Do We Need Individual Head-Related Transfer Functions for Vertical Localization? The Case Study of a Spectral Notch Distance Metric

Michele Geronazzo, *Member, IEEE*, Simone Spagnol, and Federico Avanzini

***Abstract*—This paper deals with the issue of individualizing the head-related transfer function (HRTF) rendering process for auditory elevation perception. Is it possible to find a nonindividual, personalized HRTF set that allows a listener to have an equally accurate localization performance than with his/her individual HRTFs? We propose a psychoacoustically motivated, anthropometry based mismatch function between HRTF pairs that exploits the close relation between the listener's pinna geometry and localization cues. This is evaluated using an auditory model that computes a mapping between HRTF spectra and perceived spatial locations. Results on a large number of subjects in the center for image processing and integrated computing (CIPIC) and acoustics research institute (ARI) HRTF databases suggest that there exists a nonindividual HRTF set, which allows a listener to have an equally accurate vertical localization than with individual HRTFs. Furthermore, we find the optimal parameterization of the proposed mismatch function, i.e., the one that best reflects the information given by the auditory model. Our findings show that the selection procedure yields statistically significant improvements with respect to dummy-head HRTFs or random HRTF selection, with potentially high impact from an applicative point of view.**

***Index Terms*—Spatial audio, head-related transfer functions (HRTFs), auditory models, individualized HRTFs, HRTF selection, vertical localization, spectral notch metric.**

I. Introduction

ONE of the main limitations of binaural audio through headphones, that limits its integration into commercial applications of virtual and augmented reality, lies in the lack of individualization of the rendering process. Rendering spatial sound through headphones usually involves the use of *binaural room impulse responses* (BRIRs), which are the combination of two components: the *room impulse response* (RIR), and the *head-related impulse response* (HRIR), which accounts for the acoustic transformations produced by the listener's head, pinna, torso and shoulders [1]. Having a set of HRIRs (or Head-Related Transfer Functions – HRTFs, their Fourier transforms), measured over a discrete set of spatial locations for a specific listener, allows to spatially render a dry sound by convolving it with the desired HRIR pair.

Recording *individual* HRTFs of a single listener implies a trade-off between resources and time, that takes into account several issues such as static/dynamic measurements [2], number of loudspeakers, availability of (semi-) anechoic spaces, robust placement of binaural microphones, monitoring of subject movements during acquisition [3], and repeatability in measurement [4], to name but a few. This makes HRTF recordings impractical for a real-world application, therefore different and more convenient ways to provide a listener with a set of HRTFs are highly desirable. A common practice amounts to using *generic* HRTFs, such as those that can be recorded using a dummy head (e.g., the Knowles Electronic Manikin for Acoustic Research – KEMAR [5]): in this case, the same set is used for any possible listener.

However, generic HRTFs generally result in a degradation of sound perception and localization, and in an overall poor listening experience. For this reason, recent literature is increasingly investigating the use of *personalized* HRTFs, i.e. approaches that allow to provide a listener with a HRTF set that matches as closely as possible the perceptual characteristics of his/her own individual HRTFs.

Personalized HRTFs can be derived from computational models, which generate synthetic responses from a physical [6] or structural interpretation of the acoustic contribution of head, pinna, shoulders and torso [7]. In alternative to computational models, personalization can be also achieved through HRTF selection. In this case, personalized HRTFs are chosen among the HRTF sets available in a database, by finding the "best" match between the listener and one of the subjects in the database. In this work, we follow this approach, and we define a psychoacoustically motivated, anthropometry based distance criterion to find, for a given subject, the best matching HRTF among those available. The criterion is based on a mapping between pinna geometry and localization cues, especially those for localization in the vertical dimension. The envisaged

Manuscript received August 14, 2017; revised December 20, 2017 and February 19, 2018; accepted March 27, 2018. Date of publication April 2, 2018; date of current version April 24, 2018. This work was supported in part by the research project Personal Auditory Displays for Virtual Acoustics, University of Padova, under Grant CPDA135702, and in part by the research project "Acoustically-trained 3-D audio models for virtual reality applications," Aalborg University's 2016–2021 strategic internationalization program "Knowledge for the World." The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Huseyin Hacihabiboglu. *(Corresponding author: Michele Geronazzo.)*

M. Geronazzo is with the Department of Architecture, Design, and Media Technology, Aalborg University, Kbenhavn 2450, Denmark (e-mail: mge@create.aau.dk).

S. Spagnol is with the Department of Information Engineering, University of Padova, Padova 35131, Italy (e-mail: spagnols@dei.unipd.it).

F. Avanzini is with the Department of Computer Science, University of Milano, Milano 20135, Italy (e-mail: federico.avanzini@di.unimi.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2018.2821846

application of this approach is a low-cost and non-invasive procedure for HRTF personalization, where a user can provide individual anthropometric data in the form of pictures [8], [9] or acoustic data [10], receiving back the best-matching HRTF set automatically selected from a database.

The similarity between selected HRTFs and the individual subject's HRTFs must be validated on the basis of psychoacoustic criteria, in order to ensure that the personalization procedure preserves as closely as possible perceptual features and localization abilities. Such validation is typically performed with human subjects through time-consuming psychophysical tests that require a meaningful statistical population, with issues related to number of participants, quality of the responses, homogeneity of the subject pool, and so on. The majority of these studies rely on a tradeoff between the above-mentioned factors in order to limit participant effort and learning, while keeping the experimental plan practical. Typically, available experimental data are collected on a limited pool of participants (e.g. 10–20 listeners) and/or with limited accuracy in localization performances [11] and/or using limited sets (≤ 3) of non-individual HRTFs [12]. Accordingly, the validity and statistical power of such results remain highly questionable without an extensive study on a wider population and a larger set of HRTFs.

To our knowledge, no published study faces the issue of overcoming practical limitations of massive participation of human subjects to HRTF measurements and listening tests. In this direction, we follow an alternative approach in which pairs of HRTF sets are evaluated by means of computational auditory models that simulate the human auditory system with respect to localization abilities [13], [14]. Previous studies [14] have already shown that these models are able to reproduce experimental data accurately. Therefore in this work we use one such model to define a perceptual metric that allows to automatically assess the performance of the HRTF selection procedure. This is an attractive approach in that it allows to use quantitative predictions (from auditory models) rather than qualitative responses from listening tests, overcoming the practical limitations of the latter. In sake of comparability, our approach is also evaluated keeping a connection with real-world data from listening tests, leading to individually calibrated simulations according to prior subjective evaluations [15].

The paper is organized as follows. Section II discusses relevant literature about individual sound localization, HRTF features and personalization, with a focus on vertical localization and HRTF selection approaches. Section III presents a procedure based on an auditory model for vertical localization that, given an individual HRTF, rates its perceptual similarity to all the HRTFs in a database. Section IV introduces the pinna reflection model and presents a notch-frequency mismatch function that can be used for HRTF selection. This is further analyzed in Section V, by correlating its selection results to those previously obtained using the auditory model. A general discussion of the results is provided in Section VI.

II. HRTF Personalization and Selection

A. Listening With Nonindividual HRTFs

As mentioned in the introduction, the most common approach for binaural sound rendering amounts to employing generic HRTFs measured on dummy heads. Dummy heads allow robust and easy-to-manage measurement sessions and often provide a reasonable trade-off between representativeness of a wide human population and average efficacy (see the historical review by Paul [16]), although outcomes vary considerably [4].

However, it is indisputable that our representation of auditory information is based on everyday life listening with our individual HRTFs, electing them naturally as the ground truth condition. Listening through generic dummy ears causes noticeable distortion in the listening experience, including increased front-back confusion, lack of externalization, and localization errors [17], [18]. Romigh and Simpson [19] recently identified the intraconic spectral component as the main cause of inter-individual differences where pinna acoustics become dominant.[1] Dummy-head HRTFs can thus be considered as average HRTF templates, exhibiting high variability in localization performance.

Localization in the vertical dimension is especially affected by the individual shape of the pinnae [21]. Scattering of acoustic waves in the proximity of the pinna creates listener-dependent peaks and notches that characterize the high-frequency spectrum (above 4-5 kHz) [22]. These depend strongly on the elevation angle of the incoming sound [23], [24], and to a remarkably minor extent on azimuth [6] and distance [25], [26]. The relative importance of these peaks and notches has been disputed over the past years. A recent study [27] showed that a parametric HRTF recomposed using only the first, omnidirectional peak in the HRTF spectrum (corresponding to Shaw's mode 1 [23]) coupled with the first two notches yields almost the same localization accuracy as the corresponding measured HRTF. Additional evidence in support of the relevance of the lowest-frequency notch is provided by Moore [28], who states that the threshold for perceiving a shift in the central frequency of a spectral notch is consistent with the localization blur on the median plane.

Several studies attempted to manipulate a HRTF template in order to shape individual contributions. Middlebrooks defined a *frequency scale factor* with the aim of reducing inter-subject spectral differences and computing scaled HRTFs for an arbitrary listener from anthropometric differences [29]. Brungart and Romigh proposed a HRTF enhancement procedure that increases the salience of direction-dependent spectral cues for elevation in both generic and individual HRTFs [30] through the use of an *enhancement factor* applied to the magnitude of the intraconic spectral component. HRTFs from a dummy head template can be also considered for ITD individualization with scaling procedures [31] or with linear regression analysis on anthropometry [32]. Tan and Gan allowed listeners to manipulate HRTF spectra with a series of bandpass filters to boost or attenuate 5 frequency bands [33]; similarly, a systematic manipulation of the directional bands was tested by So *et al.* [34]. Moreover, listeners can increase their localization performances through adaptation procedures [35], [36] which provide tools for remapping localization cues to spatial direction.

On the other hand, some studies have shown that listeners do not necessarily exhibit significant degradation in localization performance when using non-individual HRTFs; see for exam-

[1] The intraconic spectral component is derived by subtracting the average of equilateral directional transfer functions (DTFs) [20] from each DTF.

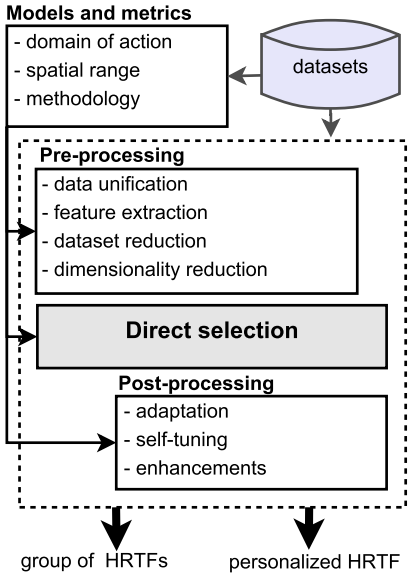


Fig. 1. General work-flow for HRTF selection.

ple the work of Iwaya [37] where non-individual HRTFs did not statistically differ from individual ones in localization performance, and were statistically better in reducing front-back confusion. Moreover, HRTFs resulting from a subjective tuning of PCA weights [38] led to better discrimination in elevation with respect to individual HRTFs.

B. Selection Approaches

The above discussion provides motivations for the development of HRTF *personalization* approaches. One of the most popular approaches in recent literature, including this work, is HRTF *selection*, in which personalized HRTFs are chosen by selecting the best match among the HRTF sets available in a database. In this section, we provide a formalization of the main elements involved in the problem of HRTF personalization, with particular focus on the selection of existing HRTF data. The scheme in Fig. 1 depicts a general work flow.

Finding models and metrics for the problem requires the definition of (i) domains of action, (ii) spatial ranges of interest, and (iii) a methodology for the techniques and processing stages within the identified domains and ranges. Domains of action belong to the following three areas: acoustics, anthropometry, and psychoacoustics. Accordingly, datasets must also include– beside HRTFs – anthropometric information and subjective ratings. The latter group includes results from localization tests (e.g. azimuth and elevation judgments, externalization ratings, front-back differentiation, etc.) with non-individual HRTFs, as well as listener preferences [11], [37]. The anthropometry of the listener can be collected in the form of direct measurements, pictures, and 3D scans [8], [39], [40]. Investigation can be limited to a specific spatial range of interest, i.e. a subspace around the listener.

Proposed methodologies are typically composed of a pre-processing phase, a direct selection, and a post-processing phase. Pre-processing include data unification, to manage heterogeneity in different HRTF databases, techniques to reduce the size of the dataset, and dimensionality reduction of the HRTFs. The direct selection phase includes several techniques, depending on the chosen domain of action: anthropometric database matching [41], linear regression models between acoustical and physical features [42], subjective selection on a reduced set of HRTFs [11], [37], [43], and optimization of errors and distances in the acoustic domain [42]. Finally, the post-selection phase usually requires the listener's involvement for self-tuning actions, such as subjective spectral manipulations and enhancement [33], choice of a preferred scale factor [42], and adjustment of weights [38], towards a training for adaptation to non-individual HRTFs [35].

III. Do We Need Individual HRTFs for Vertical Localization?

We can identify a crucial research question: is it possible to find personalization procedures for non-individual HRTF sets allowing a listener to have an equally accurate vertical localization than with his/her individual HRTFs?

In this section, we present a procedure that provides a quantitative measure of the perceptual similarity of a pair of HRTF sets with regard to vertical localization. As such, it can be used to rank all the HRTFs in a database in order of similarity with that of a specific listener. The one ranking first can be regarded as the best non-individual HRTF for the listener.

The proposed procedure exploits an auditory model for vertical localization and makes use of psychoacoustically motivated performance metrics. Therefore it may be regarded as a virtual experiment in which a virtual listener performs a vertical localization test using different HRTFs, which are ranked in order of performance. The procedure was applied to HRTF datasets from the Center for Image Processing and Integrated Computing (CIPIC) Interface Laboratory of the University of California - Davis [44], Acoustics Research Institute (ARI) of the Austrian Academy of Sciences [45], and a subset of ARI subjects. Motivations behind this choice can be summarized as follow; the CIPIC is in line with our earlier studies and findings [8], [39], [46]; on the other hand, the ARI is a larger HRTF database which supports the evaluation of the auditory model adopted in this study [14]; for 17 ARI subjects, Majdak *et al.* [15] provided individual calibration parameters thus allowing the identification of a subgroup that can represent a faithful reproduction of real subjects in both acoustic and non-acoustic factors for localization.

The results serve as a benchmark for the definition of the anthropometry-based HRTF mismatch function discussed in Sections IV and V.

A. Auditory Models for Vertical Localization

Several computational auditory models predict accuracy of human localization (and particularly vertical localization) from acoustical spectral cues. Two main categories of models exist: (i) statistical and machine learning approaches and (ii) functional models. The first group collects several studies conducted also in the field of robot audition [47], [48], and use such approaches as gaussian process regression models [49] and time-

delay neural networks [50]. The latter group adopts physiologically and psychophysically-inspired parameters following a *"template-based"* paradigm [13]: a comparison between the internal representation of an incoming sound at the eardrum and a reference template (usually an individual representation of listener HRTFs) is performed in terms of comparisons through first/second order derivatives [51] or spectral correlation [52].

In this paper, we adopt the Langendijk model [13], extended by Baumgartner *et al.* [14], where spectral features of sound events filtered with different HRTFs (target) correlate with the direction of arrival of the HRTF template, leading to a spectro-to-spatial mapping. This approach is further supported by a recent study of Van Opstal *et al.* [53] where the authors estimated the listeners' spectral-shape cues for elevation perception from the distribution of localization responses.

The model is based on two different processing phases prior to the prediction of absolute elevation. During peripheral processing, an internal representation of the incoming sound is created. The *target* sound is converted into a DTF and filtered with a gammatone filterbank simulating the auditory processing of the inner ear. In the second phase, for each target/template angle and frequency band based on equivalent rectangular bandwidth (ERB), the algorithm computes the gain at the central frequency of each band and the target/template internal representations. The *inter-spectral difference* (ISD) for each band is extracted from the differences in dB between each target angle and all template angles; for each target angle, the *spectral standard deviation* (SSD) of the ISD is computed across all template angles. The probability that a virtual listener points to a specific response angle defines the *similarity index* (SI) which receives as input the template-dependent SSD for the argument of a Gaussian distribution with zero mean and standard deviation called *uncertainty*, U. The lower the U, the higher the sensitivity of the listener in discriminating different spectral profiles resulting in a measure of probability. Simulation data are stored in probability mass vectors, where each response angle has the probability that the virtual listener points at it.

B. Model Tuning and Performance Metrics

In this work, simulations were run on the median plane only, where acoustic properties of the external ear provide vertical localization cues [24] with minimum interference from other localization cues; simulations accounted for three datasets:

- **CIPIC** - 45 virtual subjects for whom individual HRTF measurements are available in the CIPIC database [44]: 2500 HRIRs each, given by the combination of 25 azimuths $\times$ 50 elevations $\times$ 2 ears, measured at sampling rate $f_s = 44.1$ kHz (200 samples). Elevation ϕ is uniformly sampled on the range -45° to $+230.625^\circ$ in 5.625° steps. For each virtual subject, we set an uncertainty value $U = 2$, which reasonably approximates the uncertainty of a real listener in localization tasks [15].
- **ARI** - 97 virtual subjects for whom individual HRTF measurements are available in the ARI database: 1550 HRIRs each, given by the combination of 90 azimuths $\times$ 22 elevations $\times$ 2 ears, measured at sampling rate $f_s = 48$ kHz (256 samples). [2] Elevation ϕ is uniformly sampled on the range -30° to $+80^\circ$ in 5° steps. For each virtual subject, we set an uncertainty value $U = 2$.
- **ARIrU** - 17 virtual subjects for whom individual HRTF measurements from the ARI database and individual uncertainty values from [15] are available.

It is worthwhile to notice that these three datasets were considered separately; their combination could lead to biased results and misinterpretations due to the heterogeneity between HRTF databases that is a well-known issue in the literature [4], [54]. Normalization and data correction of such differences require an ad-hoc merging procedure which is beyond the scope of the paper.

DTF data were extrapolated from free-field compensated HRIRs and subsequently pre-processed by windowing with a 1-ms Hanning window centered on the maximum temporal peak in order to remove the acoustic contribution of torso reflections [46]. On the other hand, low frequency torso cues were still included; however, individual torso differences were minimized once considering simulation data related to target elevation angles in the frontal range $[-45^\circ, +45^\circ]$ where high frequency spectral peaks and notches are more prominent, reflecting the higher inter-subject variability of the pinna [23], [55]. It has to be noted that for sound source directions above the listeners (elevation $> 45^\circ$), spectral details are poorly marked due to a dominance of concha resonance [22] and HRTFs can be considered indistinguishable with a JND of 24°, thus reflecting a very poor localization performance also in real listening conditions [56].

All median plane template angles were considered in the computation of the following psychoacoustic performance metrics for vertical localization, accordingly with Middlebrooks [29] and Baumgartner *et al.* [57]:

- *local polar RMS error, PE*: quantifies the average "local" localization error, i.e. when the absolute polar error is below $\pm 90^\circ$. More precisely, we first define PE_j, the RMS angular error accounting for the precision of every j-th elevation response close to the target position:

$$PE_j = \sqrt{\frac{\sum_{i \in L} (\phi_i - \phi_j)^2 p_j[\phi_i]}{\sum_{i \in L} p_j[\phi_i]}}, \tag{1}$$

  with

$$L = \{ i \in N : 1 \leq i \leq N_\phi, |\phi_i - \phi_j| \bmod 180^\circ < 90^\circ \},$$

  defining local polar-angle responses within $\pm 90^\circ$ of the local response ϕ_i and the target position ϕ_j; $p_j[\phi_i]$ denotes the probability mass vector. Then, the *PE* for a single auditory model simulation is computed as the average of the PE_j's across target elevations.
- *quadrant error rate, QE*: quantifies the localization confusion related to the rate of "non-local" responses, i.e. where the absolute polar error exceeds $\pm 90^\circ$. We first define QE_j

[2] HRTFs (not *b* version) available at http://sofacoustics.org/data/database/ari/ (last access 18/12/2017)

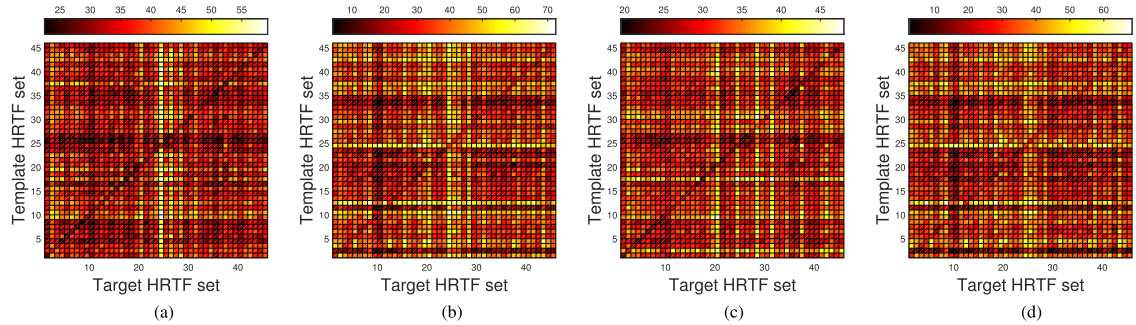


Fig. 2. Localization predictions resulting from auditory model simulations on 45 CIPIC subjects in the median plane. *"All-against-all"* matrices for (a) polar error (PE) [deg], (b) quadrant error rate (QE) [%], (c) global polar error (GPE) [deg], and (d) front-back confusion rate (FB) [%].

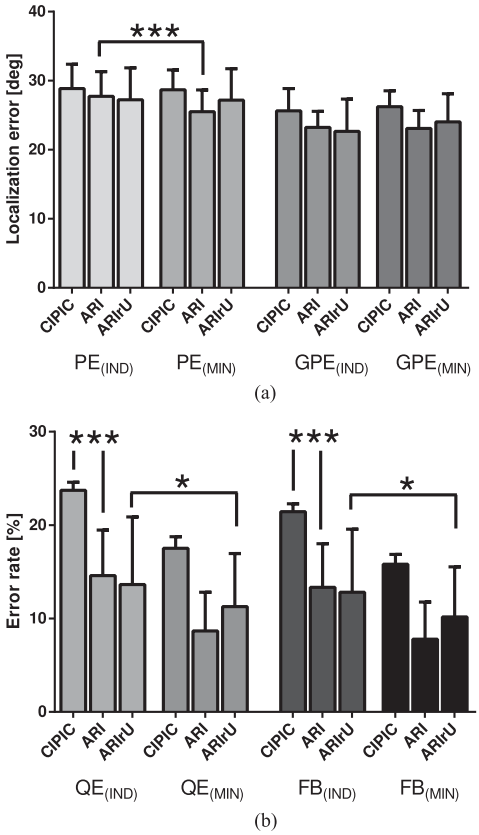


Fig. 3. Listening with individual HRTFs (IND) vs. *"best available"* non-individual HRTFs (MIN errors in metrics): global statistics for (a) angular error $[PE, GPE]$ and (b) rate of confusion $[QE, FB]$, for each the analyzed databases (CIPIC, ARI, and ARIrU). Asterisks and bars indicate, where present, a significant difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ according to paired t-test).

for the *j*-th elevation response:

$$QE_j = \sum_{i \in NL} p_j[\phi_i], \tag{2}$$

with

$$NL = \{i \in N : 1 \leq i \leq N_\phi, |\phi_i - \phi_j| \bmod 180^\circ \geq 90^\circ\}.$$

Then, the *QE* for a single auditory model simulation is computed as the average of the QE_j's across target elevations.

Moreover, we defined two additional metrics which are commonly used in the literature on sound localization [8], [58]:

- *global polar error, GPE*: quantifies the absolute angular localization error, with front-back confusions "resolved". For the j-th elevation response:

$$GPE_j = \frac{\begin{array}{c}\sum_{f \in F} |\phi_f - \phi_j| \left(p_j[\phi_f] + p_{\bar{j}}[\phi_f]\right) \\ + \sum_{b \in B} |\phi_b - \phi_j| \left(p_j[\phi_b] + p_{\bar{j}}[\phi_b]\right)\end{array}}{\begin{array}{c}\sum_{f \in F} p_j[\phi_f] + p_{\bar{j}}[\phi_f] \\ + \sum_{b \in B} p_j[\phi_b] + p_{\bar{j}}[\phi_b]\end{array}}, \tag{3}$$

with

$$F = \{i \in N : 1 \leq i \leq N_\phi, \phi_i \leq 90^\circ\},$$
$$B = \{i \in N : 1 \leq i \leq N_\phi, \phi_i > 90^\circ\},$$

and where the angle $\bar{j}$ is the front-back angle associated to j (i.e., the mirror angle of j with respect to the coronal plane). Then the *GPE* for a given auditory model simulation is computed as the average of the GPE_j's across target elevations.

- *front-back confusion rate, FB*: quantifies the localization confusion by measuring the rate of frontal responses where the target position ϕ_j is on the back region and *vice versa*, excluding elevation angles above the listener. A $\pm 30^\circ$ area is considered in this definition. For the j-th elevation response:

$$FB_j = \sum_{i \in C} p_j[\phi_i], \tag{4}$$

with

$$C = i \in N : 1 \leq i \leq N_\phi, \phi_i > 120^\circ \text{ if } \phi_j \leq 60^\circ \wedge \phi_i \leq 60^\circ \text{ if } \phi_j > 120^\circ,$$

Then the *FB* for a single auditory model simulation is computed as the average of the FB_j's across target elevations.

C. Results

Using all the available subjects for each database, a total of more than 10^4 simulations ($45 \times 45 = 2025$, $97 \times 97 = 9409$, and $17 \times 17 = 289$) were run following an *all-against-all* principle: for each virtual subject, his/her individual HRTF set were used as the template of the auditory model, and predictions on

TABLE I
STATISTICAL ANALYSIS ON METRICS FOR LOCALIZATION PERFORMANCE BETWEEN INDIVIDUAL (IND) AND *"best available"* (MIN) LISTENING CONDITIONS

	CIPIC	ARI	ARIrU
r [PE, GPE] IND	0.72 ***	0.80 ***	0.94 ***
MIN	0.76 ***	0.83 ***	0.95 ***
r [QE, FB] IND	0.92 ***	0.93 ***	0.98 ***
MIN	0.98 ***	0.93 ***	0.97 ***
$PE_{\text{IND}} - PE_{\text{MIN}}$	t(44)=0.35	t(97)=6.97	t(16)=0.06
	p = 0.724	***	p = 0.95
$GPE_{\text{IND}} - GPE_{\text{MIN}}$	W=-339	W=943	W=-73
	p = 0.056	p = 0.090	p = 0.089
$QE_{\text{IND}} - QE_{\text{MIN}}$	t(44)=6.77	t(97)=13.22	t(16)=2.03
	***	***	*
FB_{IND}-FB_{MIN}	W=873	W=4507	W=99
	***	***	*

Asterisks indicate, where present, a significant difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ according to paired t-test or Wilcoxon matched-pairs signed rank test).

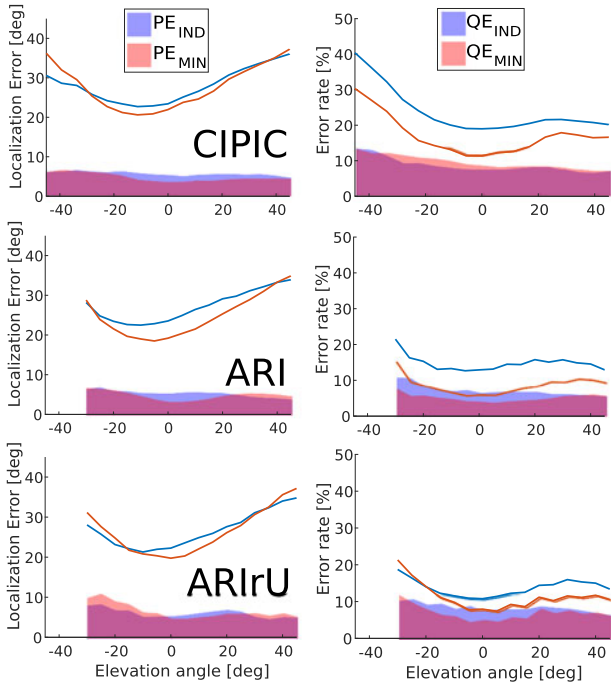


Fig. 4. Average elevation-dependent PE and QE for individual (blue lines) and *"best available"* non-individual (red lines) listening conditions in the CIPIC (first row), ARI (second row) and ARIrU (last row) datasets. Standard deviations for each angle can be determined from colored areas on the bottom of each subplot.

vertical localization performances were computed considering all the available HRTF sets as targets. Fig. 2 summarizes simulation results for the CIPIC database into four matrices for each of the above-defined metrics, where row-wise reading allows visual inspection of prediction results of all HRTF sets for a given listener, i.e. 44 non-individual HRTF sets plus the individual one (the element in the diagonal). It is worthwhile to notice that errors appeared also in the diagonal due to the U parameter able to model localization flaws that are typical of real static listening conditions.

We subsequently analyzed all the simulations by comparing predictions with two listening methods: individual HRTFs (IND) and the *"best available"* non-individual HRTFs (MIN). For every i-th row of the simulation matrix, i.e. for the i-th subject, the best available non-individual HRTF was chosen to be the one providing the minimum error value (excluding the diagonal).

Preliminary analysis of gaussianity was performed on the data by means of a Shapiro-Wilk test, which revealed violations in the distributions of GPE_{IND} and FB_{IND}. Accordingly, a paired t-test was computed for PE and QE and a Wilcoxon matched-pairs signed rank test for GPE and FB in order to assess statistical differences between individual and *"best available"* non-individual HRTFs. Results of this analysis are reported in Fig. 3 and in Table I, which groups the proposed metrics in (a) angular error [PE, GPE] and (b) rate of confusion [QE, FB] for each HRTF database. This grouping was guided by computing nonparametric Spearman correlation between data distributions, which yielded high statistical significant correlation coefficients for [PE, GPE] and for [QE, FB] in each database.

The non-significant difference in angular error between individual and best available (Fig. 3(a) and 3rd–4th rows of Table I) supports the idea that there exists a non-individual HRTF set allowing vertical localization as accurate as with individual HRTFs. Whereas the statistically significance in PE for ARI exaggerates this trend, further investigations can be found in Section IV-B. Interestingly, significant statistical differences in rate of confusion (Fig. 3(b) and 5th–6th rows of Table I) suggest that it is possible to find a non-individual HRTF set which allows better front-back discrimination than the individual one. Even if this latter result seems counter-intuitive, Asano *et al.* [24] already reported this phenomenon with two expert listeners (e.g., see performances of subject 2 in their experiment 1), concluding that macroscopic patterns, directional bands [1] and level of expertise [59] in the high-frequency regions are necessary for front-back discrimination in static conditions.

In the remainder of the paper we focus on PE and QE for the sake of consistency with previous literature which uses the same auditory model [14]. These two metrics are highly correlated to GPE and FB, respectively, suggesting interchangeability among them. Moreover, PE and QE exhibit Gaussian distributions when applied to the CIPIC database, making them more convenient metrics for statistical analysis.

Even though results of global localization metrics were similar between IND and MIN, elevation angle dependency has to be investigated in order to guarantee the actual matching between individual and *"best available"* non-individual HRTFs for each target angle. Fig. 4 depicts PE and QE as a function of the elevation angle for CIPIC, ARI, and ARIrU datasets, respectively. From this analysis, we can notice a close trend for both conditions in all datasets, thus not interfering with results obtained from global metrics. Moreover, the absence of $\phi \in [-45°, -30°)$ in the ARI database explains the average lower values for ARI and ARIrU compared to CIPIC for all the metrics; low error rates were particularly marked for QE and FB [see Fig. 3(b)] due to the absence of data for computing error rates in extreme lower elevation angles.

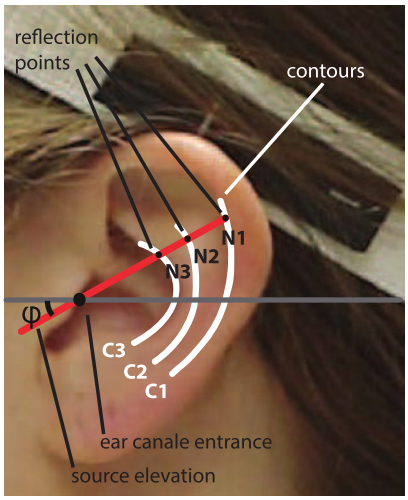


Fig. 5. Side-face picture of CIPIC subject 048. Following the pinna reflection model, a schematic representation provides the identification of ear canal entrance, main contours, and reflection points for a given elevation ϕ.

IV. An Anthropometry-Based HRTF Mismatch Function

Having established in the previous section a robust metric for HRTF similarity in the domain of psychoacoustics (predictive model), in this section we propose a second metric in the domain of acoustics, and we show that it can be derived from individual anthropometric features of the pinna. In Section V, this new metric will be validated and tuned using the first one as benchmark.

Being based on anthropometry, the new metric has the advantage that it can be used for direct HRTF selection once anthropometric data are available (e.g., from 2D pictures of the pinna). With reference to the general work-flow discussed in Section II-B, the spatial range of interest of this selection procedure is the vertical plane.

A. Notch Frequency Mismatch

According to a revised pinna reflection model [39], frequencies of the three main pinna notches (i.e. corresponding to reflection points $N_{1\ldots3}$ in Fig. 5) in a median-plane HRTF can be estimated with reasonable accuracy by calculating the distance between the ear canal entrance and points on the three pinna contours thought to be responsible for pinna reflections, i.e. the helix border (C_1), the antihelix and concha inner wall (C_2), and the concha outer border (C_3) in the schematic view of Fig. 5. More in particular, given the i-th contour C_i and an elevation ϕ, and assuming each reflection to be negative and responsible for a single notch, we estimate the frequency value where destructive interference between the direct sound and the sound reflected by the pinna contour occurs as

$$f_i(\phi) = \frac{c}{2d_i(\phi)}, \quad i = 1, 2, 3\ , \tag{5}$$

where c is the speed of the sound and $d_i(\phi)$ the distance between the pinna reflection point and the reference point. Therefore, the notch frequencies $f_i(\phi)$ can be computed from the distances $d_i(\phi)$, which in turn can be estimated from a 2D image of the pinna.

In our previous study, the notch frequencies estimated with (5) were found to accurately approximate those actually appearing in the corresponding measured HRTFs for several subjects [39]. Thus, given a subject whose personal HRTFs are not available, it is possible to select from a database the HRTF set that has the minimum mismatch between the f_i frequencies extracted from his/her own pinna contours and the F_i notch frequencies of the available median-plane HRTFs. Similarly, given two HRTF sets S_j (*template* set) and S_k (*target* set), the notch frequency mismatch between them can be defined as

$$m_{j,k} = \frac{1}{3}\sum_{i=1}^{3}\frac{w_i}{|\phi|}\sum_{\phi}\frac{|F_i(S_j,\phi) - F_i(S_k,\phi)|}{F_i(S_j,\phi)}, \tag{6}$$

where w_i are a convex combination of weights ($\sum_{i=1}^{3} w_i = 1$) and ϕ spans all the elevation angles for which the i-th notch is present in the corresponding HRTF. From this definition, it can be noticed that the mismatch function is actually non-commutative, just like the previously defined auditory model metrics.

In a preliminary study [8], we found that using the mismatch function of (6) for HRTF selection increased the average elevation performances of 17% compared to the use of a generic HRTF with average anthropometric data, significantly enhancing both the externalization and the up/down confusion rates. Furthermore, the convex combination assigning the whole weight to the first notch ($w_1 = 1$, $w_2 = w_3 = 0$, termed "all-first" weight combination hereafter) gave better average results but not statistically significant differences compared to the convex combination assigning equal weights to the three notches ($w_1 = w_2 = w_3 = 1/3$, termed "equalized" weight combination hereafter). This suggests that notches may have different relevance in elevation perception and therefore shall have different weights in the mismatch function. However, a systematic evaluation of individual weight combination for each subject was not practical in our former study which followed a typical research methodology with listening tests. Accordingly, finding the optimal convex combination of weights allows to study the connection between the characterization of individual notch patterns and localization performances. In the remainder of this paper, this main research question is addressed.

B. Extraction of Spectral Notches

In a previous study [46], we developed an algorithm that allows simultaneous estimation of notch frequencies, depths, and bandwidths from median plane HRTFs. However, since in this work we are only interested in feature extraction of notch frequencies, we apply the ad-hoc signal processing algorithm by Raykar *et al.* [60]. Briefly, the algorithm computes the autocorrelation function of the linear prediction residual and extracts notch frequencies as the local minima of its group-delay function falling beyond a fixed threshold.

Then, for each available elevation ϕ, the extracted notches are grouped in frequency tracks along adjacent elevations through the McAulay-Quatieri partial tracking algorithm [61], which is

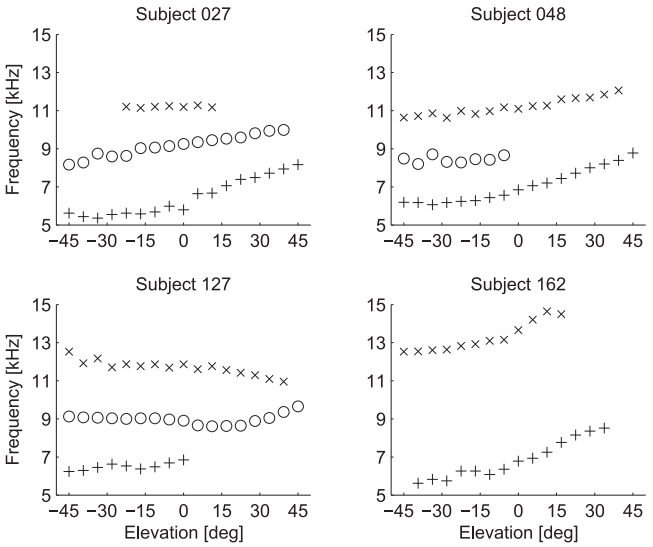


Fig. 6. Extracted notch frequency tracks of four representative CIPIC subjects (F_1:+, F_2:○, F_3:×).

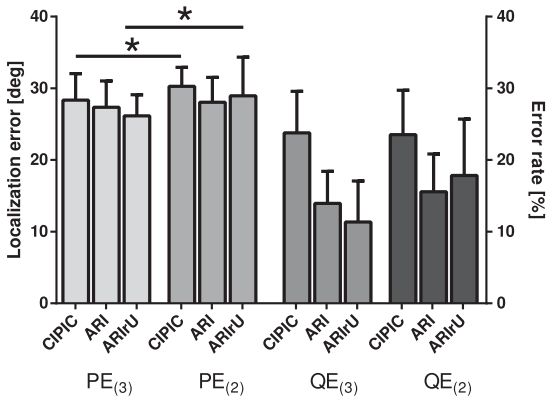


Fig. 7. Individual predicted localization performances in local polar RMS error (PE) and quadrant error rate (QE) for subjects with 3 extracted notch tracks ($PE_{(3)}, QE_{(3)}$) and subjects with 2 extracted notch tracks ($PE_{(2)}, QE_{(2)}$), for each of the three analysed databased (CIPIC, ARI, and ARIrU). Asterisks and bars indicate, where present, a significant difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ according to Mann-Whitney U test).

adapted in order to perform the tracking in space rather than in time, as in the original formulation. The very same procedure for notch extraction and grouping has been used in previous works [26], [62].

Only tracks with at least 3 points are preserved. If more than three tracks satisfying such a requirement are available, only the three longest tracks are considered and labeled F_1, F_2, and F_3 in increasing order of average frequency. Fig. 6 reports tracks of four representative subjects.

We found a significant number of cases where a subject exhibits only two tracks (12 for CIPIC, 47 for ARI, and 8 for ARIrU). This finding is due either to the actual absence of a third notch in the related HRTFs or to an insufficient notch depth, or to occasional tracking errors. To assess this, we compared individual elevation performances as predicted by the auditory model [predictions on the diagonal in Fig. 2(a) and (b)], between the group of subjects with 2 tracks ($PE_{(2)}$, $QE_{(2)}$) and all other subjects ($PE_{(3)}$, $QE_{(3)}$). Preliminary analysis of gaussianity was performed for each group by means of a Shapiro-Wilk test, which revealed violations in the distributions of $PE_{(3)}$ and $QE_{(3)}$. Accordingly, a Mann-Whitney U test was computed for both PE and QE metrics in order to assess statistical differences between subjects with 2 and 3 tracks. Results of this analysis are reported in Fig. 7. Significant statistical effects were found for $PE_{(3)} - PE_{(2)}$ ($U = 111, p = 0.026$, CIPIC; $U = 7$, $p = 0.026$, ARIrU) with the sole exception of ARI ($U = 838$, $p = 0.296$). No differences were found for $QE_{(3)} - QE_{(2)}$ in all datasets ($U = 187$, $p = 0.787$, CIPIC; $U = 761$, $p = 0.090$, ARI; $U = 12$, $p = 0.127$, ARIrU).

This result suggests that the lack of a third notch in this group of subjects is not due to tracking errors but rather reflects an actual degradation of vertical localization performances. In order to support this hypothesis, we investigated real listening performances of the 17 subjects of [15] in order to identify the connection between HRTF spectral features, i.e. notches in our study, and uncertainty values. Within the limited number of participants, 87% (7/8) subjects with 2 tracks exhibited uncertainty above the mean ($U > 2$), while 87% (7/8) with 3 tracks were below-equal the mean U ($U \leq 2$). Accordingly, the 3-track group might be the most perceptually stable group due the fine ability of its subjects in localizing; on the other hand, an analysis conducted with the 2-track group might be biased by a high uncertainty value. Practically speaking, subjects with high uncertainty typically might face an easier HRTF selection process due to a leveling of target-template correlation. We can not guarantee that those subjects belong to a homogeneous population together with the 3-track group, which requires a separate ad-hoc analysis on this aspect.

The suggested connection between number of notches and uncertainty value could also be able to explain the missing statistical significance in PE of ARI's 3- vs. 2-track HRTFs (Fig. 7), and the statistical significant improvement of PE_{MIN} compared to PE_{IND}: since the ARI analysis assigned a fixed $U = 2$, performances of the 2-track group ($\approx 50\%$ of ARI) could have been overestimated in precision, leading to a wrong balancing in auditory model predictions. This effect was minimized by

- a small number of 2-track subjects in the CIPIC (12);
- counterbalancing 2-track subjects with high U in the ARIrU.

Finally, since the notch frequency mismatch function cannot be computed for subjects with missing tracks, we safely chose to reduce the dataset for the analysis not considering subjects with 2 tracks only. Among the remaining subjects, we discarded dummy heads (if present) and subjects with less then 2 tracks (9 ARI sets which were probably corrupted by measurements errors), leaving 31 (CIPIC), 41 (ARI) and 8 (ARIrU) subjects and their extracted notch tracks as the dataset for tuning the weights of the mismatch function.[3]

[3]We preferred to analyse acoustic data measured on human subjects only in order to have an homogenous dataset with real skin response and comparable measurement conditions.

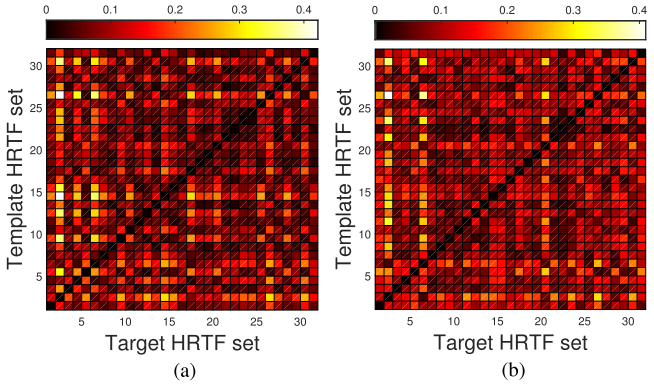


Fig. 8. Notch frequency mismatch calculated for 31 CIPIC subjects in the median plane. All-against-all matrices for (a) "all-first" weight combination ($w_1 = 1, w_2 = w_3 = 0$, and (b) "equalized" weight combination ($w_1 = w_2 = w_3 = 1/3$.

In the remaining of the paper, we chose not to describe all databases in parallel, focusing on our methodology applied to CIPIC database and thus opting for a synthetic and schematic section for a replicated analysis on ARI database (see Section V-C). Finally, we do not consider ARIrU in the following analysis due to the small number of subjects in the 3-track dataset (only 8).

V. Weight Tuning

The mismatch function described in (6) is parametrized by the three weights w_i $(i = 1, 2, 3)$. In order to use it in practice, it is necessary to estimate the best convex combination of weights.

The mismatch function was first evaluated on the complete set of spectral notch tracks for every pair of 31 test subjects and for every possible convex combination of weights w_i, in a uniform grid with step $\Delta w = 0.02$. This led to a total of 1326 convex combinations by $31^2 - 31 = 930$ subject pairs (excluding the same-subject combinations, that always give a mismatch equal to zero). Fig. 8 reports as an example the mismatch for all subject pairs in the "all-first" and the "equalized" weight combinations, used in our preliminary study [8]. Note that the matrices are not symmetric, following the non-commutative structure of the mismatch function.

A. Correlation Analysis

In order to select the optimal convex combination of weights which best reflects the information given by the auditory model, we first analyzed which of the available performance metrics (PE, QE), is the most suited for this goal.

To this aim, we initially computed the 2D correlation coefficient between either the PE or the QE auditory model matrix (obtained by subsetting the matrices reported in Fig. 2 for the 31 examined subjects) and each notch frequency mismatch matrix given by a different convex combination of weights. Formally, if A is the auditory model matrix and B is the notch frequency

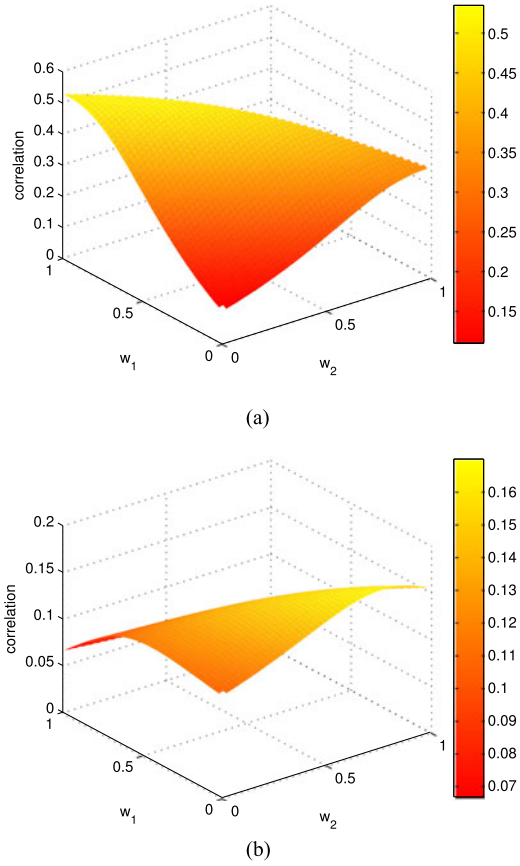


Fig. 9. Correlation between auditory model matrix and notch frequency mismatch matrices as a function of w_1 and w_2: (a) PE, (b) QE.

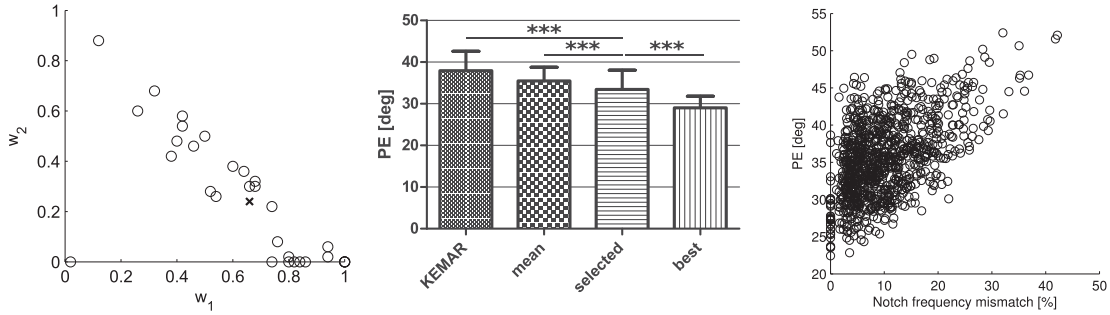
mismatch matrix, the correlation coefficient is defined as

$$r_{A,B} = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{(\sum_m \sum_n (A_{mn} - \bar{A})^2)(\sum_m \sum_n (B_{mn} - \bar{B})^2)}}, \tag{7}$$

where $\bar{A}$ and $\bar{B}$ are the mean values of A and B, respectively.

Fig. 9 shows the calculated 2D correlation coefficient as a function of w_1 and w_2 ($w_3 = 1 - w_1 - w_2$), for each of the two auditory model metrics. Note that the QE metric does not show a significant correlation with the notch frequency mismatch metric for any of the considered weight combinations ($r < .17$, $p > .36$). However, the PE metric correlates significantly ($p < .01$) with the mismatch metric when the all-first weight combination is used ($r > .5$).

By virtue of this preliminary result, we chose PE as the performance metric for subsequent analysis of the individual correlation between the PE and mismatch function for each subject. Correlation coefficients were calculated separately for every template HRTF set (matrix rows), and the convex combination of weights with maximum correlation was individually computed. We refer to this combination as *best subjective combination*. The respective correlation coefficients mostly range between .4 and .8 ($p < .05$), with five non-statistically significant cases where $r < .4$, of which only one case exhibited a negative mild correlation coefficient ($r = -.23$).



(a) (b) (c)

Fig. 10. (a) Best subjective weight combinations (○) and their center of mass (×). The all-first weight combination results as best for four different subjects. (b) Mean and standard deviation of predicted PEs with KEMAR HRTFs (*KEMAR*), mean of all predicted PEs (*mean*), predicted PEs with the non-individual HRTF set selected according to notch frequency mismatch of the first track (*selected*), and best HRTF set according to the auditory model (*best*). Asterisks and bars indicate, where present, a significant difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ according to paired t-test). (c) Notch frequency mismatch with all-first combination against PE, for all subject pairs.

Fig. 10(a) shows all the best subjective combinations along with their center of mass, calculated after individually weighting every best subjective combination by its corresponding correlation coefficient r. The center of mass corresponds to the following convex combination:

$$w_1 = 0.66, \;\; w_2 = 0.24, \;\; w_3 = 0.1, \tag{8}$$

termed "centroid" weight combination hereafter.

Then, for each subject we compared predictions in elevation localization performances (PE) between HRTF sets selected according to the best subjective weight combination, and HRTF sets selected according to the centroid combination of (8). Gaussianity was verified through a Shapiro-Wilk test. Accordingly, a paired t-test was performed, revealing no statistically significant differences on PE ($t(30) = 0.82, p = 0.416$). This finding allows to consider the centroid combination as the globally optimal combination, confirming the salience of the first notch ($w_1 = 0.66$) for vertical localization accuracy and the relatively small relevance of the third notch ($w_3 = 0.1$).

It has to be highlighted that the above results are robust with respect to the uncertainty value fixed for the auditory model simulations ($U = 2$). The very same analysis was performed with $U = .5$, yielding a globally optimal combination equal to $w_1 = .68, w_2 = .22, w_3 = .1$ almost identical to the case $U = 2$ reported in (8).

B. Comparisons Between Selection Methods

The centroid weight combination derived in the previous section can be directly used to select the HRTF set (other than the individual one) that minimizes the notch frequency mismatch function based on the extracted notch frequencies, which can in turn be estimated from individual anthropometry. However, this requires extraction of three notch tracks, whereas the all-first weight combination only requires the extraction of the first notch and is therefore very appealing for practical applications.

We compared the predicted vertical localization performances (PE) between the HRTF set (other than the individual one) selected with the centroid combination and the one selected with the all-first combination. Gaussianity was verified through a Shapiro-Wilk test for all data sets. Accordingly, a paired t-test was performed, revealing no statistically significant differences on PE ($t(30) = 0.65, p = 0.522$). On the other hand, the comparison between the centroid combination and a combination giving weight to second notch track alone ($w_1 = w_3 = 0$, $w_2 = 1$) revealed a significantly higher PE ($t(30) = 2.71$, $p < 0.05$) for the latter. This finding further reinforces the relevance of the first notch track and the results from our preliminary study using the all-first combination [8].

In light of this result, we further investigated the performance of the all-first combination. Specifically, we compared the predicted vertical localization performances (PE) between the HRTF set selected with the all-first combination and

- the performance of a generic HRTF set (i.e., the KEMAR, CIPIC subject 165);
- the mean of the PEs for all target HRTF sets;[4]
- the performance of the best HRTF set according to the auditory model (i.e., that with the minimum subjective PE).

Fig. 10(a) reports a graphical comparison between the PEs. Gaussianity was verified through a Shapiro-Wilk test for all data sets. Accordingly, paired t-tests were performed with the selected HRTF condition as reference, revealing highly significant statistical differences between generic (KEMAR) and selected HRTFs ($t(30) = 6.56, p \ll 0.001$), between mean PE and selected HRTFs ($t(30) = 3.77, p < 0.001$), and between best and selected HRTFs ($t(30) = 9.6, p \ll 0.001$). This means that elevation localization performances with the selected HRTF set are significantly more accurate than with non-individual HRTFs (generic or randomly chosen). However, the HRTF selected by the mismatch function performs worse on average than the best HRTF of the auditory model.

Nonetheless, the mismatch function is able to detect poorly performing HRTF sets. Fig. 10(c) reports a scatterplot of the mismatch metric samples (with all-first combination) against the PE metric samples for all pairs of 31 subjects. Points on the y-axis refer to individual HRTFs, whose mismatch always scores zero. The figure clearly shows that, although low mismatch values may correspond to high PEs, high mismatch val-

[4]We assume the mean PE of all target HRTFs set as the result of convergence among infinite random selections.

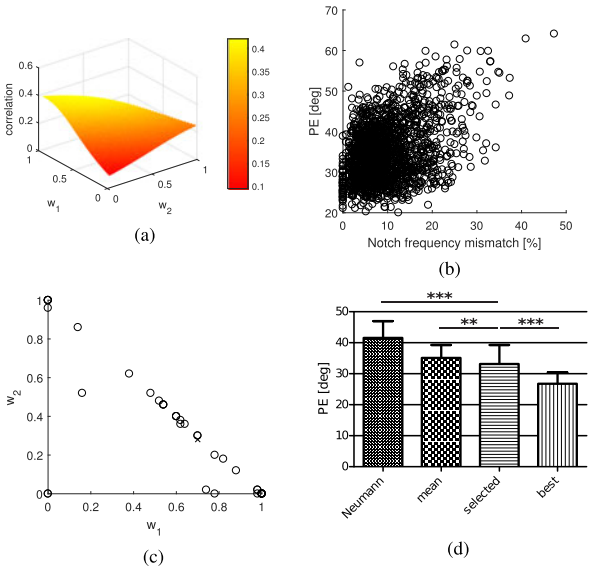


Fig. 11. Main results of the weight tuning procedure outlined in Section V applied to the ARI HRTF database: (a) correlation between PE matrix and notch frequency mismatch matrix as a function of w_1 and w_2; (b) notch frequency mismatch with all-first combination against PE; (c) best subjective weight combinations ($\circ$) and their center of mass ($\times$); (d) mean and standard deviation of predicted PEs with mannequin HRTFs (*Neumann*), mean of all predicted PEs (*mean*), predicted PEs with the non-individual HRTF set selected according to notch frequency mismatch of the first track (*selected*), and best HRTF set according to the auditory model (*best*).

ues never correspond to low PEs. Linear regressions on these data were performed separately for every template HRTF set, revealing positive slopes for all subjects except one. This means that a low mismatch is a necessary yet not sufficient condition for good localization performance, suggesting the possibility of discarding target HRTF sets whose mismatch is large. Interestingly, among 30 subjects with positive slopes, 83% (25) exhibited significant correlation coefficients $r > .036$ with $p < .05$, while the remaining 17% (5) did not. This result can occur due to irregularities of HRTF measurements, erroneous track assignments, or high influence of all mismatch weights (see Fig. 10(a) where weight combinations are far from the center of mass).

C. Weight Tuning on ARI Database

The results reported previously in this Section are based on the analysis of the CIPIC HRTF database. In order to verify whether our results are database-independent, we ran the same analysis on the ARI HRTF database (41 3-track subjects). The main results are reported in Fig. 11 as four plots with the same format as Figs. 9(a), 10(c), (a), and (b) respectively, for the sake of comparison with CIPIC results.

Despite a lower maximum correlation between PE and mismatch matrices [$r = .42$, see Fig. 11(a)] and a more scattered behavior of all-first mismatch values for low PE values [Fig. 11(b)], the similarity with CIPIC results is evident in all plots. The centroid combination [Fig. 11(c)] corresponds to

$$w_1 = 0.7, \quad w_2 = 0.28, \quad w_3 = 0.02, \tag{9}$$

and all comparisons between selection methods applied to the ARI database show the same significances found for the CIPIC database. Specifically, paired t-tests revealed no statistically significant differences on PE ($t(40) = 0.85$, $p = 0.399$) between the HRTF set selected with the centroid combination and the one selected with the all-first combination, while the comparison between the centroid combination and a combination giving weight to second notch track alone ($w_1 = w_3 = 0$, $w_2 = 1$) revealed a significantly higher PE ($t(40) = 2.46$, $p < 0.05$) for the latter. Furthermore, when comparing HRTF selection according to the all-first combination against generic (Neumann KU 100 mannequin) HRTFs, mean PE for all target HRTF sets, and best HRTF set according to the auditory model [see Fig. 11(d)], paired t-tests revealed highly significant statistical differences between generic and selected HRTFs ($t(40) = 10.61$, $p \ll 0.001$), between mean PE and selected HRTFs ($t(40) = 2.81, p < 0.01$), and between best and selected HRTFs ($t(40) = 9.53$, $p \ll 0.001$).

VI. General Discussion

With the increasing number of HRTF datasets publicly available worldwide, subjective HRTF selection becomes practically impossible (due to time constraints, subject uncertainty, and so on) without objective metrics and criteria which allow subset extraction and/or direct selection. Models of auditory perception are crucial in automating the selection process, however they usually require highly detailed characterization of listeners, making them difficult to be employed directly. As an example, the Langendijk model and its extension [14] requires an estimate of uncertainty by performing localization experiments [15] and individual DTFs which are unknown variables in the inferential problem of HRTF selection for vertical localization. In light of such issue, one might search for criteria which are less restrictive in terms of required individual information, at the same time providing performance predictions which are comparable to those of an auditory model that should take into account relevant attributes for describing the perceptual dimensions affected by HRTF set variations [63], e.g. tonal quality discrepancy, externalization, immersion, to name but a few.

Results with this auditory model suggest that if the amount of available data is large enough, it is possible to select HRTF sets that ensure comparable performances to individual ones; in our study, the available subjects are already enough to provide equal-localization performances between individual and selected HRTFs [see Fig. 3(a)]. In particular, some target HRTFs have fine spectral cues that could be easily exploited by all template HRTFs. On the other hand, template HRTFs with rich spectral differences in elevation angles have so many features that could ease adaptation to any target HRTF.

Moreover, there is a clear correlation between number of notch tracks and U: template HRTF sets with 3 tracks naturally performed with high precision and low uncertainty. Accordingly, systematic small errors for certain template HRTFs could easily occur and could be reduced even further by setting $U < 2$. A separate analysis on HRTF sets with 2 tracks will be conducted in a future study, where the focus will be on the in-

terdependent evaluation between weight tuning and individual calibration of non-acoustic factors. Relevant practical aspects will be the definition of a more detailed model for localization uncertainty and effective listening tests able to parametrize it. All this new information could be related to a 2- or 3-track user profile; for identified 2-track subjects, the development of ad-hoc training procedures with non-individual target HRTFs (e.g. those with fine spectral cues) could also reduce individual U.

Further studies are thus mandatory in order to find guidelines in HRTF database design, defining requirements such as number of HRTFs related to gender differences or specific signal patterns (e.g. connected to anthropometry and ethnicity [40]); in this research direction, the main goal could be the definition of a *corpus* in order to perform statistical analysis and data mining for HRTF selection purposes.

Our case study of a spectral notch metric provides a simple HRTF selection method with potentially high impact from an applicative point of view: using the all-first weight combination (i.e. using only 17 numbers for a single listener, the frequencies of the first notch at 17 elevations in the median plane), one can choose a HRTF set which has significantly better predicted performance than (i) a generic set (i.e. the KEMAR HRTF set), and (ii) random selection. Moreover, the correlation analysis between mismatch and PE [see Fig. 10(c)] suggests that the mismatch function with all-first weight combination can quickly aid to compute a personalized short list of N candidate HRTFs, that include with high probability the best available HRTF set according to the auditory model. These N candidates may be further analyzed by a subsequent selection step, employing subjective selection procedures or different data analysis. Given 31 CIPIC / 41 ARI subjects and the Langendijk model, we estimated that a short-list of $N = 10 \pm 7$ (CIPIC) / $N = 20 \pm 11$ (ARI) would be sufficient. These numbers results from the average position ($\pm$ standard deviation) of the best possible HRTF set according to the auditory model on the ordered notch frequency mismatch ranking for each subject. Moreover, one can consider the 95% confidence interval identifying the upper limit of $N = 22$ (CIPIC) and $N = 37$ (ARI) for whom the best possible HRTF set is in the subject-dependent ranking with high-probability, thus allowing a statistically reliable data-set reduction. Differences in N among databases reflect the correlation between PE and mismatch matrices (see Fig. 9(a) for CIPIC, and Fig. 11(a) for ARI).

The practical validity of our approach relies on robust procedures able to compute individual notch frequencies from acoustic and non-acoustic data which should be easily obtainable and handled by users. In this direction, a recent study proposed an easy-to-use tool that guides users in performing an image-guided HRTF selection without previous knowledge on binaural audio technologies [9]; moreover, we have shown that direct acoustic estimation of the first pinna notch could be provided through a self-adjustable procedure which roughly acquires individual HRTFs making use of a smartphone loudspeaker as sound source and binaural microphones as receivers [10]. On the other hand, further investigations are required in order to assess the potential of learning and adaptation effects towards a *localization equivalent* performance with non-individual HRTFs with particular attention to multimodal virtual reality. This issue is even more crucial for immersive augmented reality technologies that require a high level of personalization in order to guarantee the best match between real and virtual acoustic scenes.

It is worthwhile to note that final results are dependent on the choice of a specific auditory model which serves as a ground truth in our research framework. From a methodological point of view, nothing prevents to replicate our study employing different auditory models [51], [52], [57] or other similarity metrics based on HRTF subjective ratings which might be able to characterize human spatial hearing. A strong connection between real listening evaluation and auditory model calibration is the key element for obtaining reliable and meaningful results. Similarly, our research methodology can be adapted to many other mismatch metrics such as those employing CIPIC anthropometric data [64], [65], or multiple regression analysis on both CIPIC anthropometric data and pinna pictures [62]. In principle, it will be possible to quantitatively compare performances of our notch frequency mismatch with other HRTF selection approaches. Finally, we would like to stress that our approach can be extended to HRTF sets for which listeners are no longer available (e.g. no interest in the study anymore, age-related hearing degeneration, death of the subject, etc.), thus building ever-growing and valid databases for research in this field.

VII. Conclusion

Our main goal is to provide reliable and replicable results without performing listening tests or impractical massive subjective evaluations. In the context of this work, we avoided 1089 and 9409 full-sphere localization experiments within CIPIC and ARI subject pools, respectively. The main contributions of this paper can be summarized as follows:

1) using the Langendijk model and its extension [14] with CIPIC and ARI databases, there exists a non-individual HRTF set which allows a listener to have an equally accurate vertical localization than with his/her individual HRTF set; moreover, once detected, this non-individual HRTF might also reduce front-back confusion rate;
2) given the extended Langendijk model and Raykar's notch extraction algorithm [60], we were able to exclude 12 CIPIC and 47 ARI subjects with poor spectral cues (i.e. only 2 notch tracks) and predicted localization performances ($U > 2$) in order to perform data reduction and strengthen data pool for HRTF selection purposes;
3) we investigated in detail the mismatch function of (6), computing the globally optimal combination of weights as the centroid combination reported in (8) for 31 CIPIC subjects and (9) for 41 ARI subjects; from an applicative point of view, we also demonstrated the relevance of the all-first weight combination (considering the first notch only), and showed that it does not cause statistically significant degradation of localization performance compared to the optimal combination, while it still provides a statistically significant improvement with respect to generic HRTFs (i.e., the KEMAR and Neumann KU 100 dummy heads) or random selection;

4) the all-first weight combination correlates significantly with the predicted performance of the auditory model; this allows to define a subject-dependent criterion for dataset reduction.

Our results suggest that the use of auditory models can effectively simulate a virtual localization experiment, thus providing an alternative means to listening tests for assessing the performance of our HRTF selection procedure. Nonetheless, this research might benefit from future listening tests with a large amount of subjects (e.g., through online tests) as well as large HRTF datasets.

ACKNOWLEDGMENT

The authors would like to thank A. Bedin, for his initial contribution to our preliminary study on the notch metric, M. Laroze, and A. Carraro for their help in the development of the experimental framework with auditory models.

REFERENCES

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1983.

[2] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," *J. Audio Eng. Soc.*, vol. 55, no. 7/8, pp. 623–637, Jul. 2007. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14190

[3] J. G. Bolaños and V. Pulkki, "HRIR database with measured actual source direction data," in *Proc. 133rd Audio Eng. Soc. Convection Conf.*, San Francisco, CA, USA, Oct. 2012, pp 1–8.

[4] A. Andreopoulou, D. Begault, and B. Katz, "Interlaboratory round robin HRTF measurement comparison," *IEEE J. Select. Topics Signal Process.*, vol. 9, no. 5, pp. 895–906, Aug. 2015.

[5] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Amer.*, vol. 97, no. 6, pp. 3907–3908, Jun. 1995.

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

[7] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 6, no. 5, pp. 476–488, Sep. 1998.

[8] M. Geronazzo, S. Spagnol, A. Bedin, and F. Avanzini, "Enhancing vertical localization with image-guided selection of nonindividual head-related transfer functions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 4496–4500.

[9] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini, "Improving elevation perception with a tool for image-guided head-related transfer function selection," in *Proc. 20th Int. Conf. Digital Audio Effects*, Edinburgh, U.K., Sep. 2017, pp. 397–404.

[10] M. Geronazzo, J. Fantin, G. Sorato, G. Baldovino, and F. Avanzini, "Acoustic selfies for extraction of external ear features in mobile audio augmented reality," in *Proc. 22nd ACM Symp. Virtual Reality Softw. Technol.*, Munich, Germany, Nov. 2016, pp. 23–26.

[11] B. F. G. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *J. Acoust. Soc. Amer.*, vol. 131, no. 2, pp. EL99–EL105, Feb. 2012.

[12] G. Parseihian and B. F. G. Katz, "Rapid head-related transfer function adaptation using a virtual auditory environment," *J. Acoust. Soc. Amer.*, vol. 131, no. 4, pp. 2948–2957, 2012. [Online]. Available: http://link.aip.org/link/?JAS/131/2948/1

[13] E. H. A. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *J. Acoust. Soc. Amer.*, vol. 112, no. 4, pp. 1583–1596, 2002. [Online]. Available: http://link.aip.org/link/?JAS/112/1583/1

[14] R. Baumgartner, P. Majdak, and B. Laback, "Assessment of sagittal-plane sound localization performance in spatial-audio applications," in *The Technology of Binaural Listening* (Modern Acoustics and Signal Processing Series), J. Blauert, Ed. Berlin, Germany: Springer, Jan. 2013, pp. 93–119.

[15] P. Majdak, R. Baumgartner, and B. Laback, "Acoustic and nonacoustic factors in modeling listener-specific performance of sagittal-plane sound localization," *Frontiers Psychology*, vol. 5, pp. 1–10, Apr. 2014.

[16] S. Paul, "Binaural recording technology: A historical review and possible future developments," *Acta Acust. United Acust.*, vol. 95, no. 5, pp. 767–788, Sep. 2009.

[17] H. Møller, M. Sørensen, J. Friis, B. Clemen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?" *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, 1996.

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

[19] G. D. Romigh and B. D. Simpson, "Do you hear where I hear?: Isolating the individualized sound localization cues," *Frontier Neurosci.*, vol. 8, 2014, Art. no. 370.

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

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

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

[23] E. A. G. Shaw and R. Teranishi, "Sound pressure generated in an external-ear replica and real human ears by a nearby point source," *J. Acoust. Soc. Amer.*, vol. 44, no. 1, pp. 240–249, 1968.

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

[25] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. Head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1465–1479, Sep. 1999.

[26] S. Spagnol, "On distance dependence of pinna spectral patterns in head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 137, no. 1, pp. EL58–EL64, Jan. 2015.

[27] 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, Aug. 2007.

[28] B. C. J. Moore, S. R. Oldfield, and G. J. Dooley, "Detection and discrimination of spectral peaks and notches at 1 and 8 kHz," *J. Acoust. Soc. Amer.*, vol. 85, no. 2, pp. 820–836, Feb. 1989.

[29] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1493–1510, 1999.

[30] D. S. Brungart and G. D. Romigh, "Spectral HRTF enhancement for improved vertical-polar auditory localization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New York, NY, USA, Oct. 2009, pp. 305–308.

[31] A. Lindau, J. Estrella, and S. Weinzierl, "Individualization of dynamic binaural synthesis by real time manipulation of ITD," in *Proc. Audio Eng. Soc. Conv.*, 2010, pp. 1–10.

[32] M. Aussal, F. Alouges, and B. F. Katz, "ITD interpolation and personalization for binaural synthesis using spherical harmonics," in *Proc. Audio Eng. Soc. UK Conf.*, 2012, pp. NN1–N10.

[33] C.-J. Tan and W.-S. Gan, "User-defined spectral manipulation of HRTF for improved localisation in 3-D sound systems," *Electron. Lett.*, vol. 34, no. 25, pp. 2387–2389, Dec. 1998.

[34] R. H. Y. So, N. M. Leung, A. B. Horner, J. Braasch, and K. L. Leung, "Effects of spectral manipulation on nonindividualized head-related transfer functions (HRTFs)," *Human Factors: J. Human Factors Ergonom. Soc.*, vol. 53, no. 3, pp. 271–283, Jun. 2011.

[35] C. Mendonça, G. Campos, P. Dias, and J. A. Santos, "Learning auditory space: Generalization and long-term effects," *PLoS One*, vol. 8, no. 10, Oct. 2013, Art. no. e77900.

[36] R. Trapeau, V. Aubrais, and M. Schnwiesner, "Fast and persistent adaptation to new spectral cues for sound localization suggests a many-to-one mapping mechanism," *J. Acoust. Soc. Amer.*, vol. 140, no. 2, pp. 879–890, Aug. 2016.

[37] Y. Iwaya, "Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears," *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 340–343, 2006.

[38] S. Hwang, Y. Park, and Y.-S. Park, "Customization of spatially continuous head-related impulse responses in the median plane," *Acta Acust. United Acust.*, vol. 96, no. 2, pp. 351–363, Mar. 2010.

[39] 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.
[40] B. Xie, X. Zhong, and N. He, "Typical data and cluster analysis on head-related transfer functions from Chinese subjects," *Appl. Acoust.*, vol. 94, pp. 1–13, Jul. 2015.
[41] D. Zotkin, R. Duraiswami, and L. Davis, "Rendering localized spatial audio in a virtual auditory space," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 553–564, Aug. 2004.
[42] J. C. Middlebrooks, E. A. Macpherson, and Z. A. Onsan, "Psychophysical customization of directional transfer functions for virtual sound localization," *J. Acoust. Soc. Amer.*, vol. 108, no. 6, pp. 3088–3091, Dec. 2000.
[43] A. Honda, H. Shibata, J. Gyoba, K. Saitou, Y. Iwaya, and Y. Suzuki, "Transfer effects on sound localization performances from playing a virtual three-dimensional auditory game," *Appl. Acoust.*, vol. 68, no. 8, pp. 885–896, Aug. 2007.
[44] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New York, NY, USA, Oct. 2001, pp. 1–4.
[45] P. Majdak, M. J. Goupell, and B. Laback, "3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Attention Perception Psychophys.*, vol. 72, no. 2, pp. 454–469, Feb. 2010.
[46] M. Geronazzo, S. Spagnol, and F. Avanzini, "Estimation and modeling of pinna-related transfer functions," in *Proc. 13th Int. Conf. Digital Audio Effects*, Graz, Austria, Sep. 2010, pp. 431–438.
[47] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robots—Building audio-motor maps based on the HRTF," in *Proc. 2006 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2006, pp. 1170–1176.
[48] H. Nakashima and T. Mukai, "3-D sound source localization system based on learning of binaural hearing," in *Proc. 2005 IEEE Int. Conf. Syst.*, Man, Cybern., Oct. 2005, vol. 4, pp. 3534–3539.
[49] Y. Luo, D. Zotkin, and R. Duraiswami, "Gaussian process data fusion for heterogeneous HRTF datasets," in *Proc. 2013 IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2013, pp. 1–4.
[50] C. Jin, M. Schenkel, and S. Carlile, "Neural system identification model of human sound localization," *J. Acoust. Soc. Amer.*, vol. 108, no. 3, pp. 1215–1235, Sep. 2000.
[51] P. Zakarauskas and M. S. Cynader, "A computational theory of spectral cue localization," *J. Acoust. Soc. Amer.*, vol. 94, no. 3, pp. 1323–1331, Sep. 1993.
[52] P. M. Hofman and A. J. V. Opstal, "Spectro-temporal factors in two-dimensional human sound localization," *J. Acoust. Soc. Amer.*, vol. 103, no. 5, pp. 2634–2648, 1998.
[53] A. J. V. Opstal, J. Vliegen, and T. V. Esch, "Reconstructing spectral cues for sound localization from responses to rippled noise stimuli," *PLOS One*, vol. 12, no. 3, Mar. 2017, Art. no. e0174185.
[54] 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 Interactions Virtual Environ.*, Reutlingen, Germany, Mar. 2018, pp 1–5.
[55] S. Spagnol, M. Hiipakka, and V. Pulkki, "A single-azimuth pinna-related transfer function database," in *Proc. 14th Int. Conf. Digital Audio Effects*, Paris, France, Sep. 2011, pp. 209–212.
[56] P. Minnaar, J. Plogsties, and F. Christensen, "Directional resolution of head-related transfer functions required in binaural synthesis," *J. Audio Eng. Soc*, vol. 53, no. 10, pp. 919–929, Oct. 2005.
[57] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *J. Acoust. Soc. Amer.*, vol. 136, no. 2, pp. 791–802, 2014.
[58] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. II: Psychophysical validation," *J. Acoust. Soc. Amer.*, vol. 85, no. 2, pp. 868–878, Feb. 1989.
[59] G. Andéol, S. Savel, and A. Guillaume, "Perceptual factors contribute more than acoustical factors to sound localization abilities with virtual sources," *Front. Neurosci.*, vol. 8, 2015, Art. no. 451.
[60] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *J. Acoust. Soc. Amer.*, vol. 118, no. 1, pp. 364–374, Jul. 2005.
[61] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.
[62] S. Spagnol and F. Avanzini, "Frequency estimation of the first pinna notch in head-related transfer functions with a linear anthropometric model," in *Proc. 18th Int. Conf. Digital Audio Effects*, Trondheim, Norway, Dec. 2015, pp. 231–236.
[63] L. S. R. Simon, N. Zacharov, and B. F. G. Katz, "Perceptual attributes for the comparison of head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 140, no. 5, pp. 3623–3632, Nov. 2016.
[64] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Rendering localized spatial audio in a virtual auditory space," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 553–564, Aug. 2004.
[65] K. Iida, Y. Ishii, and S. Nishioka, "Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae," *J. Acoust. Soc. Amer.*, vol. 136, no. 1, pp. 317–333, Jul. 2014.

Michele Geronazzo received the M.S. degree in computer engineering and the Ph.D. degree in information and communication technology from the University of Padova, Padova, Italy, in 2009 and 2014, respectively. Between 2014 and 2017, he was a Postdoctoral Researcher with the University of Padova and the University of Verona, Verona, Italy, in the fields of ICT and neurosciences. He is currently a Postdoctoral Researcher with Aalborg University Copenhagen, Kbenhavn, Denmark, where he is with the Multisensory Experience Laboratory in developing his research project "Acoustically-trained 3-D audio models for virtual reality applications" (main topics: virtual acoustics, headphones, and binaural hearing). He is a coauthor of more than fifty scientific publications. His main research interests include binaural spatial audio modeling and synthesis, multimodal virtual/augmented reality, and sound design for human–computer interaction. His Ph.D. thesis was honored by the Acoustic Society of Italy with the "G. Sarcedote" award for best Ph.D. thesis in acoustics. He is a member of the organizing committee of the IEEE VR Workshop on Sonic Interactions for Virtual Environments since 2015 (Chair of the 2018 edition). He served as a Guest Editor for *Wireless Communications and Mobile Computing* (Wiley and Hindawi, 2018). He is a corecipient of four best paper/poster awards.

Simone Spagnol received the Ph.D. degree in information and communication technology from the University of Padova, Padova, Italy, in 2012. Since 2012, he has been a Postdoctoral Researcher with the Iuav University of Venice, Venice, Italy, the University of Iceland, Reykjavk, Iceland, and the University of Padova, where he is currently a Senior Postdoctoral Fellow. He has authored more than 50 scientific publications, and has served as Guest Editor for *Wireless Communications and Mobile Computing* in 2017–2018. His research interests include 3-D sound technologies and sonic interaction design applied to multimodal assistive technologies. He has been a key researcher and principal investigator in several international and national research projects, including two Horizon 2020 EU projects. He has participated in the organization of national and international conferences, and he received four best paper awards as first author.

Federico Avanzini received the Ph.D. degree in computer science from the University of Padova, Padova, Italy, in 2002, and he was there until 2017, as a Postdoctoral Researcher, an Assistant Professor, and an Associate Professor. He is as an Associate Professor with the University of Milano, Milano, Italy, since January 2018. He has authored about 150 publications on peer-reviewed international journals and conferences, and has served in several program and editorial committees. His main research interests include algorithms for sound synthesis and processing, nonspeech sound in human–computer interfaces, and multimodal interaction. He has been a key researcher and principal investigator in several national and international research projects. He was the General Chair of the 2011 International Conference on Sound and Music Computing, and is currently an Associate Editor for the international journal *Acta Acustica United With Acustica*.