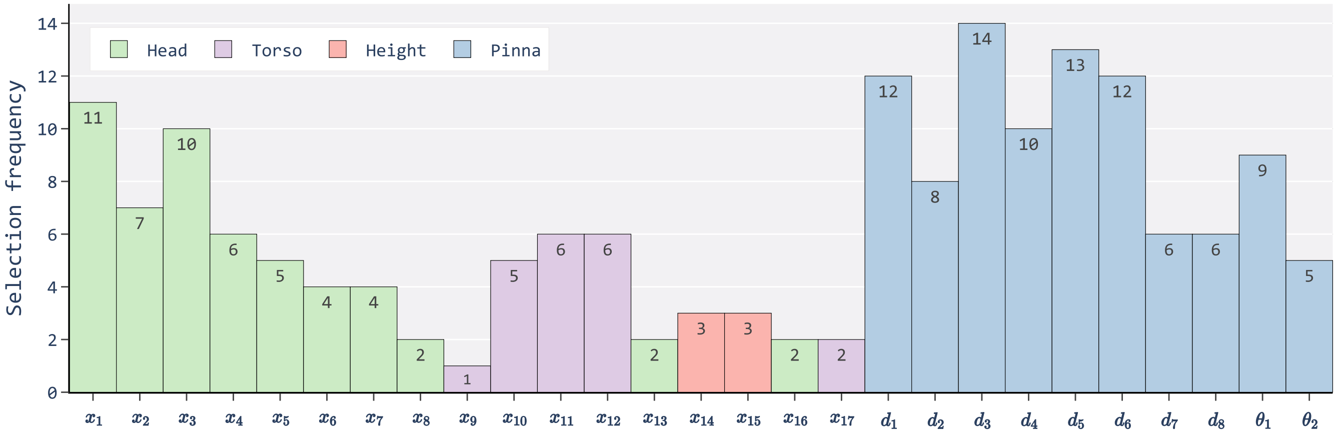

FIGURE 7. Frequency of the CIPIC anthropometric parameters selected across the 17 analyzed works.

In addition, some works neglected the higher frequencies as they are marginal for sound localization [151]. The typical highest considered frequencies are:

- 12 kHz [70], [83],
- 13 kHz [121],
- 15 kHz [77], [84], [103], [105], [107], [132], [142], [147], [148],
- 16 kHz [79], [93], [120], [135], [149], [150],
- 18 kHz [86], [104].

VII. OUTPUT PREPROCESSING

A. FEATURE SCALING

In the field of HRTF individualization, a common approach to scaling the HRTF values is to compute their log-magnitude. He et al. [95] suggested that the log-magnitude provides improved performances over power [88] and no preprocessing in predicting the individual HRTF using a sparse representation. Other feature scaling approaches include min-max normalization [79], [84], [87], [97], [98], [99], [102], standardization [132], [152], and RMS normalization [130].

Preprocessing methods based on signal processing include HRIR truncation [120], [136], also to isolate the pinna influence [143], [144], and frequency smoothing [83], [84], [103], [104] which can be also based on auditory critical bands [130] or an equivalent rectangular bandwidth (ERB) filter [86]. Alotaibi and Wickert [120] proposed to preprocess the HRIR removing the ITD, and to restore it as a postprocessing step.

B. DIMENSIONALITY REDUCTION

HRTFs are characterized by a high dimensionality, typically comprising more than 100 frequency bins. Consequently, many HRTF individualization methods encompass dimensionality reduction techniques to facilitate the training of ML models. Fig. 5 shows the distribution of the dimensionality reduction techniques used in the analyzed publications.

1) PCA

One of the earliest and most enduring dimensionality reduction algorithms is PCA. Kistler and Wightman [153] found that the first five PCs, which explained 90% of the variance, yield a sound localization similar to the original HRTF. Many HRTF individualization studies selected a number of PCs to explain 90% or more of the variance. However, the actual number of PCs varied across the studies including 5 [131], 6 [102], [142], 7 [72], 8 [122], 10 [124], [128], [129], [134], [138], [139], [152], 12 [89], [98], [119], 15 [121], 20 [141], [150].

Some studies have compared the application of PCA to the entire or grouped HRTF data. Xu et al. [152] reported an SD improvement of 1.2 dB by applying the PCA independently for each spatial direction compared to applying the PCA for all spatial positions when predicting the HRTF magnitude from anthropometry using MLR. Bomhardt et al. [121] reported that the application of PCA on HRTFs grouped by direction yielded a lower reconstruction error compared to the ipsi- and contralateral grouping and the ungrouped condition only when less than 20 PCs were considered. A variant of the PCA is the spatial PCA (SPCA) [154], which is applied in the spatial domain instead of the time or frequency domains. SPCA decomposes the HRTF into a weighted combination of spatial PCs (SPCs). Zhang et al. [85] employed SPCA to decompose the HRTF magnitude into 200 SPCs and to predict their weights from anthropometric parameters using NNs. The same method was then employed in combination with a distance-dependent HRTF model [155], and later refined integrating numerical simulation [156]. Chen et al. [157] applied PCA both in the spectrum and spatial domains in order to reduce the reconstruction error.

Additional studies specifically investigated the influence of various factors on HRTF compression using PCA. These included the comparison between different HRTF spectrum representations [158], [159], the application of PCA by jointly handling the HRTFs of the left and right ears [160], and the influence of other factors such as the input structure (signal or space), domain (time or frequency), the HRTF smoothing, and the HRTF dataset [161]. To assess the suitability of the PCA for retaining the individual information of the HRTF, Fayek et al. [132] trained a NN with one hidden layer to classify the subject based on the PCs' weights of the HRTF. They found that classification accuracy decreased as the number of PCs decreased, achieving a test accuracy of 22% with 85% of the variance explained. Consequently, they suggested that inter-subject variation in a set of HRTFs has a relatively minor impact on the overall variance compared to the variance between the spatial directions. Although this may suggest a potential inappropriateness of PCA for HRTF individualization, the observed low performances could be attributed to other factors, such as the poor accuracy of the classifier. Thus, further studies are necessary to confirm this hypothesis.

2) ICA

Another dimensionality reduction technique used in HRTF individualization is independent component analysis (ICA). Huang and Zhuang [99] employed ICA to reduce the HRTF dimensionality to 18 independent components and estimated them from anthropometry using support vector regression (SVR). Wang and Chan [109] proposed a similar approach using 2D common factor decomposition followed by ICA for dimensionality reduction and used SVR as a regression model. Further, Liu et al. [145] reduced the median plane HRTF to ten independent components and predicted the obtained weights through MLR from anthropometry. Similarly, Liu et al. [146] used two independent components, and trained three generalized regression neural networks (GRNN) based on anthropometry of head, torso, and pinna, respectively.

3) TENSOR-BASED

Some researchers employed tensor generalizations of dimensionality reduction techniques. These included singular value decomposition (SVD), which is generalized to tensors by higher-order SVD (HOSVD). Some studies suggested that HOSVD yields lower SD than PCA in the individualization of HRTF using MLR [149] and radial basis function neural network (RBFNN) [126]. Similarly, Rothbucher et al. [138] reported a slight improvement of SD using tensor-based techniques such as HOSVD, two-dimensional PCA (2DPCA), and generalized low rank approximations of matrices (GLRAM), in comparison to standard PCA. Also, HOSVD and GLRAM led to a higher reduction rate than PCA and 2DPCA. In a previous study, the authors reported that GLRAM and HOSVD exhibited lower SD compared to PCA in the lone HRTF dimensionality reduction task [162]. HOSVD has also been used along with higher-order partial least squares (HOPLS) to predict HRTF [127].

4) NON-NEGATIVE MATRIX FACTORIZATION

Tang et al. [100] used non-negative matrix factorization (NMF) along with SVR to predict low-dimensional HRTFs from anthropometry. The authors reported a mean SD of 4.7 dB compared to 5.1 dB obtained with MLR combined with PCA [124]. However, these results were obtained evaluating the method for only one subject at four azimuths in the horizontal plane.

5) FILTER APPROXIMATION

Other less investigated approaches for HRTF dimensionality reduction include those based on signal processing techniques. Gupta et al. [133] employed Prony's signal modeling method to approximate the HRIR with a set of time delays and resonant frequencies. Then, this set was predicted from a linear combination of pinna anthropometric parameters using MLR. The authors reported better localization performances for the estimated HRTF compared to a non-individual HRTF.

6) SPHERICAL HARMONICS

Given the spherical nature of HRTFs, some researchers have investigated the use of spherical harmonics (SH) for dimensionality reduction. Xi et al. [105] used SH to combine CIPIC and HUTUBS datasets. They predicted the SH coefficients from anthropometry using a DNN and reported a mean SD of 4.46 dB for CIPIC HRTFs, which was slightly lower than existing approaches [84], [85], [163]. Further, Zhao et al. [77] employed head and torso anthropometry and pinna image features to train an NN predicting SH coefficients of HRTFs. They reported a mean SD of 5.31 dB, representing an improvement over the average HRTF and a value similar to that obtained by Zhi et al. [61].

7) AUTOENCODER

In one of the earliest studies using autoencoders with HRTFs, Luo et al. [87] reported enhanced SDR values in comparison to PCA, albeit with a subtle difference. The authors used perceptual feedback as input similarly to Yamamoto and Igarashi [88], who also trained an autoencoder to obtain an HRTF generator. Later, the training of NNs to predict the low-dimensional HRTF representation obtained with an autoencoder became a widespread approach. Chen et al. [84] observed a reduction of SD by 0.5 dB on the horizontal plane compared to a previous study which did not consider dimensionality reduction [163]. Lu and Qi [108] trained a user-independent DNN to predict the low-dimensional HRTFs in the latent space obtained with an autoencoder. The model was then fine-tuned with user-dependent anthropometry for the purpose of individualization. Miccini and Spagnol [79] trained variational autoencoders (VAE) on both pinna images and HRTFs to train a DNN mapping the two latent spaces. The VAE trained on HRTFs was conditioned on the spatial coordinates. Yao et al. [104] used an autoencoder and a VAE to reduce the dimensionality of the anthropometry and HRTF, respectively. These latent representations were used to train a DNN, which yielded an SD improvement of almost 0.5 dB compared to PCA and SH for HRTF dimensionality reduction. Zurale and Dubnov [164] proposed a vector quantized VAE (VQ-VAE) for HRTF dimensionality reduction. Unlike VAEs, a VQ-VAE incorporates a quantization phase between the encoder and the decoder in which the latent space is quantized into a fixed number of vectors. The VQ-VAE model yielded lower SD for the HRTF reconstruction in comparison to both PCA and a standard autoencoder.

8) OTHER NON-LINEAR TECHNIQUES

Among non-linear dimensionality reduction techniques, Grijalva et al. [147], [148] indicated that Isomap outperforms PCA when using NNs to individualize HRTF from anthropometry, reporting a mean SD of 4.76 dB in the horizontal plane [147], and 4.6 dB for the full sphere [148]. The authors adapted the graph construction of Isomap to incorporate existing correlations among HRTFs. However, unlike PCA, Isomap is not capable of projecting new points into the low-dimensional space, although some approximations exist [165]. Studies dedicated to HRTF dimensionality reduction reported that Isomap and local linear embedding (LLE) outperformed PCA in the correlation with spatial direction [166] and in localization accuracy [167].

VIII. ML MODELS

After data preprocessing, an ML model is trained to learn the underlying relationship between the input and the output. As previously mentioned, in HRTF individualization, regression algorithms are selected, which represent a type of supervised learning. The distribution of the ML models used in the analyzed publications is shown in Fig. 5.

A. LINEAR REGRESSION

Linear regression approaches are widespread, particularly in early HRTF individualization methods (see Fig. 4). Given the high-dimensional nature of HRTFs, multivariate linear regression (MLR) is the typical choice. The MLR model is rarely trained to directly predict the raw HRTF magnitude values, although some approaches do exist [143], [144]. Several studies predicted PCs weights obtained from the HRTF [102], [110], [121], [131], [152], [168], PRTF [141], DTF [72], [124], [128], [130], [138], [139] and HRIR [129], [134]. Other HRTF dimensionality reduction techniques used in conjunction with MLR include HOSVD [127], [149] and ICA [145]. Chen et al. [136] divided the HRIR into three segments, each corresponding to the influence of a different body part: the head and pinnae, the torso, and the knees. Then, they constructed a distinct MLR model for each segment, using three distinct sets of anthropometric parameters.

Given its simplicity, MLR has been found to lack the capacity of adequately describe the complex relationship between input data, usually anthropometry, and the HRTF. This is demonstrated by the findings of studies that have employed MLR as baseline conditions in comparison to non-linear methods, which yielded superior performances [100], [101], [137], [142].

B. SPARSE REPRESENTATION

Sparse representation, which is a further linear approach, learns a sparse vector representing the anthropometry of the test subject as a linear combination of the training anthropometry. The same sparse vector is then directly applied to the HRTF magnitude. Following this approach, Bilinski et al. [94] reported lower SD than ridge regression and a non-individual HRTF, whereas the nearest HRTF in the training set achieved only a slightly lower SD. He et al. [95] used sparse representation to predict HRTF magnitude, with a particular emphasis on the impact of preprocessing and postprocessing operations for HRTF. Zhu et al. [96] proposed a method for weighting the anthropometric parameters in the sparse representation based on their influence on HRTF magnitude. They obtained a mean SD of 5.5 dB, which was lower than 1.8 dB compared to other literature methods [59], [117], [169], but similar to the unweighted approach.

Qi and Tao [97] questioned the underlying assumption of previous works that the same sparse vector can be used for both anthropometry and HRTF. Thus, they trained a DNN to map the learned weights for anthropometry and HRTF. They reported better objective and subjective performance compared to classical sparse representation [94], [95]. Lu et al. [73] reduced the dimensionality of morphological landmarks with sparse PCA and subsequently used sparse representation to select the best-match HRTF. The authors reported improved SD compared to other selection-based approaches [59], [170], especially between 0 and 8 kHz. Later, the authors proposed a modified version of their approach employing PCA for HRTF dimensionality reduction [171] resulting in lower SD values compared to their previous work.

C. NEURAL NETWORKS

Neural networks (NNs) can overcome the limitations of linear approaches. Hu et al. [119] proposed the first HRTF individualization method based on a feedforward NN (FNN) with one hidden layer predicting the PCs weights of horizontal plane HRTF from anthropometry. Following studies employed a single-layer FNN to predict the raw HRTF [101], [132], or low-dimensional representations obtained with PCA [142], spatial PCA [85], and Isomap [147], [148]. The number of neurons in the hidden layer varied including 16 [119], 18 [142], 20 [101], [147], 35 [148], and 512 [132]. In light of the recent proliferation of deep learning, several HRTF individualization studies proposed to use DNNs, i.e., NNs with several hidden layers, to predict HRTF from anthropometry. Nevertheless, in these studies, the depth of DNNs remained contained, likely due to the limited size of HRTF training data. The proposed DNN architectures included three [84], [105], five [103], [104], [106], or seven [120] hidden layers. The number of neurons for each layer included 40 [84], 48 [103], 64 [106] and 128 [104], [120]. DNNs were also used to map low-dimensional representations of pinna depth images to HRTFs [79].

Other types of NNs have been employed for HRTF individualization. Among recurrent neural networks (RNNs), rarely investigated in this field, Lee et al. [172] proposed a bidirectional long short-term memory (LSTM) to predict horizontal plane HRTF from anthropometry, although their paper was not completely peer reviewed. Conversely, CNNs have received considerable interest by HRTF individualization studies. CNNs have been used to predict the HRTF from anthropometry [107] and 3D head meshes [86]. Further, some works trained CNNs by combining anthropometry with pinna images [76], [77], [82] or 3D pinna meshes [83]. Lu et al. [74], [75] trained CNNs to predict HRTF from a set of landmarks placed on the torso, head, and pinnae, whereas Zandi et al. [90] and Jayaram et al. [91] trained CNNs with binaural recordings as input.

RBFNNs and GRNNs are a particular type of NN commonly employed for HRTF individualization. RBFNNs are composed of a single hidden layer with a non-linear radial basis function (RBF) as the activation function. Given their simplicity, RBFNNs are well-suited to HRTF datasets given their limited sizes. RBFNNs were used to predict low-dimensional HRTF obtained with PCA [157], [173] and HOSVD [126], using anthropometry as input. GRNNs were used to predict DTF on the median plane [60] and HRTF PC weights on the horizontal plane [122]. Liu et al. [146] trained three GRNNs for torso, head, and pinna to predict two ICA components of median plane HRTF.

Recently, more complex NN architectures have been proposed. Lu and Qi [108] trained an autoencoder to obtain low-dimensional HRTF features, which were then predicted from spatial coordinates using a user-independent model. The model was then fine-tuned incorporating user-dependent information represented by anthropometry. The authors reported a mean SD improvement greater than 0.5 dB and improved localization performances compared to random HRTF and other literature works [95], [150], [174]. Qiu et al. [175] proposed a multi-stage model combining the modeling of global and local spectrum features. First, they trained a LightGBM model for each HRTF frequency, with anthropometry as input. Then, they trained a transformer encoder to learn global spectrum features from the previously predicted HRTFs. The authors reported a mean SD of 4.54 dB, which was slightly lower than the 4.79 dB obtained by replacing the transformer with a classical NN. Comparable SD values were found for LightGBM alone and existing literature methods [156], [176]. Zhang et al. [156] and Javeri et al. [93] proposed to integrate the principles of HRTF numerical simulation approaches, such as the boundary element method (BEM), into NNs.

D. SUPPORT VECTOR REGRESSION

Among regression algorithms, support vector regression (SVR) has received considerable interest for HRTF individualization. SVR with an RBF kernel has been employed to map anthropometry to low-dimensional HRTFs in the horizontal plane, obtained with PCA [98], ICA [99], and NMF [100]. SVR yielded improved SDR compared to a single-layer NN [98], [99] and improved SD over linear regression [100]. Wang and Chan [109] proposed a joint optimization of SVR to exploit the correlation between the HRTF dimensions. They reported a mean SD of 4.6 dB, which was lower than only 0.3 dB over the standard SVR.

E. TREE-BASED MODELS

Decision trees and random forests have been less investigated in the HRTF individualization field. Teng and Zhong [70] trained a random forest to predict the HRTF magnitude from anthropometry, but the model was evaluated at only five directions. Qiu et al. [177] employed LightGBM, a gradient boosting framework based on decision trees, to predict HRTF from anthropometry. They reported a mean SD of 2.3 dB, which was lower than existing approaches [76], [156]. Later, the authors integrated LightGBM in a multi-stage model [175]. Despite the limited interest, decision trees and random forests could represent an effective approach and a simpler alternative to DNNs. Angelucci et al. [137] compared the efficacy of different ML models in predicting the HRIR from anthropometry. These models included linear regression, kernel regression with RBF kernel, SVM with RBF kernel, regression tree, random forest, and DNN. The results indicated that the latter three models exhibited lower error values.

F. REINFORCEMENT LEARNING

Reinforcement learning is an ML paradigm based on the actions taken by an intelligent agent in an environment with the goal of maximizing a given reward [178]. Reinforcement learning has rarely been applied to HRTF individualization, as it is less suited to this task than supervised learning. Nambu et al. [135] employed the actor-critic paradigm using a dummy head HRIR as the initial HRIR. They reported that this approach resulted in improved localization on the horizontal plane compared to the dummy-head HRIR.

IX. MODEL TRAINING AND VALIDATION

A. MANAGE HRTF STRUCTURE

HRTF individualization methods must address the complex structure of HRTF data. HRTFs depend on multiple variables, including the subject's anatomy, the left and right ears, the sound direction (azimuth and elevation), and the frequency (time for HRIRs). Many ML models are unable to accommodate the multidimensional structure of HRTF data. To manage the HRTF structure, two high-level approaches exist, which are identified by the number of trained models.

- *Single ML model:* when using a single ML model, one or more HRTF variables are employed as input features in conjunction with the actual features (e.g., anthropometry). For example, the direction can be used as an additional input feature of a model predicting a multidimensional output for the given direction [74], [82], [104], [120], [132], [148], [150]. The multidimensional output is represented by the HRTF frequency bins, the HRIR time sample, or their low-dimensional representation. In a similar manner, the frequency bin can be utilized as an input feature to predict the SH coefficients of the HRTF [77]. Moreover, both direction and frequency can be used as additional input features to ML models that predict a unidimensional output [101]. A single model can be used even without additional input features by employing ML models that predict a 2D matrix as output, which represents the HRTF values for each frequency bin and direction [83], [107]. Similarly, when considering the full sphere HRTF, the ML model can yield a 3D output with dimensions corresponding to frequency, azimuth, and elevation [86].

- *Multiple ML models:* some studies trained a separate model for each direction to estimate a multidimensional output which represents the frequency bins [60], [75], [76], [84], [103], [119], [141], [142]. Alternatively, one model is trained for each frequency bin, with the model's multidimensional output corresponding to the directions [85]. A further approach is to train one model for each frequency bin, with the direction used as an additional input feature [175], [177] or vice-versa [70].

Angelucci et al. [137] conducted a comparative analysis of the two approaches: (*a*) training a one ML model for each direction and (*b*) training a single ML model with azimuth and elevation as additional input features. Results varied according to the ML algorithms. However, the authors concluded that, in general, the approach based on a single model for each direction performed better in terms of objective metrics.

B. DATASET SPLITTING

In the ML pipeline, the employed dataset is split into different partitions, namely training, validation, and test sets, used at different stages of the pipeline. The training set is used to train the ML model to predict the desired output in response to the input. The validation set is composed of data that were not used in the training step. Its purpose is to provide an unbiased evaluation of the models trained with different configurations of the hyperparameters. In the studies analyzed in this survey, the validation set has rarely been considered [74], [83], [85], [101], [105], [132], [175]. The test set is employed to assess the extent to which the final ML model is able to generalize on unseen data.

A number of strategies exist for dataset splitting into different partitions. The simplest one is the holdout method, which involves performing a single split. The holdout method is the most frequent approach in HRTF individualization studies. The typical percentages of data retained for the test set are approximately:

- 5% [73], [79], [100], [124], [157],
- 10% [70], [101], [122], [143], [144], [173],
- 15% [82], [87], [97], [98], [108], [120], [142], [145],
- 20% [85], [99], [104], [107], [109], [119], [137], [138], [177].

An alternative approach to holdout is represented by cross-validation (CV) techniques, which evaluate the model's performance with multiple partitions of the dataset. The use of CV techniques typically provides a more thorough evaluation than holdout, as they serve to prevent overfitting and selection bias. A common CV technique is k-fold CV which involves the splitting of the dataset into k equal-sized partitions. For each of the k trained models, one partition is designated as the test set, whereas the remaining $k - 1$ partitions constitute the training set. Common values of k in the analyzed HRTF individualization studies are 4 [60], 5 [83], [148] and 7 [126], [147] and 10 [86], [150]. Nevertheless, the most prevalent CV approach is the leave-one-out CV (LOOCV), which can be regarded as a particular case of k-fold CV, where k equals the number of instances in the dataset [76], [77], [84], [88], [90], [94], [95], [96], [102], [103], [105], [132], [146], [149]. Moreover, some studies utilized CV without providing details regarding the specific methodology [131], [168], [179].

C. HYPERPARAMETERS TUNING

The training of ML models is influenced by their hyperparameters, which control the learning process, in contrast to the model's parameters learned during such process. Hyperparameter tuning, or optimization, is the procedure of finding the optimal values of the model's hyperparameters for a given problem. In the analyzed HRTF individualization methods, hyperparameter tuning is often overlooked, and the used hyperparameter values are omitted. Some studied conducted an informal tuning of the hyperparameters without a specific strategy [83], [150]. In several studies employing a single-layer FNN, the number of hidden nodes is varied and the value that minimizes the error on the validation set is selected [148]. However, the majority of these studies omit the set of data used to compute such error [99], [119], [142], [147]. Lu et al. [75] tuned the dropout rate for a CNN by identifying the value that minimized the error on the validation set. Some studies tuned the hyperparameters based on error minimization during CV, but without specifying the employed search method [94], [96], [97], [101], [127]. Other studies conducted a grid search to tune the hyperparameters but did not provide details on the validation strategy for the models trained with different configurations of the hyperparameters [70], [146]. Qiu et al. [177] employed Bayesian optimization with k-fold CV to tune the parameters of a LightGBM model.

D. DATASET COMBINATION

HRTF datasets are usually small-sized due to the difficulties of HRTF acoustic measurement. Therefore, ML models for HRTF individualization can be improved by combining multiple datasets. This can be beneficial for several reasons, including the availability of larger sets of data to train the model. However, the HRTF measurement procedures adopted for different datasets result in considerable acoustic differences [180], [181]. This has the potential to enhance the generalization capacity of ML models, yet their adequate training is challenging, as cross-dataset differences must first be mitigated. This allows to obtain harmonized HRTF data and prevent the training of biased ML models.

Despite several works devoted to the mitigation of cross-dataset differences exist in the literature [182], [183], [184], [185], few HRTF individualization studies have adopted one of these approaches to date. Lu et al. [73], [171] evaluated the proposed HRTF individualization method separately on the CIPIC and Chinese pilots' dataset, rather than performing a joint training and evaluation. Lu et al. [74] trained and evaluated their HRTF individualization method with the Chinese pilots' dataset. Then, they reported an evaluation of the trained model on the HRTFs of two subjects from the SYMARE dataset. Xi et al. [105] performed an evaluation of the proposed HRTF individualization method combining the CIPIC and HUTUBS datasets. These datasets were harmonized by aligning the mean and standard deviation of the HUTUBS HRTFs magnitude to those of the CIPIC dataset. Lu and Qi [108] used the PKU&IOA and the CIPIC datasets to train their model for HRTF individualization. The only reported operation to harmonize the datasets is resampling to the same sample rate.

X. EVALUATION METRICS

Following the training of the ML model, its performance is quantified by computing some evaluation metrics. In the context of HRTF individualization, the type of evaluation can be objective, perceptual, or based on auditory models.

A. OBJECTIVE METRICS

The evaluation through objective metrics quantifies the error introduced by the HRTF estimation with respect to the individual HRTF. Typically, such error is computed between the magnitude spectra of the target and the predicted HRTFs. Although several metrics exist, spectral distortion (SD) is the prevalent one. Other less common metrics include:

- root mean square error (RMSE) in addition to SD [76], [97], [106], [107], [108] or alone [137],
- signal-to-distortion ratio (SDR) [87], [98], [99], [127],
- mean absolute error (MAE) [101],
- mean squared error (MSE) [88], also in percentage [128], [129], [134], [139],
- inter-subject spectral difference (ISSD) [121],
- spectral distance error (SDE) [83],
- R^2 [72],
- Itakura–Saito divergence [93].

In the field of HRTF individualization, it is a common practice to compare the objective results obtained with the proposed method against other conditions to demonstrate the improved results achieved by the former. These control conditions include:

- variations to the proposed approach [77], [79], [82], [84], [85], [104], [109], [125], [126], [138], [147], [148], [149], [150], [175],
- other approaches proposed in the literature [73], [74], [76], [77], [82], [83], [86], [97], [105], [136], [152], [156], [171], [173], [175],
- different ML models [75], [98], [99], [100], [101], [127], [137], [142],
- numerical simulation [156],
- baseline conditions, e.g., average [76], [77], [83], random [130], and generic [82], [83], [85], [107], [145] HRTFs.

However, only a minority of these studies conducted a statistical analysis to assess the significance of the reported improvements [73], [103], [125], [142], [148], [152], [156], [171].

Although objective metrics provide a clear quantification of the fit of the estimated HRTF magnitude to the individual one, the relationship between such metrics and perceptual outcomes has not yet been demonstrated. Tommasini et al. [142] questioned the suitability of SD for evaluating the localization accuracy on the median plane. To this end, they extracted the central frequencies and the amplitudes of the pinna peaks and notches since their relationship with the localization on the median plane has been demonstrated [24], [62], [186], [187], [188], [189]. A comparison of linear regression with an NN in an HRTF individualization task found that the NN exhibits lower SD. However, they found similar errors on the predicted peaks and notches, large enough to have a perceptual impact. Thus, they suggested that SD is an inadequate metric for assessing the localization accuracy of an estimated HRTF. To obtain a similar indication of the localization accuracy through objective metrics, Bomhardt et al. [121] reported the correlation coefficient between the peaks and notches extracted from the individual and the estimated HRTFs.

B. PERCEPTUAL EXPERIMENTS

HRTF individualization should ultimately yield HRTFs that provide a perceptual experience as close as possible to the individual HRTF. Consequently, a perceptual evaluation of HRTF individualization methods is more appropriate than the computation of objective metrics. Despite that, the majority of the examined studies reported only objective evaluation, and those that conduct perceptual experiments rarely exceed ten participants, with few exceptions [85], [88]. Further, there is currently no standard protocol for perceptually evaluating the estimated HRTFs. However, a recent review discussed the methodology and the metrics to evaluate HRTF perceptual performances [190].

In perceptual experiments reported in HRTF individualization studies, participants are presented with auditory stimuli delivered through headphones. These stimuli are spatialized in different directions using the individualized HRTF and other HRTFs as control conditions. The participant provides feedback, typically in the form of perceived source direction, which is then used to compute performance metrics. Such experiments usually concentrate on the lone horizontal [74], [76], [90], [102], [106], [107], [119], [124], [127], [131], [134], [136], [173] or median planes [89], [133], [145], [157]. Some studies evaluated separately multiple planes [85] or the full sphere [91], [97], also at different distances [108]. The stimulus is spatialized in a number of directions that varies from 6 [133], [145] to 24 [124], whereas a greater number is rarely considered [108]. Then, each direction can be considered once [90], [102], [124], [136] or repeated up to ten times [108]. A noise signal of approximately one second [74], [89], [131], [145], [173] or noise bursts [72], [85], [107], [119], [124], [127], [133], [134] are employed as auditory stimuli. Other sound sources such as speech and music are less frequently considered [76], [88], [90], [108], [179]. In some experiments, a training session is conducted by presenting the participant with the stimulus spatialized in known directions. This training is performed in a dedicated session prior to the actual experiment [72], [91], [97], [107], [108], [134] or in repeated sessions preceding each experimental run [85], [119], [124]. Once the participants have listened to the spatialized stimulus, they are asked to provide feedback on the direction of arrival of the sound. To collect such feedback, a graphical user interface with a circle for angle selection is commonly adopted [85], [89], [90], [107], [119], [134], [136]. Other methods include the angle selection from a list [74], [102] or the use of a laser pen to measure the angle pointed by the participant [145].

In perceptual experiments, the auditory stimulus is spatialized with different HRTF conditions. One condition is the individualized HRTF obtained with the proposed methods, eventually with some variations [76], [85], [107]. The individual HRTF can be used as a control condition, which represents the target performance to be achieved [89], [91], [127], [131], [133], [134]. Notably, none of the analyzed studies used real sound sources as a control condition. Conversely, a generic HRTF is frequently used as a baseline control condition. A generic HRTF can be represented by an HRTF selected from a dataset [90], [108], [119], [124], an average HRTF [76], or a dummy head HRTF, such as KEMAR [74], [85], [89], [91], [102], [107], [136], [145], [157], [173] or B&K [133]. Furthermore, other literature approaches can be used for comparison [74], [97], [106], [107], [108].

To assess the localization performances, some metrics can be computed from the participant responses. These metrics include:

- absolute difference between the target and perceived directions [85], [89], [91], [102], [133], [145],
- ratio of correctly localized stimuli [74], [76], [90], [97], [102], [107], [108], [119], [124], [131], [134], [136], [157],
- front-back confusion ratio [72], [76], [85], [89], [91], [97], [107], [108], [117], [119], [124], [131], [134], [136], [145], [157],
- up-down confusion ratio [85], [97], [108], [145].

The distance confusion ratio can be considered in experiments involving different distances [108]. None of the analyzed studies considered the inside-the-head localization ratio, which measures the perceived externalization. Other studies have overlooked localization and have instead asked participants to rate the similarity between the estimated and individual HRTFs [106], [127], or the preference between the estimated and a non-individual HRTF [88], [173]. Further, Wang et al. [102] conducted a test in which participants were aware of the azimuth of the spatialized stimulus and were asked to rate the obviousness of such an angle. Whereas most experiments relied on audio-only stimuli, Lu and Qi [108] proposed an experiment with visual stimuli as well represented by a virtual reality (VR) scene in which realistic sound sources were rendered in 6 degrees of freedom. The authors reported an improved correct localization ratio over an audio-only experiment.

C. AUDITORY MODELS

Conducting perceptual experiments presents several practical challenges. Therefore, some HRTF individualization studies have relied on auditory models, i.e., computational models of the auditory system that can be used to estimate the localization responses of a subject with a given HRTF. Most of the employed auditory models are those included in the auditory modeling toolbox (AMT) [191]. Among the AMT models, the one proposed by Baumgartner et al. [81] has been used to predict localization performances in the median plane [60], [79], [85], [86]. Models for localization on the horizontal plane have been developed as well [192], [193]. The main limitation of using computational auditory models in the evaluation stage is related to the disjoint prediction of horizontal and vertical localization. However, a Bayesian spherical sound localization model has recently been proposed by Barumerli et al. [194]. The model jointly evaluates the two dimensions and has already been used to assess the similarity between predicted HRTFs and their measured counterparts [93].

XI. DISCUSSION

The preceding analysis of HRTF individualization based on ML revealed several commonalities across different approaches, despite considerable variability. Throughout the survey, we have endeavored to delineate an evolutionary trajectory across ML-based HRTF individualization methods. For example, the approaches based on NNs followed a nearly linear path over time, beginning with simple single-layer FNNs and progressing to deep NNs with increasing size and more sophisticated architectures. However, in many cases, the body of work on ML-based HRTF individualization does not follow a distinct evolution or progressive development. Instead, it comprises various studies that each adopts a specific approach or technique. Consequently, we grouped these references together to highlight the diffusion of a certain technique, rather than to present a chronological or methodological evolution between studies, given that such an evolution does not exist. This represents a potential manifestation of fragmentation and lack of standardization in the use of ML for HRTF individualization. This is in contrast to other fields related to audio and ML, which have been subjected to a well-structured standardization process, also encompassing scientific challenges. These challenges interest topics such as acoustic scene classification, e.g. DCASE[4], acoustic source localization and tracking, e.g. LOCATA[5], room acoustics characterization, e.g. ACE[6], and speech enhancement for 3D audio, e.g. L3DAS[7]. It is our hope that the organization of the literature provided in this survey will encourage future standardization actions in the ML-based HRTF individualization field.

[4] https://dcase.community/

[5] https://www.locata.lms.tf.fau.de/

[6] http://www.ee.ic.ac.uk/naylor/ACEweb/

[7] https://www.l3das.com/index.html

A. HRTF DATASETS

With regard to data, our findings indicate that the majority of studies relies on the CIPIC dataset, as illustrated in Fig. 5. However, Fig. 4 shows that this tendency has diminished in the last years, in favor of more recent and larger HRTF datasets. A comprehensive assessment of the CIPIC's suitability for the training of HRTF individualization methods has not yet been conducted. The limited size of CIPIC (45 subjects) and similar HRTF datasets renders the findings based on these data less generalizable. The training of ML models, in particular DNNs, would be favored by larger datasets, thereby preventing overfitting, which has been scarcely investigated in the HRTF literature. In addition, larger datasets would also contribute to the spread in the HRTF individualization field of the recent developments in ML. Examples of this include the latest deep learning architectures, such as transformers [107], [175] and few-shot learning [195], [196]. Furthermore, the explainable artificial intelligence (XAI) paradigm [197], [198] should be considered to foster the comprehension of the interrelationship between human anatomy and HRTF. In this regard, datasets of numerically simulated HRTFs comprising approximately one thousand subjects [31], [39] could serve as a first step toward more generalized ML models. However, simulated HRTFs exhibit perceptual differences with acoustically measured ones [37, Sec. 3.3] [43], [44]. In addition, ML methods dealing with limited data could be further investigated, such as transfer learning [77], [199] and data augmentation [76], [77], [79]. An extensive analysis of an HRTF individualization method should encompass multiple HRTF datasets for training and evaluation, in order to overcome the limitations of the single datasets. This is of particular importance in the HRTF field where the datasets are limited in size and the differences in the acoustic measurement across the datasets are significant [180], [181]. These differences are caused by (*a*) the employed equipment, (*b*) the measurement conditions (e.g., subject standing or seating, room characteristics, temperature, and humidity), (*c*) the considered spatial coordinates, (*d*) the reflection from the measurement system, and (*e*) the postprocessing. Recently, Pauwels and Picinali [200] found that these differences can be identified by ML algorithms. This topic is directly related to the repeatability of HRTF measurement, which is compromised by a number of factors. These include the placement of the in-ear microphones, the background noise, the accidental subject's movements and misalignment [180], [201], [202], [203]. The repeatability issue has also been observed for numerically simulated HRTFs [204]. Regarding this topic, some researchers investigated the datasets merging and harmonization by mitigating their spectral differences [183], [185] or standardizing the HRIR's sample rate and length, and obtaining a common spatial spherical grid finding the shared angles [184] or via interpolation [182]. Additionally, toolkits that facilitate the management and preprocessing of different HRTF datasets for an ML pipeline have been proposed [205]. Despite that, HRTF individualization studies have rarely addressed the topic of datasets merging so far (see Sec. IX.D). In addition to the technical perspective, other differentiating aspects between HRTF datasets should be considered, such as the subjects' characteristics (e.g., sex, ethnicity, age) [73]. The scarce heterogeneity of the subpopulations represented in such datasets prevents the development of fair artificial intelligence (AI).

B. ANTHROPOMETRY

The analysis of the input data types revealed that the majority of the studies are based on the CIPIC anthropometric parameters, as illustrated in Fig. 5. This choice is likely driven by the availability of these parameters in the CIPIC dataset and several other HRTF datasets [35], [36], [38], [39], [66], [206], [207], [208], which include anthropometry measured following CIPIC specification, as shown in Table 1 and in Table A.1 of the supplementary materials. Despite the prevalence of CIPIC parameters in the literature, the relevance and the comprehensiveness of such parameters for HRTF have been subject to debate [27], [69]. For instance, the CIPIC specification fails to sufficiently describe the fossa triangularis, despite its influence on HRTF has been demonstrated [64], [69]. Therefore, in the literature, some novel anthropometric parameters have been proposed to overcome the limitations of CIPIC specification [37], [39], [69], [70], [71]. Further investigation is needed to determine a set of anthropometric parameters that exhaustively describe the relationship between anatomy and HRTF, which is crucial for improving anthropometry-based HRTF individualization methods.

Furthermore, anthropometry-based methods should consider the inherent limitations of this input data type, in addition to those of the CIPIC specification. Anthropometric specifications should be defined in a more precise manner, rather than just relying on 2D sketches that may result in ambiguous measurement points. The lack of rigorous definitions impedes the replication of such measurements by other researchers and may result in errors that significantly affect the HRTF estimation [17, Sec. 7.5]. Anthropometry-based studies typically circumvent such repeatability issues by exclusively relying on anthropometric data included in the datasets. Thus, no additional anthropometric measurements are conducted on new subjects to evaluate the proposed HRTF individualization method. Further, as with HRTF acoustic recording, the measurement of anthropometry entails time-consuming sessions and experienced personnel, albeit the required equipment is generally less expensive. Thus, although the HRTF recording is yet more impractical than the anthropometric measurement, the latter is far from being accessible in end-user applications. This issue could be addressed by methods for the automatic extraction of anthropometric parameters [56], [57], [58], [59], [60], [61]. Alternative proposals of relevant landmarks such as pinna contours [209] or a bendy bone armature [210] can also promote reproducible research once supported by robust automatic extraction methods.

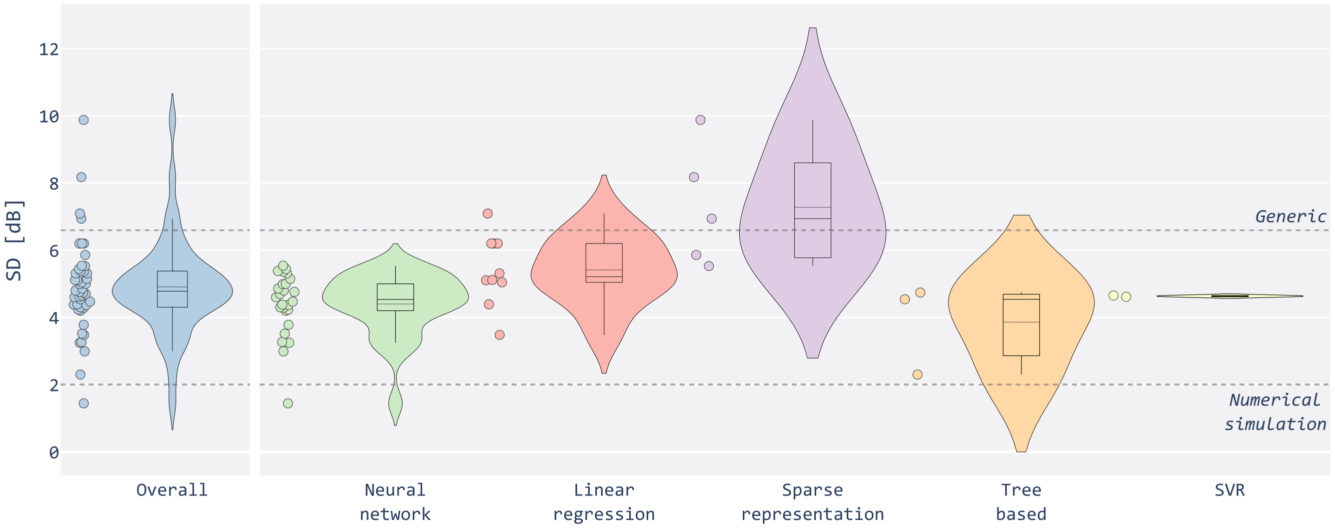

FIGURE 8. Distributions of the SD values reported in the analyzed HRTF individualization studies grouped by ML model. The SD of the reference conditions for generic and numerically simulated HRTFs are also shown (horizontal dashed lines).

C. EVALUATION OF INDIVIDUALIZED HRTFS

1) SD ANALYSIS

Another topic worth discussing is the evaluation methodologies of the HRTFs estimated by the proposed individualization methods. The majority of the analyzed publications consider exclusively objective evaluation. A rigorous and systematic comparison of the results reported in the analyzed studies is challenging to achieve due to the absence of a standard evaluation protocol. In most cases, the mean SD across directions, frequencies, and subjects of the predicted HRTF in comparison to the individual one is reported, yet some studies omit this information. The SD is often computed for the entire spectrum, although it is recommended to exclude the lowest (below 100 Hz) and highest (above 20 kHz) frequencies from the analysis, as the measurement equipment is less accurate in these bands [17, Sec. 7.5]. Thus, some studies reduced the frequency range considered for individualization, sometimes even further than the aforementioned limits, intending to focus on specific body parts or discard the frequencies irrelevant for localization (see Section VI). However, there is no common strategy to select such a frequency range. Further, the spatial coordinates under consideration vary. For instance, some studies focus on the horizontal or median planes, which makes it more challenging to compare different approaches.

Despite the discussed dissimilarities, in this survey, we provide a qualitative analysis of the publications reporting the mean SD obtained by evaluating the proposed HRTF individualization methods. A total of 46 out of 76 publications were identified as eligible for this analysis. Fig. 8 shows the SD distributions for these publications, grouped by the employed ML model. Two reference SD conditions are also reported to facilitate the interpretation of SD values. The first one is represented by the SD between a measured HRTF and the corresponding numerical simulation. We set this reference to 2 dB corresponding to the SD up to 18 kHz, obtained for HRTFs simulated using finite difference time domain (FDTD) on a 3D submillimeter mesh [211]. The second reference condition is represented by the SD between a measured HRTF and a generic one measured on a dummy head. We set this reference to 6.6 dB corresponding to the mean SD obtained comparing measured CIPIC HRTFs and KEMAR HRTFs with small [85] and large [107] pinnae. Although direct comparisons between different studies are difficult, we notice from the SD distributions that methods employing NN-based approaches tend to exhibit lower SD compared to other approaches such as linear regression and sparse representation. Tree-based models and SVR also yield promising results, although they have been employed in only a limited number of studies. From Fig. 8, we noticed a couple of SD values considerably lower than the rest of the literature. Son and Choi [106] reported a mean SD of 1.44 dB, but this value was computed for only two azimuth angles on the horizontal plane. Qiu et al. [177] reported a mean SD of 2.3 dB, computed on the full sphere using LightGBM. The authors hypothesized that the low SD is attributable to the efficacy of LightGBM in preventing overfitting. Nevertheless, further studies are necessary to reproduce the reported results and to confirm the effectiveness of LightGBM for HRTF individualization. The interpretation of these results is currently hindered by the lack of standardization in the field, which makes it challenging to draw a fair comparison with the state-of-the-art. This involves aspects such as the absence of common reference performances from state-of-the-art approaches, and the limited use of validation approaches like CV, which provide a more accurate estimate of how the ML models will generalize on unknown data. In addition, the practice of publicly releasing code repositories and trained models has been adopted by only a few authors [150], [176]. This practice should become established in future studies in favor of the open science paradigm, which is currently scarcely embraced in the HRTF research field.

Scientific challenges can represent a tentative approach to a common and open methodology for evaluating HRTFs. In such challenges, the approaches proposed by researchers undergo a shared benchmark using standardized metrics. The Listener Acoustic Personalization (LAP) challenge[8] is an example of such a community effort to develop a shared platform for the evaluation of personalized spatial audio technologies. In the first edition, the challenge focused on merging different HRTF datasets and spatially upsampling HRTFs. However, the organizers, some of whom are also authors, have included HRTF individualization as a natural challenge on the roadmap for future editions.

2) LIMITATIONS OF OBJECTIVE METRICS

The exclusive reliance on objective metrics to evaluate individualized HRTFs is a substantial limitation, given the inherently perceptual nature of HRTFs. There is evidence in the literature that SD is not straightforwardly related to perceptual cues such as the localization ones [142]. A comprehensive assessment of the performances of the predicted HRTFs necessitates perceptual experiments which, however, have only been conducted in a limited number of studies. These studies typically reported improved localization accuracy of the proposed method over non-individual HRTFs or other methods from the literature, yet inferior to the individual HRTF, though statistical analyses are sporadic [17, Sec. 7.5] [15]. To date, no study has demonstrated to provide estimated HRTFs that are perceptually indistinguishable from individual HRTFs. Given the impracticality of conducting localization tests, alternative approaches to perceptual experiments could be considered. Auditory models represent the prevalent alternative, although some NN-based metrics that capture localization perception have been proposed [212].

3) BEYOND LOCALIZATION EXPERIMENTS

The improved localization accuracy demonstrated by estimated HRTFs in comparison to the non-individual ones represents an encouraging result towards the employment of personalized HRTFs by end-users. However, the extent to which such improved performances will result in a superior spatial audio experience is contingent upon the specific end-user application. To this end, future research should consider perceptual aspects beyond mere localization, including externalization, tone color, crispiness, immersion, realism, and speech/music perception [11], [12], [44], [213], [214], [215] [37, Sec. 3.3]. For example, the perceived realism of the sound sources can be addressed by criteria such as authenticity [23], plausibility [216], transfer-plausibility [217], and co-immersion [218]. Moreover, more complex experiments are necessary to assess the performances of the individualized HRTFs in ecological virtual environments that encompass a multitude of factors [219], such as visual stimuli [108]. In this regard, the experimental settings could investigate the various levels of the reality-virtuality continuum [220], ranging from pure virtual auditory environments to audio augmented reality (AAR) [221] or virtuality (AAV) [222].

Future research should also investigate the role of HRTF personalization in conjunction with other facets of spatial audio contributing to a more realistic simulation. The existing literature provides evidence that metrics such as localization error, reversal rates, and externalization are influenced by HRTF as well as head tracking, HpTF, reverberation, and divergence between real and virtual listening environments [9], [223], [224], [225], [226], [227] [17, Ch. 8,11,12,13]. It also important to note that adequate hearing training contributes to improving the localization accuracy with non-individual HRTFs [228], [229], [230], [231], [232], a phenomenon known as HRTF accommodation [233].

XII. CONCLUSION

This survey systematically categorized and examined the ML-based HRTF individualization methods reported in the literature. The methods were categorized according to the typical steps involved in the ML workflow, i.e., dataset, input and output data, preprocessing, ML model, and evaluation. The analysis revealed the prevalent approaches for each step, which include CIPIC as the dataset, anthropometry as input, HRTF magnitude as output, possibly PCA for HRTF preprocessing, MLR or NNs as ML model, and SD for evaluation. Subsequently, we discussed the main gaps existing in the literature which could comprise topics of future studies, i.e.

1) the limitations of anthropometry-based methods,
2) the reported performances, which are still inferior to the individual HRTFs,
3) the scarce applications of recent ML developments, including XAI,
4) the lack of a standardized evaluation protocol, and
5) the infrequent investigation of perceptual metrics, especially in the context of ecologically-valid experimental settings, which encompass multiple aspects of spatial audio beyond HRTF individualization.

REFERENCES

[1] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning* (Adaptive Computation and Machine Learning Series). Cambridge, MA, USA: MIT Press, 2018.

[2] J. Veltman, A. Oving, and A. W. Bronkhorst, “3-D audio in the fighter cockpit improves task performance,” *Int. J. Aviation Psychol.*, vol. 14, no. 3, pp. 239–256, 2004.

[3] J. M. Loomis, R. G. Golledge, and R. L. Klatzky, “Navigation system for the blind: Auditory display modes and guidance,” *Presence*, vol. 7, no. 2, pp. 193–203, 1998.

[4] K. Sunder and S. Jain, “Virtual studio production tools with personalized head related transfer functions for mixing and monitoring dolby atmos and multichannel sound,” in *Audio Engineering Society Convention*, vol. 152, New York, NY, USA: Audio Eng. Soc., May 2022.

[5] E. M. Wenzel, F. L. Wightman, and D. J. Kistler, “Localization with non-individualized virtual acoustic display cues,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA: Assoc. Comput. Machinery, 1991, pp. 351–359.

[8] https://www.sonicom.eu/lap-challenge/

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

[7] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, “Using a typical human subject for binaural recording,” in *Proc. 100th Audio Eng. Soc. Conv., Copenhagen*, May 1996, pp. 1–18.

[8] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, “Binaural technique: Do we need individual recordings?,” *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, 1996.

[9] D. R. Begault, E. M. Wenzel, and M. R. Anderson, “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source,” *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916, Oct. 2001.

[10] J. Oberem, J.-G. Richter, D. Setzer, J. Seibold, I. Koch, and J. Fels, “Experiments on localization accuracy with non-individual and individual HRTFs comparing static and dynamic reproduction methods,” in *Proc. DAGA*, 2018, pp. 702–7015.

[11] L. S. R. Simon, N. Zacharov, and B. F. G. Katz, “Perceptual attributes for the comparison of head-related transfer functions,” *J. Acoustical Soc. Amer.*, vol. 140, no. 5, pp. 3623–3632, 2016.

[12] C. Jenny et al., “Usability of individualized head-related transfer functions in virtual reality: Empirical study with perceptual attributes in sagittal plane sound localization,” *JMIR Serious Games*, vol. 8, no. 3, Sep. 2020, Art. no. e17576.

[13] S. Xu, Z. Li, and G. Salvendy, “Individualization of head-related transfer function for three-dimensional virtual auditory display: A review,” in *Proc. Int. Conf. Virtual Reality*, Berlin, Heidelberg: Springer, 2007, pp. 397–407.

[14] P. Nowak, V. Zimpfer, and U. Zölzer, “3D virtual audio with headphones: A literature review of the last ten years,” *Fortschritte der Akustik–DAGA München*, 2018.

[15] C. Guezenoc and R. Seguier, “HRTF individualization: A survey,” in *Proc. AES Conv.*, New York, Oct. 2018, pp. 1–7.

[16] K. Iida, *Individuality of HRTF*. Singapore: Springer Singapore, 2019, pp. 59–105.

[17] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display*, 2nd ed. New York, NY, USA: J. Ross Publishing, 2013.

[18] R. Bomhardt, “Anthropometric individualization of head-related transfer functions analysis and modeling,” Ph.D. dissertation, Inst. Tech. Acoust., Aachen Univ., Berlin, Germany, 2017.

[19] S. Li and J. Peissig, “Measurement of head-related transfer functions: A review,” *Appl. Sci.*, vol. 10, no. 14, 2020, Art. no. 5014.

[20] K. Pollack, W. Kreuzer, and P. Majdak, “Perspective chapter: Modern acquisition of personalised head-related transfer functions–an overview,” in *Advances in Fundamental and Applied Research on Spatial Audio*, B. F. Katz and P. Majdak, Eds. Rijeka: IntechOpen, 2022, ch. 2.

[21] M. Cobos, J. Ahrens, K. Kowalczyk, and A. Politis, “An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2022, no. 1, May 2022, Art. no. 10.

[22] K. McMullen and Y. Wan, “A machine learning tutorial for spatial auditory display using head-related transfer functions,” *J. Acoustical Soc. Amer.*, vol. 151, no. 2, pp. 1277–1293, 2022.

[23] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Source Localization*. Cambridge, MA, USA: The MIT Press, 1996.

[24] E. A. Shaw and R. Teranishi, “Sound pressure generated in an external-ear replica and real human ears by a nearby point source,” *J. Acoustical Soc. America*, vol. 44, no. 1, pp. 240–249, 1968.

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

[26] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, “Extracting the frequencies of the pinna spectral notches in measured head related impulse responses,” *J. Acoustical Soc. Amer.*, vol. 118, no. 1, pp. 364–374, 2005.

[27] S. Ghorbal, T. Auclair, C. Soladie, and R. Seguier, “Pinna morphological parameters influencing HRTF sets,” in *Proc. 20th Int. Conf. Digit. Audio Effects*, 2017, pp. 1–7.

[28] J. C. Middlebrooks, “Individual differences in external-ear transfer functions reduced by scaling in frequency,” *J. Acoustical Soc. Amer.*, vol. 106, no. 3, pp. 1480–1492, 1999.

[29] I. Engel et al., “The SONICOM HRTF dataset,” *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 241–253, May 2023.

[30] S. Spagnol, M. Hiipakka, and V. Pulkki, “A single-azimuth pinna-related transfer function database,” in *Proc. 14th Int. Conf. Digit. Audio Effects*, 2011, pp. 209–212.

[31] C. Guezenoc and R. Seguier, “A wide dataset of ear shapes and pinna-related transfer functions generated by random ear drawings,” *J. Acoustical Soc. Amer.*, vol. 147, no. 6, pp. 4087–4096, 2020.

[32] M. Geronazzo, S. Spagnol, and F. Avanzini, “Estimation and modeling of pinna-related transfer functions,” in *Proc. 13th Int. Conf. Digit. Audio Effects*, Graz, Austria, Sep. 2010, pp. 431–438.

[33] T. Nishino, Y. Nakai, K. Takeda, and F. Itakura, “Database of head related transfer functions,” 1999, Itakura Laboratory and CIAIR. [Online]. Available: http://www.sp.m.is.nagoya-u.ac.jp/HRTF/

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

[35] O. Warusfel, “LISTEN HRTF database,” 2003, room Acoustics Team, IRCAM. [Online]. Available: http://recherche.ircam.fr/equipes/salles/listen/

[36] X. Guo, D. Xiong, Y. Wang, Y. Ma, D. Lu, and Q. Liu, “Head-related transfer function database of Chinese male pilots,” in *Proc. Int. Conf. Man- Mach.-Environ. Syst. Eng.*, Springer, 2016, pp. 3–11.

[37] 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, 2019.

[38] F. Brinkmann et al., “The HUTUBS head-related transfer function (HRTF) database,” 2019.

[39] S. Ghorbal, X. Bonjour, and R. Séguier, “Computed HRIRs and ears database for acoustic research,” in *Audio Engineering Society Convention 148*, Audio Engineering Society, 2020.

[40] M. Vorländer, “Past, present and future of dummy heads,” in *Proc. Acústica, Guimarães, Portugal*, pp. 13–17, 2004.

[41] M. Burkhard and R. Sachs, “Anthropometric manikin for acoustic research,” *J. Acoustical Soc. Amer.*, vol. 58, no. 1, pp. 214–222, 1975.

[42] N. A. Gumerov, R. Duraiswami, and D. N. Zotkin, “Fast multipole accelerated boundary elements for numerical computation of the head related transfer function,” in *Proc. 2007 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 1, pp. I–165.

[43] P. Mokhtari, R. Nishimura, and H. Takemoto, “Toward HRTF personalization: An auditory-perceptual evaluation of simulated and measured HRTFs,” in *Proc. 14th Int. Conf. Auditory Display*, Int. Community Auditory Display, 2008, pp. 1–8.

[44] R. Pelzer, M. Dinakaran, F. Brinkmann, S. Lepa, P. Grosche, and S. Weinzierl, “Head-related transfer function recommendation based on perceptual similarities and anthropometric features,” *J. Acoustical Soc. Amer.*, vol. 148, no. 6, pp. 3809–3817, 2020.

[45] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, “HRTF personalization using anthropometric measurements,” in *Proc. 2003 IEEE Workshop Appl. Signal Process. Audio Acoust. (IEEE Cat. No 03TH8684)*, 2003, pp. 157–160.

[46] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini, “Applying a single-notch metric to image-guided head-related transfer function selection for improved vertical localization,” *J. Audio Eng. Soc.*, vol. 67, no. 6, pp. 1–15, Jun. 2019.

[47] B. U. Seeber and H. Fastl, “Subjective selection of non-individual head-related transfer functions,” in *Proc. 9th Int. Conf. Auditory Display, Georgia Inst. Technol.*, 2003, pp. 259–262.

[48] B. F. G. Katz and G. Parseihian, “Perceptually based head-related transfer function database optimization,” *J. Acoustical Soc. Amer.*, vol. 131, no. 2, pp. EL99–105, Feb. 2012.

[49] J. C. Middlebrooks, E. A. Macpherson, and Z. A. Onsan, “Psychophysical customization of directional transfer functions for virtual sound localization,” *J. Acoustical Soc. Amer.*, vol. 108, no. 6, pp. 3088–3091, 2000.

[50] 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.

[51] V. Gudivada, A. Apon, and J. Ding, “Data quality considerations for Big Data and machine learning: Going beyond data cleaning and transformations,” *Int. J. Adv. Softw.*, vol. 10, no. 1, pp. 1–20, 2017.

[52] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*, 4th ed. Cambridge, MA, USA: Elsevier Science, 2022.

[53] A. Kulkarni, S. Isabelle, and H. Colburn, “On the minimum-phase approximation of head-related transfer functions,” in *Proc. 1995 IEEE Workshop Appl. Signal Process. Audio Acoust.*, 1995, pp. 84–87.

[54] N. R. Haddaway, M. J. Page, C. C. Pritchard, and L. A. McGuinness, “PRISMA2020: An R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis,” *Campbell Systematic Rev.*, vol. 18, no. 2, Jun. 2022, Art. no. e1230.

[55] M. J. Page et al., “PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews,” *BMJ*, vol. 372, pp. 1–36, 2021.

[56] K. J. Faller II and K. P. Hoang, “Estimation of parameters of a head-related transfer function (HRTF) customization model,” in *Proc. Meetings Acoust.*, vol. 18. AIP Publishing, 2012, pp. 1–8.

[57] 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*. New York, NY, USA: Audio Engineering Society, May 2016.

[58] M. T. Islam and I. Tashev, “Anthropometric features estimation using integrated sensors on a headphone for HRTF personalization,” in *Proc. Audio Eng. Soc. Conf.: 2020 AES Int. Conf. Audio Virtual Augmented Reality*, Audio Engineering Society, 2020, pp. 1–10.

[59] E. A. Torres-Gallegos, F. Orduña-Bustamante, and F. Arámbula-Cosío, “Personalization of head-related transfer functions (HRTF) based on automatic photo-anthropometry and inference from a database,” *Appl. Acoust.*, vol. 97, pp. 84–95, 2015.

[60] D. Fantini, F. Avanzini, S. Ntalampiras, and G. Presti, “HRTF individualization based on anthropometric measurements extracted from 3D head meshes,” in *Proc. 2021 IEEE Immersive 3D Audio: From Architecture Automot.*, 2021, pp. 1–10.

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

[62] M. B. Gardner and R. S. Gardner, “Problem of localization in the median plane: Effect of pinnae cavity occlusion,” *J. Acoustical Soc. Amer.*, vol. 53, no. 2, pp. 400–408, 1973.

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

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

[65] K. Iida, O. Nishiyama, and T. Aizaki, “Estimation of the category of notch frequency bins of the individual head-related transfer functions using the anthropometry of the listener’s pinnae,” *Appl. Acoust.*, vol. 177, 2021, Art. no. 107929.

[66] B. Xie, X. Zhong, D. Rao, and Z. Liang, “Head-related transfer function database and its analyses,” *Sci. China Ser. G: Phys., Mechan. Astron.*, vol. 50, no. 3, pp. 267–280, 2007.

[67] N. S. A. o. C. GB/T 2428-1998, “GB/T 2428-1998 standard head-face dimensions of adults,” Gen. Admin. Qual. Supervision Inspection China, Tech. Rep., 1998.

[68] M. S. P. H. GJB 4856-2003, “GJB 4856-2003: Human dimensions of Chinese male pilot population,” Chinese People’s Liberation Army Gen. Armament Dept., Beijing, Tech. Rep. GJB4856, 2003.

[69] P. Stitt and B. F. G. Katz, “Sensitivity analysis of pinna morphology on head-related transfer functions simulated via a parametric pinna model,” *J. Acoustical Soc. Amer.*, vol. 149, no. 4, pp. 2559–2572, 2021.

[70] Y. Teng and X. Zhong, “An individualized HRTF model based on random forest and anthropometric parameters,” in *Proc. 2023 15th Int. Conf. Intell. Hum.- Mach. Syst. Cybern.*, 2023, pp. 143–146.

[71] D. Fantini, S. Ntalampiras, G. Presti, and F. Avanzini, “Toward a novel set of pinna anthropometric features for individualizing head-related transfer function,” in *Proc. 21th Sound Music Comput. Conf.*, Porto, Portugal, Jul. 2024, pp. 285–292.

[72] C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile, “Enabling individualized virtual auditory space using morphological measurements,” in *Proc. 1st IEEE Pacific-Rim Conf. Multimedia (2000 Int. Symp. Multimedia Inf. Process.)*, 2000, pp. 235–238.

[73] D. Lu, X. Zeng, X. Guo, and H. Wang, “Personalization of head-related transfer function based on sparse principle component analysis and sparse representation of 3D anthropometric parameters,” *Acoust. Aust.*, vol. 48, no. 1, pp. 49–58, 2020.

[74] D. Lu, X. Zeng, X. Guo, and H. Wang, “Head-related transfer function reconstruction with anthropometric parameters and the direction of the sound source,” *Acoust. Aust.*, vol. 49, no. 1, pp. 125–132, 2021.

[75] D. Lu, J. Zhang, H. Gao, and C. Liu, “Personalized HRIR based on PointNet network using anthropometric parameters,” in *Proc. Int. Conf. Man- Mach.- Environ. Syst. Eng.*, Springer, 2023, pp. 54–59.

[76] G. W. Lee and H. K. Kim, “Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear,” *Appl. Sci.*, vol. 8, no. 11, 2018, Art. no. 2180.

[77] M. Zhao, Z. Sheng, and Y. Fang, “Magnitude modeling of personalized HRTF based on ear images and anthropometric measurements,” *Appl. Sci.*, vol. 12, no. 16, 2022, Art. no. 8155.

[78] Ž. Emeršič, V. Štruc, and P. Peer, “Ear recognition: More than a survey,” *Neurocomputing*, vol. 255, pp. 26–39, 2017.

[79] R. Miccini and S. Spagnol, “A hybrid approach to structural modeling of individualized HRTFs,” in *Proc. 2021 IEEE Conf. Virtual Reality 3D User Interfaces Abstr. Workshops*, 2021, pp. 80–85.

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

[81] R. Baumgartner, P. Majdak, and B. Laback, “Modeling sound-source localization in sagittal planes for human listeners,” *J. Acoustical Soc. Amer.*, vol. 136, no. 2, pp. 791–802, 2014.

[82] B.-Y. Ko, G.-T. Lee, H. Nam, and Y.-H. Park, “PRTFNet: HRTF individualization for accurate spectral cues using a compact PRTF,” *IEEE Access*, vol. 11, pp. 96119–96130, 2023.

[83] Y. Zhou, H. Jiang, and V. K. Ithapu, “On the predictability of HRTFs from ear shapes using deep networks,” in *Proc. ICASSP 2021-2021 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 441–445.

[84] T.-Y. Chen, T.-H. Kuo, and T.-S. Chi, “Autoencoding HRTFs for DNN based HRTF personalization using anthropometric features,” in *Proc. ICASSP 2019-2019 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 271–275.

[85] M. Zhang, Z. Ge, T. Liu, X. Wu, and T. Qu, “Modeling of individual HRTFs based on spatial principal component analysis,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 785–797, 2020.

[86] J. Zhao, D. Yao, J. Gu, and J. Li, “Efficient prediction of individual head-related transfer functions based on 3D meshes,” *Appl. Acoust.*, vol. 219, 2024, Art. no. 109938.

[87] Y. Luo, D. N. Zotkin, and R. Duraiswami, “Virtual autoencoder based recommendation system for individualizing head-related transfer functions,” in *Proc. 2013 IEEE Workshop Appl. Signal Process. to Audio Acoust.*, 2013, pp. 1–4.

[88] K. Yamamoto and T. Igarashi, “Fully perceptual-based 3D spatial sound individualization with an adaptive variational autoencoder,” *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–13, Nov. 2017.

[89] S. Hwang, Y. Park, and Y.-s. Park, “Customization of spatially continuous head-related impulse responses in the median plane,” *Acta Acustica United With Acustica*, vol. 96, no. 2, pp. 351–363, 2010.

[90] N. H. Zandi, A. M. El-Mohandes, and R. Zheng, “Individualizing head-related transfer functions for binaural acoustic applications,” in *Proc. 21st ACM/IEEE Int. Conf. Inf. Process. Sensor Netw.*, 2022, pp. 105–117.

[91] V. Jayaram, I. Kemelmacher-Shlizerman, and S. M. Seitz, “HRTF estimation in the wild,” in *Proc. 36th Annu. ACM Symp. User Interface Softw. Technol.*, New York, NY, USA: Assoc. Comput. Machinery, 2023, pp. 1–9.

[92] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, “Dataset of head-related transfer functions measured with a circular loudspeaker array,” *Acoustical Sci. Technol.*, vol. 35, no. 3, pp. 159–165, 2014.

[93] N. Javeri, P. B. Dutta, K. Sunder, and K. Jain, “A machine learning approach to predicting personalized head related transfer functions and headphone equalization from video capture data,” in *Proc. 2023 IEEE Immersive 3D Audio: From Architecture Automot.*, 2023, pp. 1–9.

[94] 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. 2014 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4468–4472.

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

[96] M. Zhu, M. Shahnawaz, S. Tubaro, and A. Sarti, "HRTF personalization based on weighted sparse representation of anthropometric features," in *Proc. 2017 Int. Conf. 3D Immersion*, 2017, pp. 1–7.

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

[98] Q.-H. Huang and Y. Fang, "Modeling personalized head-related impulse response using support vector regression," *J. Shanghai Univ. (English Edition)*, vol. 13, no. 6, pp. 428–432, 2009.

[99] Q. Huang and Q. Zhuang, "HRIR personalisation using support vector regression in independent feature space," *Electron. Lett.*, vol. 45, no. 19, 2009, Art. no. 1.

[100] Y. Tang, Y. Fang, and Q. Huang, "Audio personalization using head related transfer function in 3DTV," in *Proc. 2011 3DTV Conf.: True Vis.-Capture, Transmiss. Display 3D Video*, 2011, pp. 1–4.

[101] R. Fernandez Martinez, P. Jimbert, E. M. Sumner, M. Riedel, and R. Unnthorsson, "Prediction of head related transfer functions using machine learning approaches," *Acoustics*, vol. 5, no. 1, pp. 254–267, Mar. 2023.

[102] L. Wang, X. Zeng, and X. Ma, "Advancement of individualized head-related transfer functions (HRTFs) in perceiving the spatialization cues: Case study for an integrated HRTF individualization method," *Appl. Sci.*, vol. 9, no. 9, pp. 1–9, 2019.

[103] T.-Y. Chen, P.-W. Hsiao, and T.-S. Chi, "Exploring redundancy of HRTFs for fast training DNN-based HRTF personalization," in *Proc. 2018 IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2018, pp. 1929–1933.

[104] D. Yao et al., "An individualization approach for head-related transfer function in arbitrary directions based on deep learning," *JASA Exp. Lett.*, vol. 2, no. 6, 2022, Art. no. 064401.

[105] J. Xi, W. Zhang, and T. D. Abhayapala, "Magnitude modelling of individualized HRTFs using DNN based spherical harmonic analysis," in *Proc. 2021 IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2021, pp. 266–270.

[106] J. S. Son and S. H. Choi, "A DNN-based personalized HRTF estimation method for 3D immersive audio," *Int. J. Internet, Broadcast. Commun.*, vol. 13, no. 1, pp. 161–167, 2021.

[107] R. Zhang, R. Meng, J. Sang, Y. Hu, X. Li, and C. Zheng, "Modelling individual head-related transfer function (HRTF) based on anthropometric parameters and generic HRTF amplitudes," *CAAI Trans. Intell. Technol.*, vol. 8, no. 2, pp. 364–378, 2023.

[108] 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.

[109] Z. Wang and C. F. Chan, "HRIR customization using common factor decomposition and joint support vector regression," in *Proc. 21st IEEE Eur. Signal Process. Conf.*, 2013, pp. 1–5.

[110] D. Lu, J. Zhu, Z. Shen, C. Liu, and Q. Xin, "A hybrid algorithm for predicting head related transfer function based on physiological parameters," in *Proc. J. Phys.: Conf. Ser.*, vol. 2784. IOP Publishing, 2024, Art. no. 012015.

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

[112] K. Iida, M. Yairi, and M. Morimoto, "Role of pinna cavities in median plane localization," *J. Acoustical Soc. Amer.*, vol. 103, no. 5 Supplement, pp. 2844–2844, 1998.

[113] M. Zhang, R. Kennedy, T. Abhayapala, and W. Zhang, "Statistical method to identify key anthropometric parameters in HRTF individualization," in *Proc. IEEE Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, 2011, pp. 213–218.

[114] R. Wu, G. Yu, and R. H. So, "Key anthropometric parameters of pinna correlate with individualized head-related transfer functions," in *Proc. INTER-NOISE NOISE-CON Congr. Conf. Proc.*, vol. 255. Institute of Noise Control Engineering, 2017, pp. 4023–4028.

[115] K. Watanabe, K. Ozawa, Y. Iwaya, Y. Suzuki, and K. Aso, "Estimation of interaural level difference based on anthropometry and its effect on sound localization," *J. Acoustical Soc. Amer.*, vol. 122, no. 5, pp. 2832–2841, 2007.

[116] S. Spagnol and F. Avanzini, "Anthropometric tuning of a spherical head model for binaural virtual acoustics based on interaural level differences," in *Proc. 21st Int. Conf. Auditory Display*, Georgia Inst. Technol., 2015, pp. 204–209.

[117] 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," *J. Sound Vib.*, vol. 329, no. 19, pp. 4093–4106, 2010.

[118] C. S. Reddy and R. M. Hegde, "A joint sparsity and linear regression based method for customization of median plane HRIR," in *Proc. 2015 49th Asilomar Conf. Signals, Syst. Comput.*, 2015, pp. 785–789.

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

[120] S. S. Alotaibi and M. Wickert, "Modeling of individual head-related transfer functions (HRTFs) based on spatiotemporal and anthropometric features using deep neural networks," *IEEE Access*, vol. 12, pp. 14620–14635, 2024.

[121] R. Bomhardt, H. Braren, and J. Fels, "Individualization of head-related transfer functions using principal component analysis and anthropometric dimensions," in *Proc. Meetings Acoust. 172ASA*, vol. 29, Acoustical Society of America, 2016, Art. no. 050007.

[122] L. Wang and X. Y. Zeng, "New method for synthesizing personalized head-related transfer function," in *Proc. 2016 IEEE Int. Workshop Acoust. Signal Enhancement*, 2016, pp. 1–5.

[123] S. Spagnol, "HRTF selection by anthropometric regression for improving horizontal localization accuracy," *IEEE Signal Process. Lett.*, vol. 27, pp. 590–594, 2020.

[124] H. Hu, L. Zhou, J. Zhang, H. Ma, and Z. Wu, "Head related transfer function personalization based on multiple regression analysis," in *Proc. 2006 Int. Conf. Comput. Intell. Secur.*, vol. 2, 2006, pp. 1829–1832.

[125] S. Xu, Z. Li, and G. Salvendy, "Identification of anthropometric measurements for individualization of head-related transfer functions," *Acta Acustica United With Acustica*, vol. 95, no. 1, pp. 168–177, 2009.

[126] L. Li and Q. Huang, "HRTF personalization modeling based on RBF neural network," in *Proc. 2013 IEEE Int. Conf. Acoust., Speech Signal Process*, 2013, pp. 3707–3710.

[127] Q. Huang and L. Li, "Modeling individual HRTF tensor using high-order partial least squares," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, pp. 1–14, 2014.

[128] W. W. Hugeng and D. Gunawan, "Improved method for individualization of head-related transfer functions on horizontal plane using reduced number of anthropometric measurements," *J. Telecommun.*, vol. 2, pp. 31–41, May 2010.

[129] H. Hugeng, W. Wahab, and D. Gunawan, "A new selection method of anthropometric parameters in individualizing HRIR," *TELKOMNIKA (Telecommun. Comput. Electron. Control)*, vol. 13, no. 3, pp. 1014–1020, 2015.

[130] D. Schönstein and B. F. G. Katz, "HRTF selection for binaural synthesis from a database using morphological parameters," in *Proc. 20th Int. Congr. Acoust.*, 2010, pp. 1–6.

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

[132] H. Fayek, L. van der Maaten, G. Romigh, and R. Mehra, "On data-driven approaches to head-related-transfer function personalization," in *Audio Engineering Society Convention 143*. New York, NY, USA: Audio Eng. Soc., Oct. 2017.

[133] N. Gupta, A. Barreto, and M. Choudhury, "Modeling head-related transfer functions based on pinna anthropometry," in *Proc. 2nd Int. Latin Amer. Caribbean Conf. Eng. Technol.*, Boca Raton, FL, USA, 2004.

[134] W. W. Hugeng and D. Gunawan, "Enhanced individualization of head-related impulse response model in horizontal plane based on multiple regression analysis," in *Proc. IEEE 2nd Int. Conf. Comput. Eng. Appl.*, vol. 2, 2010, pp. 226–230.

[135] I. Nambu et al., "Reinforcement-learning-based personalization of head-related transfer functions," *J. Audio Eng. Soc.*, vol. 66, no. 5, pp. 317–328, 2018.

[136] W. Chen, R. Hu, X. Wang, C. Yang, and L. Meng, "Individualization of head related impulse responses using division analysis," *China Commun.*, vol. 15, no. 5, pp. 92–103, 2018.

[137] S. Angelucci, C. Rinaldi, F. Franchi, and F. Graziosi, “Comparison of ML solutions for HRIR individualization design in binaural audio,” in *Proc. Int. Conf. Adv. Inf. Netw. Appl*, Springer, 2023, pp. 271–278.

[138] M. Rothbucher, M. Durkovic, H. Shen, and K. Diepold, “HRTF customization using multiway array analysis,” in *Proc. IEEE 18th Eur. Signal Process. Conf.*, 2010, pp. 229–233.

[139] H. Hugeng, W. Wahab, and D. Gunawan, “The effectiveness of chosen partial anthropometric measurements in individualizing head-related transfer functions on median plane,” *J. ICT Res. Appl.*, vol. 5, no. 1, pp. 35–56, Sep. 2011.

[140] F. Grijalva, L. C. Martini, B. Masiero, and S. Goldenstein, “A recommender system for improving median plane sound localization performance based on a nonlinear representation of HRTFs,” *IEEE Access*, vol. 6, pp. 24829–24836, 2018.

[141] S. G. Rodríguez and M. A. Ramírez, “Linear relationships between spectral characteristics and anthropometry of the external ear,” in *Proc. ICAD 05-11th Meeting Int. Conf. Auditory Display*, 2005.

[142] F. C. Tommasini, O. A. Ramos, M. X. Hüg, and F. Bermejo, “Usage of spectral distortion for objective evaluation of personalized HRTF in the median plane,” *Int. J. Acoust. Vib.*, vol. 20, no. 2, pp. 81–89, 2015.

[143] K. Iida, H. Shimazaki, and M. Oota, “Generation of the individual head-related transfer functions in the upper median plane based on the anthropometry of the listener’s pinnae,” in *Proc. IEEE 7th Glob. Conf. Consum. Electron.*, Oct. 2018, pp. 79–80.

[144] K. Iida, H. Shimazaki, and M. Oota, “Generation of the amplitude spectra of the individual head-related transfer functions in the upper median plane based on the anthropometry of the listener’s pinnae,” *Appl. Acoust.*, vol. 155, pp. 280–285, 2019.

[145] X. Liu, H. Song, and X. Zhong, “A hybrid algorithm for predicting median-plane head-related transfer functions from anthropometric measurements,” *Appl. Sci.*, vol. 9, no. 11, 2019, Art. no. 2323.

[146] X. Liu, W. Huang, H. Zhang, and X. Zhong, “Median-plane head-related transfer function personalization using two-dimensional independent component analysis,” in *Proc. IEEE 8th Int. Conf. Comput. Commun.*, 2022, pp. 2308–2312.

[147] F. Grijalva, L. Martini, S. Goldenstein, and D. Florencio, “Anthropometric-based customization of head-related transfer functions using isomap in the horizontal plane,” in *Proc. 2014 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4473–4477.

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

[149] G. Grindlay and M. A. O. Vasilescu, “A multilinear (tensor) framework for HRTF analysis and synthesis,” in *Proc. 2007 IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, 2007, pp. I–161–I–164.

[150] R. Miccini and S. Spagnol, “HRTF individualization using deep learning,” in *Proc. 2020 IEEE Conf. Virtual Real. 3D User Interfaces Abstr. Workshops*, 2020, pp. 390–395.

[151] R. B. King and S. R. Oldfield, “The impact of signal bandwidth on auditory localization: Implications for the design of three-dimensional audio displays,” *Hum. Factors*, vol. 39, no. 2, pp. 287–295, 1997.

[152] S. Xu, Z. Li, and G. Salvendy, “Improved method to individualize head-related transfer function using anthropometric measurements,” *Acoustical Sci. Technol.*, vol. 29, no. 6, pp. 388–390, 2008.

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

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

[155] M. Zhang, X. Wu, and T. Qu, “Individual distance-dependent HRTFs modeling through a few anthropometric measurements,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2020)*, 2020, pp. 401–405.

[156] M. Zhang, J.-H. Wang, and D. L. James, “Personalized HRTF modeling using DNN-augmented BEM,” in *Proc. ICASSP 2021-2021 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 451–455.

[157] W. Chen, H. Zhang, J. Yu, and F. Luo, “Individualization of head related transfer function based on PCA and RBF network,” in *Proc. 11th Int. Conf. Netw., Commun. Comput.*, New York, NY, USA: Assoc. Comput. Machinery, 2023, pp. 109–116.

[158] Z. Liang, B. Xie, and X. Zhong, “Comparison of principal components analysis on linear and logarithmic magnitude of head-related transfer functions,” in *Proc. IEEE 2nd Int. Congr. Image Signal Process.*, 2009, pp. 1–5.

[159] S. Takane, “Effect of domain selection for compact representation of spatial variation of head-related transfer function in all directions based on spatial principal components analysis,” *Appl. Acoust.*, vol. 101, pp. 64–77, 2016.

[160] G. Marentakis, “Principal component analysis of binaural HRTF pairs,” in *Proc. 20th Sound Music Comput. Conf.*, Stockholm, Sweden, Jun. 2023, pp. 178–185.

[161] G. Marentakis and J. Hôlzl, “Compression efficiency and signal distortion of common PCA bases for HRTF modelling,” in *Proc. 18th Sound Music Comput. Conf.*, Axea sas/SMC Network, 2021, pp. 60–67.

[162] M. Rothbucher, H. Shen, and K. Diepold, “Dimensionality reduction in HRTF by using multiway array analysis,” in *Human Centered Robot Systems: Cognition, Interaction, Technology*. Berlin, Germany: Springer, 2009, pp. 103–110.

[163] C. J. Chun, J. M. Moon, G. W. Lee, N. K. Kim, and H. K. Kim, “Deep neural network based HRTF personalization using anthropometric measurements,” in *Audio Engineering Society Convention 143*. Boca Raton, FL, USA: Audio Eng. Soc., Oct. 2017.

[164] D. Zurale and S. Dubnov, “Learning sub-dimensional HRTF representations towards individualization applications-traditional and deep learning approaches,” in *Proc. 2023 IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2023, pp. 1–5.

[165] Y. Bengio, J.-f. Paiement, P. Vincent, O. Delalleau, N. Roux, and M. Ouimet, “Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering,” in *Proc. Adv. Neural Inf. Process. Syst.*, S. Thrun, L. Saul, and B. Schölkopf, Eds., vol. 16. MIT Press, 2003.

[166] B. Kapralos and N. Mekuz, “Application of dimensionality reduction techniques to HRTFs for interactive virtual environments,” in *Proc. Int. Conf. Adv. Comput. Entertainment Technol.*, 2007, pp. 256–257.

[167] B. Kapralos, N. Mekuz, A. Kopinska, and S. Khattak, “Dimensionality reduced HRTFs: A comparative study,” in *Proc. 2008 Int. Conf. Adv. Comput. Entertainment Technol.*, New York, NY, USA: Assoc. Comput. Machinery, 2008, pp. 59–62.

[168] N. Inoue, T. Kimura, T. Nishino, K. Itou, and K. Takeda, “Evaluation of HRTFs estimated using physical features,” *Acoustical Sci. Technol.*, vol. 26, no. 5, pp. 453–455, 2005.

[169] M. Geronazzo, S. Spagnol, A. Bedin, and F. Avanzini, “Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions,” in *Proc. 2014 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4463–4467.

[170] X. Zeng, S. Wang, and L. Ga, “Individualization of head-related transfer function based on anthropometric parameters selection and database matching,” *J. Tech. Acoust.*, vol. 28, pp. 16–20, 2009.

[171] D. Lu, X. Zeng, X. Guo, and H. Wang, “Head-related transfer function personalization based on modified sparse representation with matching in a database of Chinese pilots,” *Acoust. Aust.*, vol. 48, no. 3, pp. 463–471, Dec. 2020.

[172] G. W. Lee, H. K. Kim, C. J. Chun, and K. M. Jeon, “Personalized HRTF estimation based on one-to-many neural network architecture,” in *Audio Engineering Society Convention 153*. New York, NY, USA: Audio Eng. Soc., Oct. 2022.

[173] L. Meng, X. Wang, W. Chen, C. Ai, and R. Hu, “Individualization of head related transfer functions based on radial basis function neural network,” in *Proc. 2018 IEEE Int. Conf. Multimedia Expo*, 2018, pp. 1–6.

[174] X. Qi and L. Wang, “Parameter-transfer learning for low-resource individualization of head-related transfer functions,” in *Proc. INTERSPEECH*, 2019, pp. 3865–3869.

[175] Y. Qiu, Z. Li, and J. Wang, “Individual HRTF prediction based on anthropometric data and multi-stage model,” in *Proc. 2023 IEEE Int. Conf. Multimedia Expo Workshops*, 2023, pp. 314–319.

[176] Y. Wang, Y. Zhang, Z. Duan, and M. Bocko, “Global HRTF personalization using anthropometric measures,” in *Audio Engineering Society Convention 150*. New York, NY, USA: Audio Eng. Soc., May 2021.

[177] Y. Qiu, J. Wang, and Z. Li, “Personalized HRTF prediction based on LightGBM using anthropometric data,” *China Commun.*, vol. 20, no. 6, pp. 166–177, 2023.

[178] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction, 2nd Ed., Ser. Adaptive Computation and Machine Learning Series*. Cambridge, MA, USA: MIT Press, 2018.

[179] N. Inoue, T. Nishino, K. Itou, and K. Takeda, “HRTF modeling using physical features,” in *Proc. Forum Acusticum 2005*, 2005, pp. 199–202.

[180] A. Andreopoulou, D. R. Begault, and B. F. G. Katz, “Inter-laboratory round robin HRTF measurement comparison,” *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 895–906, Aug. 2015.

[181] R. Barumerli, M. Geronazzo, and F. Avanzini, “Round robin comparison of inter-laboratory HRTF measurements–assessment with an auditory model for elevation,” in *Proc. IEEE 4th VR Workshop Sonic Interact. Virtual Environments*, Reutlingen, Germany: IEEE Comput. Soc., Mar. 2018, pp. 1–5.

[182] A. Andreopoulou and A. Roginska, “Towards the creation of a standardized HRTF repository,” in *Audio Engineering Society Convention 131*, Audio Engineering Society, 2011.

[183] Y. Luo, D. N. Zotkin, and R. Duraiswami, “Gaussian process data fusion for heterogeneous HRTF datasets,” in *Proc. 2013 IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2013, pp. 1–4.

[184] B. Tsui and G. Kearney, “A head-related transfer function database consolidation tool for high variance machine learning algorithms,” in *Audio Engineering Society Convention 145*. New York, NY, USA: Audio Eng. Soc., Oct. 2018.

[185] Y. Wen, Y. Zhang, and Z. Duan, “Mitigating cross-database differences for learning unified HRTF representation,” in *Proc. 2023 IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2023, pp. 1–5.

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

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

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

[189] 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.

[190] D. Poirier-Quinot, M. S. Lawless, P. Stitt, and B. F. Katz, “HRTF performance evaluation: Methodology and metrics for localisation accuracy and learning assessment,” in *Proc. Adv. Fundam. Appl. Res. Spatial Audio*, B. F. Katz and P. Majdak, Eds. Rijeka: IntechOpen, 2022, ch. 3.

[191] P. Majdak, C. Hollomey, and R. Baumgartner, “Amt 1. x: A toolbox for reproducible research in auditory modeling,” *Acta Acustica*, vol. 6, 2022, Art. no. 19.

[192] M. Dietz, S. D. Ewert, and V. Hohmann, “Auditory model based direction estimation of concurrent speakers from binaural signals,” *Speech Commun.*, vol. 53, no. 5, pp. 592–605, 2011.

[193] H. Wierstorf, A. Raake, and S. Spors, *Binaural Assessment of Multichannel Reproduction*. Berlin, Germany: Springer, 2013, pp. 255–278.

[194] R. Barumerli, P. Majdak, M. Geronazzo, D. Meijer, F. Avanzini, and R. Baumgartner, “A Bayesian model for human directional localization of broadband static sound sources,” *Acta Acustica*, vol. 7, 2023, Art. no. 12.

[195] S. Ntalampiras, “One-shot learning for acoustic diagnosis of industrial machines,” *Expert Syst. With Appl.*, vol. 178, Sep. 2021, Art. no. 114984.

[196] S. Ntalampiras and A. Scalambrino, “Automatic prediction of disturbance caused by inter-floor sound events,” *IEEE Trans. Cogn. Devlop. Syst.*, early access, Jul. 08, 2024, doi: 10.1109/TCDS.2024.3424457.

[197] S. Ntalampiras, “Speech emotion recognition via learning analogies,” *Pattern Recognit. Lett.*, vol. 144, pp. 21–26, 2021.

[198] S. Ntalampiras, “Explainable Siamese neural network for classifying pediatric respiratory sounds,” *IEEE J. Biomed. Health Informat.*, vol. 27, no. 10, pp. 4728–4735, 2023.

[199] S. Ntalampiras, “A transfer learning framework for predicting the emotional content of generalized sound events,” *J. Acoustical Soc. Amer.*, vol. 141, no. 3, pp. 1694–1701, Mar. 2017.

[200] J. Pauwels and L. Picinali, “On the relevance of the differences between HRTF measurement setups for machine learning,” in *Proc. ICASSP 2023-2023 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[201] K. A. Riederer, “Repeatability analysis of head-related transfer function measurements,” in *Audio Engineering Society Convention 105*. Audio Eng. Soc., 1998.

[202] X.-l. Zhong and B.-s. Xie, “Consistency among the head-related transfer functions from different measurements,” in *Proc. Meetings Acoust.*, vol. 19. Acoustical Soc. Amer., 2013, Art. no. 050014.

[203] A. Andreopoulou, A. Rogińska, and H. Mohanraj, “A database of repeated head-related transfer function measurements,” in *Proc. 19th Int. Conf. Auditory Display*, Georgia Inst. Technol., 2013.

[204] R. Greff and B. F. Katz, “Round robin comparison of HRTF simulation systems: Preliminary results,” in *Audio Engineering Society Convention 123*. New York, NY, USA: Audio Eng. Soc., 2007.

[205] J. Pauwels, “The Hartufo toolkit for machine learning with HRTF data,” in *Proc. Audio Eng. Soc. Conf.: AES 2023 Int. Conf. Spatial Immersive Audio*, Audio Engineering Society, 2023, pp. 1–12.

[206] N. Gupta, A. Barreto, M. Joshi, and J. C. Agudelo, “HRTF database at FIU DSP lab,” in *Proc. 2010 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 169–172.

[207] P. Majdak and M. Mihocic, “ARI HRTF database,” 2016, Acoustics Research Institute (ARI). [Online]. Available: http://www.kfs.oeaw.ac.at/hrtf

[208] R. Bomhardt, M. de la Fuente Klein, and J. Fels, “A high-resolution head-related transfer function and three-dimensional ear model database,” in *Proc. Meetings Acoust.*, vol. 29. AIP Publishing, 2016, Art. no. 050002.

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

[210] K. Pollack, F. Pausch, and P. Majdak, “Parametric pinna model for a realistic representation of listener-specific pinna geometry,” in *Proc. 24th Int. Congr. Acoust.: ICA*, Gyeongju, 2022, pp. 168–178.

[211] S. T. Prepelit, ă, J. Gómez Bolaños, M. Geronazzo, R. Mehra, and L. Savioja, “Pinna-related transfer functions and lossless wave equation using finite-difference methods: Validation with measurements,” *J. Acoustical Soc. Amer.*, vol. 147, no. 5, pp. 3631–3645, 2020.

[212] I. Ananthabhotla, V. K. Ithapu, and W. O. Brimijoin, “A framework for designing head-related transfer function distance metrics that capture localization perception,” *JASA Exp. Lett.*, vol. 1, no. 4, 2021, Art. no. 044401.

[213] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, “A spatial audio quality inventory (SAQI),” *Acta Acustica United With Acustica*, vol. 100, no. 5, pp. 984–994, 2014.

[214] C. Jenny and C. Reuter, “Can I trust my ears in VR? Literature review of head-related transfer functions and valuation methods with descriptive attributes in virtual reality,” *Int. J. Virtual Reality*, vol. 21, no. 2, pp. 29–43, Oct. 2021.

[215] M. Oehler, M. do VM da Costa, M. Regener, and T. Minh Voong, “Relevance of individual numerically simulated head-related transfer functions for different scenarios in virtual environments,” in *Proc. Audio Eng. Soc. Conf.: 2022 AES Int. Conf. Audio Virtual Augmented Reality*, Audio Eng. Soc., 2022.

[216] A. Lindau and S. Weinzierl, “Assessing the plausibility of virtual acoustic environments,” *Acta Acustica United With Acustica*, vol. 98, no. 5, pp. 804–810, 2012.

[217] S. A. Wirler, N. Meyer-Kahlen, and S. J. Schlecht, “Towards transfer-plausibility for evaluating mixed reality audio in complex scenes,” in *Proc. Audio Eng. Soc. Conf.: 2020 AES Int. Conf. Audio Virtual Augmented Reality*, Audio Eng. Soc., Aug. 2020.

[218] G. C. Stecker, T. M. Moore, M. Folkerts, D. Zotkin, and R. Duraiswami, “Toward objective measures of auditory co-immersion in virtual and augmented reality,” in *Proc. Audio Eng. Soc. Conf.: 2018 AES Int. Conf. Audio Virtual Augmented Reality*, Audio Eng. Soc., Aug. 2018, pp. 1–6.

[219] S. Serafin, M. Geronazzo, C. Erkut, N. C. Nilsson, and R. Nordahl, “Sonic interactions in virtual reality: State of the art, current challenges, and future directions,” *IEEE Comput. Graph. Appl.*, vol. 38, no. 2, pp. 31–43, Mar./Apr. 2018.

[220] P. Milgram and F. Kishino, “A taxonomy of mixed reality visual displays,” *IEICE Trans. Inf. Syst.*, vol. 77, no. 12, pp. 1321–1329, 1994.

[221] J. Yang, A. Barde, and M. Billinghurst, “Audio augmented reality: A systematic review of technologies, applications, and future research directions,” *J. Audio Eng. Soc.*, vol. 70, no. 10, pp. 788–809, Oct. 2022.

[222] D. Fantini, G. Presti, M. Geronazzo, R. Bona, A. G. Privitera, and F. Avanzini, "Co-immersion in audio augmented virtuality: The case study of a static and approximated late reverberation algorithm," *IEEE Trans. Visual. Comput. Graph.*, vol. 29, no. 11, pp. 4472–4482, Nov. 2023.

[223] W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd, "The contribution of head movement to the externalization and internalization of sounds," *PLoS One*, vol. 8, no. 12, 2013, Art. no. e83068.

[224] J. Catic, S. Santurette, and T. Dau, "The role of reverberation-related binaural cues in the externalization of speech," *J. Acoustical Soc. Amer.*, vol. 138, no. 2, pp. 1154–1167, 2015.

[225] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *Proc. IEEE 8th Int. Conf. Qual. Multimedia Experience*, 2016, pp. 1–6.

[226] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. Katz, and C. de Boishéraud, "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *J. Acoustical Soc. Amer.*, vol. 141, no. 3, pp. 2011–2023, 2017.

[227] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo, "Sound externalization: A review of recent research," *Trends Hear.*, vol. 24, 2020, Art. no. 2331216520948390.

[228] B. A. Wright and Y. Zhang, "A review of learning with normal and altered sound-localization cues in human adults: Revisión del aprendizaje en adultos con claves de localización sonora normales o alteradas," *Int. J. Audiol.*, vol. 45, no. sup1, pp. 92–98, 2006.

[229] C. Mendonça, G. Campos, P. Dias, J. Vieira, J. P. Ferreira, and J. A. Santos, "On the improvement of localization accuracy with non-individualized HRTF-based sounds," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 821–830, 2012.

[230] G. Parseihian and B. F. G. Katz, "Rapid head-related transfer function adaptation using a virtual auditory environment," *J. Acoustical Soc. Amer.*, vol. 131, no. 4, pp. 2948–2957, 2012.

[231] C. Mendonça, "A review on auditory space adaptations to altered head-related cues," *Front. Neurosci.*, vol. 8, 2014, Art. no. 92880.

[232] F. Klein and S. Werner, "Auditory adaptation to non-individual HRTF cues in binaural audio reproduction," *J. Audio Eng. Soc.*, vol. 64, no. 1/2, pp. 45–54, 2016.

[233] L. Picinali and B. F. G. Katz, *System-to-User and User-to-System Adaptations in Binaural Audio*. Cham, Switzerland: Springer, 2023, pp. 115–143.

DAVIDE FANTINI received the M.Sc. and Ph.D. degrees in computer science from the University of Milan, Milan, Italy, in 2019 and 2024, respectively. He is currently a Research Fellow with the Department of Computer Science, University of Milan. He is with the SONICOM Project, funded through the EU's Horizon 2020 Programme, which leverages artificial intelligence to design immersive audio technologies. His research interests include machine learning, signal processing, extended reality, and their application to spatial audio, with a focus on head-related transfer functions, binaural rendering, artificial reverberation, and their influence on auditory spatial perception.

MICHELE GERONAZZO (Senior Member, IEEE) received the M.Sc. degree in 2009, and the Ph.D. degree in 2014. He is currently an Associate Professor with the University of Padova, Padua, Italy, and part of the coordination unit of the EU-H2020 Project SONICOM with Imperial College London, London, U.K. From 2014 to 2021, he was an Assistant Professor with the University of Udine, Udine, Italy and Postdoctoral Researcher with Aalborg University Copenhagen, Copenhagen, Denmark. He has authored or coauthored more than ninety scientific publications. His research mainly focuses on modeling and simulation of complex human-machine sonic interactions. Since 2015, he has been a part of the organizing committee of the IEEE VR Workshop on Sonic Interactions for Virtual Environments and Chair of the 2018 and 2020 editions. He is also an Associate Editor for the IEEE Open Journal of Signal Processing, *ACM Transactions on Applied Perception,* and Editor of the *Sonic Interactions in Virtual Environments* (Springer-Nature). He is also the co-recipient of 6 best paper or poster awards.

FEDERICO AVANZINI is currently a Full Professor with the Computer Science Department, University of Milan, Milan, Italy. He was a Principal Investigator and scientific responsible of EU, national, and industry funded projects. He is also a Conference Coordinator with the permanent Steering Committee of the Sound and Music Computing Conference and Summer School, and President of the Italian Music Informatics Association. His research interests mainly include concern algorithms for sound synthesis and processing, and 3D sound rendering, with applications to such domains as assistive technologies, virtual musical instruments, digital cultural heritage, and digital learning.

STAVROS NTALAMPIRAS received the engineering and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Patras, Patras, Greece, in 2006 and 2010, respectively. He is currently an Associate Professor with the Department of Computer Science, University of Milan, Milan, Italy. He has carried out research or didactic activities with Politecnico di Milano, Milan, Joint Research Center of the European Commission, Brussels, Belgium, National Research Council of Italy, Rome, Italy, and Bocconi University, Milan. His research interests include content-based signal processing, machine learning, audio pattern recognition, bioacoustics, medical acoustics, and cyber-physical systems. He is currently an Associate Editor for IEEE Transactions on Neural Networks and Learning Systems, *PLOS One*, *IET Signal Processing,* and *CAAI Transactions on Intelligence Technology*. He is also a Member of the IEEE Computational Intelligent Society Task Force on Computational Audio Processing.