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, (2019 June.).
DOI: https://doi.org/10.17743/jaes.2019.0010

Applying a Single-Notch Metric to Image-Guided Head-Related Transfer Function Selection for Improved Vertical Localization

MICHELE GERONAZZO,[1] *AES Member*, **ENRICO PERUCH**[2], **FABIO PRANDONI**[2], **AND FEDERICO AVANZINI**[3]
(mge@create.aau.dk)

[1]*Dept. of Architecture, Design, and Media Technology, Aalborg University, Copenhagen, Denmark*
[2]*Dept. of Information Engineering, University of Padova, Padova, Italy*
[3]*Dept. of Computer Science, University of Milano, Milano, Italy*

This paper proposes an image-guided HRTF selection procedure that exploits the relation between features of the pinna shape and HRTF notches. Using a 2D image of a user's pinna, the procedure selects from a database the HRTF set that best fits the anthropometry of that user. The proposed procedure is designed to be quickly applied and easy to use for a user without previous knowledge on binaural audio technologies. The entire process is evaluated by means of (i) an auditory model for sound localization in the mid-sagittal plane available from previous literature, and (ii) a short localization test in virtual reality. Using both virtual and real subjects from an HRTF database, predictions and the experimental evaluation aimed to assess the vertical localization performance with HRTF sets selected by the proposed procedure. Our results report a statistically significant improvement in predictions of the auditory model for localization performance with selected HRTFs compared to KEMAR HRTFs, which is a commercial standard in many binaural audio solutions. Moreover, the proposed localization test with human listeners reflect the model's predictions, further supporting the applicability of our perceptually-motivated metrics with anthropometric data extracted by pinna images.

0 INTRODUCTION

Our auditory system continuously captures everyday acoustic scenes and acquires spatial information by processing temporal and spectral features of sound sources related to both the environment and the listeners themselves. Knowledge of such a complex process is needed in order to develop accurate and realistic artificial sound spatialization algorithms (see [1, 2] for a systematic review) in several application domains, including music listening, entertainment, immersive virtual and augmented reality (VR/AR), sensory substitution or aid devices, tele-operation, teleconferencing, and so on (see [3] for trends in binaural technologies and models, and [4] for an overview of possible application areas applied to VR/AR).

Many of the above mentioned scenarios require spatial sound to be delivered through headphones. This 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. Having a set of HRIRs (or Head-Related Transfer Functions - HRTFs, their Laplace transforms) measured over a discrete set of spatial locations allows to spatially render a dry sound by convolving it with the desired HRIR pair. Moving sound sources can also be rendered by suitably interpolating spatially neighboring HRIRs.

The ability to localize sound sources is important in several everyday activities. Accordingly, localization accuracy is a relevant auditory quality even in *Virtual Auditory Displays* (VADs) [5]. This paper deals in particular with elevation localization cues, which are mainly provided by monaural spectral features of the HRTF [6]. Specifically, the scattering of acoustic waves in the proximity of the pinna creates a complex and individual topography of pressure nodes that is not completely understood [7, 8], and results in elevation- and listener-dependent peaks and notches that appear in the HRTF spectrum in the range [3,16] kHz. This monaural information complements binaural cues such as *interaural time difference* (ITD) and *interaural level difference* (ILD), which are mainly related to localization in

the horizontal plane and are almost constant with varying elevations. However, both ITD and ILD cues connected to the source lateral angle determine the binaural weighting of monaural spectral cues in sagittal planes [9]. For the sake of completeness, the contribution of torso reflections and acoustic shadow below 3 kHz should also be considered as an additional elevation cue, which is less perceptually relevant than high-frequency content [10] but still relevant for other perceptual sound field qualities such as source width or envelopment [11, 12].

Individual anthropometric features of the human body have a key role in shaping individual HRTFs (see the discussion in Sec. 1 below). This paper proposes an image-guided HRTF selection technique that builds on previous work on the relation between features of the pinna shape and HRTF notches [13]. Using a 2D image of a subject's pinna, the procedure selects from a database the HRTF set that best fits the anthropometry of that subject. One of the challenging issues with this approach is the trade off between handiness of pinna feature acquisition and localization performance in elevation; since the procedure in our previous research [13] relied on expert operators for the extraction of anthropometric information, this work provides an easy to use tool for a user without previous knowledge on pinna acoustics and spatial hearing.

Auditory localization performance with HRTF sets is usually assessed through psychoacoustic experiments with human subjects. However, an attractive alternative approach consists in using computational auditory models able to simulate the human auditory system. If the auditory model is well calibrated to the reality, a perceptual metric can be developed to predict the perceptual performance of a VAD. This approach was successfully employed by some of the authors in a systematic investigation of salient spectral features of HRTFs for elevation perception with particular attention to notch contribution in localization cues from 150 HRTF sets [14]. In this paper the applicability of those findings is further studied deducting spectral features indirectly from pinna geometry, i.e., contours, rather than extracting notch frequencies from HRTF data.

The proposed HRTF selection procedure is here validated through both an auditory model and a psychoacoustic listening test. Specifically, this paper extends a recent publication [15] that provided a preliminary validation of the procedure based on an auditory model only.[1]

Regarding the first point, i.e., auditory model predictions, we adopted the CIPIC database [16] in which HRTFs and side-pictures of the pinna are available. The applicability of the proposed notch distance metric are also discussed in terms of individual HRTF identification from images. Predictions in elevation perception are evaluated by means of an auditory model for sound localization in the mid-sagittal plane [17] (i.e., the vertical plane dividing the listener's head in left and right halves) provided by the Auditory Modeling Toolbox.[2] Using virtual subjects from the CIPIC database, we present a virtual experiment that simulates the vertical localization performance of CIPIC subjects when they are provided with HRTF sets selected by the proposed procedure. Finally, we assessed results from model predictions through a short listening test in virtual reality aimed at investigating the localization ability of participants. The long-term goal of our approach is to replace time- and resource-consuming psychoacoustic tests with auditory model predictions, in order to provide a fast assessment system for different HRTF selection criteria.

1 RELATED WORKS

One of the main limitations of binaural audio technologies for commercial uses is the hard work needed to estimate individual HRTFs that capture all of the physical effects creating a personal perception of immersive audio. The measurement of a listener's individual HRTFs in all directions requires a special measuring apparatus and a long measurement time, often a too heavy task to perform for users of real-world applications. That is the main reason why alternative ways are preferred, that provide listeners with personalized, albeit not individual, HRTF sets: a trade off between quality and costs of the acoustic data for audio rendering [18].

1.1 Individual HRTFs

The standard setup for individual HRTF measurement requires an anechoic chamber with a set of loudspeakers mounted on a geodesic sphere (with a radius of at least one meter in order to avoid near-field effects) at fixed intervals in azimuth and elevation. Listeners, seated in the center of the sphere, have microphones in their ears. After subject preparation, HRIRs are measured playing analytic signals and recording responses collected at the ears for each loudspeaker position in space (see Geronazzo [2] for a systematic review on this topic).

The main goal is to extract the set of HRTFs for every listener thus providing them individual transfer functions. In addition to the above mentioned high demanding requirements (time and equipment), there are some more critical aspects in HRTF measurements; listener's pose is usually limited to a few positions (standing or sitting), a relatively small number of specific locations around the body are measured, and time-invariance of the measurements is implicitly assumed (e.g., without considering that the external ear is one of the parts of the human body that always grows during lifetime [19]). Moreover, repeatability [20] and required accuracy of HRTF and anthropometric [21] measurements are still delicate issues.

1.2 Personalized and Generic HRTFs

Personalized HRTFs can be chosen among those available in a dataset, instead of performing individual measurements. This procedure is based on a match between an

[1] The present paper was invited for submission to the J. of the Audio Eng. Soc. as a result of our previous publication [15] winning the 3rd Best Paper Award of the 2017 Int. Conf. on Digital Audio Effects (DAFx).

[2] http://amtoolbox.sourceforge.net/

external subject (the one without individual HRTFs) and a set of internal subjects (i.e., belonging to a database), for whom acoustics and anthropometric information is available. The most interesting and important point in HRTF selection is thus the method by which a specific set of HRTFs is selected from the database to match as closely as possible those of the external subject. Researchers are finding different ways to deal with this issue and there is a variety of alternatives using common hardware and/or software tools. The main benefit of this approach is that users can autonomously select of their best HRTF set without needing special equipment or knowledge. Personalized HRTFs cannot guarantee the same performance as individual ones, but they usually provide better performance than the generic dummy-head HRTFs such as those of the Knowles Electronic Manikin for Acoustic Research (KEMAR) [22].

In a recent survey on HRTF individualization, Guezenoc and Séguier [23] classify existing approaches into four main groups: acoustic measurements, numerical simulations, anthropometry-based methods, and perception-based methods. The two latter groups are further subdivided into approaches relying on adaptation of non-individual sets, and those employing automatic selection from a database, which is the focus of this paper. Anthropometry-based selection methods (see, e.g., [24]) are based on finding the best-matching HRTF in the anthropometric domain, i.e., those that best match the external ear shape of a subject using anthropometric measurements available in the database (see Sec. 2.2 for further details). Perception-based methods on the other hand are based on listener's feedback: examples are the DOMISO system [25] (the acronym stands for Determination method of OptimuM Impulse-response by Sound Orientation) in which subjects choose their preferred HRTFs from a database through tournament-style listening tests, and reinforcement-learning-based personalization [26] in which participants' evaluation of virtual source location guides the parameterization of an auto-regressive moving average (ARMA) model for generic HRTFs. Results of both studies show that personalized HRTFs selected in this way perform comparably to individualized HRTFs (and better than generic ones). Two-step approaches also exist [27]: the first step typically selects one subset from an initial larger pool of HRTF sets, removing those that perform worst from a perceptual point of view, while the second step refines the selection in order to obtain the best match among generic HRTFs of a dataset that is reduced in size compared to the complete database.

2 IMAGE-GUIDED HRTF SELECTION

Our approach to the selection problem consists in mapping anthropometric features into the HRTF domain, following a ray-tracing modeling of pinna acoustics [28, 29]. The main idea is to draw pinna contours on an image. Distances from the ear canal entrance define reflections on pinna borders generating spectral notches in the HRTF. Accordingly, one can use such anthropometric distances and corresponding notch parameters to choose the best match among a pool of available HRTFs [13].

2.1 Notch Distance Metrics

The extraction of HRTFs using reflections and contours is based on an approximate description of the acoustical effects or the pinna on incoming sounds. In particular, the distance d_c between a reflection point on the pinna and the entrance of the ear canal (the "focus point" hereafter) is given by:

$$d_c(\phi) = \frac{c t_d(\phi)}{2}, \tag{1}$$

where $t_d(\phi)$ is elevation-dependent temporal delay between the direct and the reflected wave, and c is the speed of sound.

The corresponding notch frequency depends on the sign of the reflection. Assuming the reflection coefficient to be positive, a notch is created at all frequencies such that the phase difference between the reflected and the direct wave is equal to a half-wavelength:

$$f_n(\phi) = \frac{2n+1}{2t_d(\phi)} = \frac{c(2n+1)}{4d_c(\phi)}, \tag{2}$$

where $n \in \mathbb{N}$. Thus, the first notch frequency is found when $n = 0$, giving the following result:

$$f_0(\phi) = \frac{c}{4d_c(\phi)}. \tag{3}$$

In fact, a previous study [29] on the CIPIC database [16] proved that almost 80% of the subjects in the database exhibit a clear negative reflection in their HRIRs. Under this assumption notches are found at full-wavelength delays resulting in the following equation:

$$f_n(\phi) = \frac{n+1}{t_d(\phi)} = \frac{c(n+1)}{2d_c(\phi)}, \tag{4}$$

where $n \in \mathbb{N}$, and

$$f_0(\phi) = \frac{c}{2d_c(\phi)}. \tag{5}$$

In particular, a recent work by Zonooz and colleagues [30] suggested that the human auditory system performs a weighted spectral analysis within the $6-9$ kHz frequency range in which the pinna's most prominent elevation-related notch (sometimes referred to as the first notch, $N1$ [7]) occurs. Accordingly, it has been shown [13] that this first HRTF notch is typically associated to the most external pinna contour on the helix border (the C_1 contour hereafter). Based on this finding, we proposed a mismatch function between sets of first-notch frequencies belonging to pairs of HRTFs, which provides a measure of the distance between these HRTFs.

More precisely, assume that N estimates of the C_1 contour and K estimates of the focus point are available from the pinna image of a subject, as depicted in Fig. 1. We define the basic notch distance metric in the form of a mismatch function between the corresponding notch frequencies computed with Eq. (4), and the notch frequencies of an HRTF in the database:

$$m_{(k,n)} = \frac{1}{N_\varphi} \sum_{\varphi} \frac{f_0^{(k,n)}(\varphi) - F_0(\varphi)}{F_0(\varphi)}, \tag{6}$$

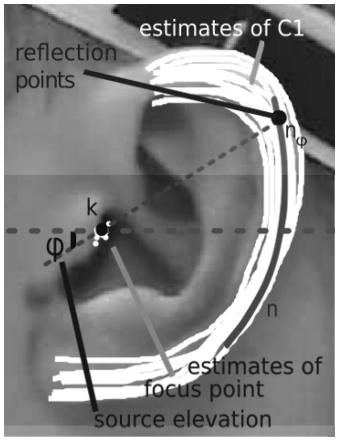


Fig. 1. Schematic view of the required parameters from pinna anthropometry. The $n-th$ contour (in dark grey) belongs to the N estimates (in white) of the external pinna contour C_1, $k-th$ point belongs to the K estimates of the focus point, and n_ϕ is one of the N_ϕ reflection points for each single contour.

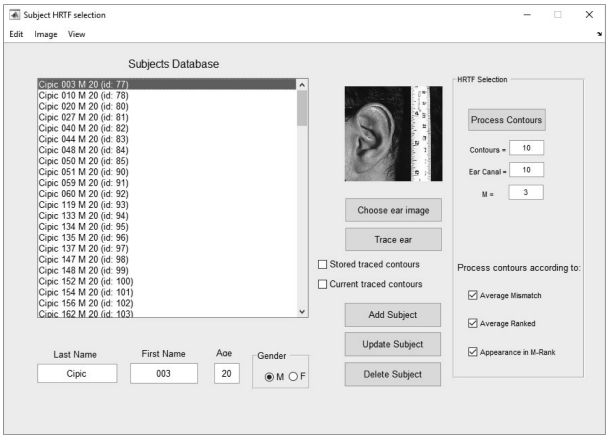


Fig. 2. The proposed tool for HRTF selection: main Graphical User Interface. Users can manipulate the main parameters for the HRTF ranking: number of estimations of C_1, N, number of estimations of the focus point, K, and the restriction of M positions for ranking purposes (see Sec. 3).

where $f_0^{(k,n)}(\varphi)$ are the frequencies extracted from the image and contours of the subject using Eq. (4), and F_0 are the notch frequencies extracted from the HRTF in the database with *ad-hoc* algorithms [28, 31, 24]; (k, n) with $(0 \leq k < K)$ and $(0 \leq n < N)$ refers to a one particular pair of traced C_1 contour and focus point; φ spans all the $[-45°, +45°]$ elevation angles for which the notch is present in the corresponding HRTF; N_φ is the number of elevation angles on which the summation is performed. Notches extracted from HRTFs at single elevations need to be grouped into a track evolving through elevation consistently [29].

If the notches extracted from the subject's pinna image are to be compared with a set of HRTFs taken from a database, various notch distance metrics can be defined based on this mismatch function, to rank database HRTFs in order of similarity. In particular, we define two metrics:

- **Mismatch:** each HRTF is assigned a similarity score that corresponds exactly to increasing values of the mismatch function calculated with Eq. (6) (for a single (k, n) pair).
- **Ranked position:** each HRTF is assigned a similarity score that is an integer corresponding to its ranked position taken from the previous mismatch values (for a single (k, n) pair).

2.2 A HRTF Selection Tool

Based on the concepts outlined above, we propose a tool for selecting from a database an HRTF set that best fits the image of a subject's pinna. The C_1 contour and the focus point are traced manually on the pinna image by an operator, and then the HRTF sets in the database are automatically ranked in order of similarity with the subject. The tool is implemented in Matlab and it is freely available at https://github.com/msmhrtf/sel under GPL3.0 license.

2.2.1 Graphical User Interface

Fig. 2 provides a screenshot of the main GUI, which is responsible for managing subjects and organizing them in a list (on the left of the screen). The list can be managed using the buttons, "*Add Subject,*" "*Update Subject,*" and "*Delete Subject,*" as well as some text fields used to assign to each subject their own information. For each subject stored in the list, an image of the left pinna can be assigned with the button "*Choose ear image*": the image will be shown in the middle of the GUI when a name from the list is clicked.

After loading the pinna image of a subject, the main pinna contour C_1 and the focus point can be traced manually by clicking on the "*Trace Ear*" button. Two parameters N and K can be specified, which are the number of estimates that the operator will trace for the C_1 contour and the focus point, respectively. One last parameter to be set, M, refers to the top-M appearance metrics discussed in Sec. 3.

Two checkboxes under the "*Trace Ear*" button aid the usability of the tracing task: the first one is the "*Stored traced contours*" that shows the already drawn contours in the previous drawing session. The second one, called "*Current traced contours*" is about visualizing on the pinna image the contours drawn in the current session.[3] By clicking on the "*Process Contours,*" the application returns the ranked positions of the database HRTFs according to defined three metrics.

2.2.2 Database of Generic HRTFs and F_0 Estimation

The public database used for our purpose is the CIPIC [16]. The first release provided HRTFs for 43 subjects (plus 2 dummy-head KEMAR HRTF sets with small and large pinnae, respectively) at 25 different azimuths and 50 different elevations, to a total of 1250 directions. In addition, this database includes a set of pictures of external ears and anthropometric measures for 37 subjects. Information of

[3] The default tracing procedure allows drawing a single contour/focus point at a time, that visually disappears once traced; for every estimate, our tool shows pinna images clean from traced information.

the first prominent notch in each HRTF was extracted from outputs of the *structural decomposition algorithm* [31] that allows the separation of resonances (peaks) from reflections (notches) through an *analysis-by-synthesis* approach. An iterative compensation of the HRTF magnitude, $|H_0|$, follows these steps at each i^{th} −iteration :

1. Computation of the spectral envelope for $|H_i|$;
2. Removal of the spectral envelope from $|H_i|$;
3. Fitting a multi-notch filter to the result of step 2;
4. Removal the filter obtained in step 3 from $|H_i|$;
5. $|H_{i+1}| \leftarrow$ output of step 4.

The process converges in the condition of no local notches above a given amplitude threshold found in the filter computed in step 3. The resulting final spectrum $|H_{end}|$ contains the resonant component alone, while the reflective component is given by cascade combination of all multi-notch filters computed at step 3. Finally, the notch center frequencies extracted by the algorithm can be arranged into "tracks" that describe the evolution of such frequencies with elevation. In this work only $F_0(\varphi)$ tracks were estimated using an *ad hoc* notch tracking algorithm [29].

2.2.3 Guidelines for Contour Tracing

In the GUI, the user has to draw by hand N estimates of the C_1 contours on top of a pinna image. After that, the user has to click on K estimates of the focus point. The rationale behind this is that by averaging over repeated attempts we aim at reducing errors due to operator's mistakes and inherent ambiguities of the tracing task (as an example, the true location of the ear canal entrance is not known from the 2D image and must be guessed). By working on the application, we have derived some empirical guidelines for the tracing task that can be useful for future non-expert operators. In particular, the most effective way to trace the C_1 contour from the image is to cover the area of C_1 with N curves, starting from the internal edge to the external edge of C_1 and vice versa, while the most effective way to trace the focus point is to guess the ear canal entrance with K points in the most likely area. In other words, the tracing procedure is a simplified version of the *optimal focus* estimation procedure proposed in a previous work [29] where a minimization problem was solved by searching in a wide area near the pinna *tragus* tracing several specific contours. On the other hand, real case applications allow the operator to easily localize where the ear canal is on the physical human ear, reducing also uncertainty for the estimation of external pinna contours.

2.3 Evaluation

The main aim of the proposed validation procedure is to verify the effectiveness of our HRTF selection tool in providing participants with HRTFs that are reasonably close to their individual HRTF by only using a picture of their external ear. Strengths and limits of such an approach are discussed (i) in terms of the notch frequency mismatch in Eq. (6), (ii) with the support of an auditory model to predict performance in elevation perception, and (iii) through a real localization experiment with human subjects in virtual reality. Fig. 3 depicts a schematic view of this three-stage validation process.

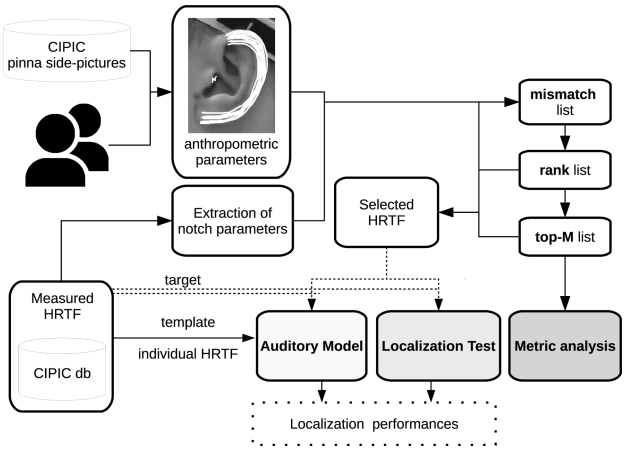


Fig. 3. Schematic view of the proposed validation procedure with metric analysis, auditory model predictions, and localization experiments in VR.

3 DATA ACQUISITION AND METRIC ANALYSIS

Our experimental subjects were taken from the CIPIC database. In particular, we selected the 22 CIPIC subjects for which a complete set of data was available (HRTF, anthropometry, and ear pictures). We chose to draw $N = 10$ estimates of the C_1 contour and $K = 10$ estimates of the focus point, a good trade off that guarantees enough accuracy and fast completion of the selection procedure. The parameter M was set to 3. The entire procedure of creating a subject, retrieving the picture and anthropometric measures, and drawing the contours and focus points takes about five minutes for each subject. Data processing time is negligible. With these settings each subject has $N \times K = 100$ pairs of contours and focus points ready to be processed.

The results of the computation are three rankings of 43 HRTF sets (CIPIC's dummy heads were excluded for homogeneity) derived from our metrics:

- **Average mismatch:** CIPIC subjects are sorted according to their mismatch values (averaged over the $N \times K$ estimates), in increasing order of mismatch.
- **Average rank:** CIPIC subjects are sorted according to their ranks (averaged over the $N \times K$ estimates) in the mismatch ranking, in increasing order of rank.
- **Top-M appearance:** for a given integer M, CIPIC subjects are sorted according to the number of times in which they appear in the first M positions of the ranking, over the $N \times K$ estimates, in decreasing order of occurrence count.

We defined three *best fitting HRTFs* by choosing the HRTFs ranking first in each of the metrics: the best average

mismatch (**best m**), best average rank (**best r**), and best top-3 rank (**best top3**) selected HRTFs.

A preliminary analysis on data distributions of mismatch and rank values showed that normality assumption was violated according to a Shapiro-Wilk test; thus, two Kruskal Wallis nonparametric one-way ANOVAs with three levels of feedback condition (individual, dummy-head KEMAR, best m) and (individual, dummy-head KEMAR, best r) were performed to assess the statistical significance of mismatch and rank metrics, respectively, on all traced pinna contours and ear-canal points. Pairwise post-hoc Wilcoxon tests for paired samples with Holm-Bonferroni correction procedures on p-values provided statistical significance in performance between conditions.

3.1 Results

A preliminary analysis on data distribution of rank values derived from mismatches between $f_0^{(k,n)}(\varphi)$ and individual HRTF's $F_0(\varphi)$ (22×100 observations) was conducted in order to identify the existence of outliers for our metrics. Samples in the last quartile of this distribution were considered cases of limited applicability for the proposed notch distance metric, showing a rank position greater than 27.25 of a total of 43.

Leaving aside for a moment the discussion on applicability of our metrics, we considered the last quartile value as a threshold for the average rank position of each individual HRTF in order to discard CIPIC subjects that cannot be classified according to our criteria and for which no firm conclusions can be drawn. After the application of such threshold, the same analysis was performed on 17×100 observations, i.e., 5 subjects were removed; the 75% of the observations had a rank position less than 18 which is in the first half of the available positions. Moreover, the median value for rank position is 8, which suggests data convergence to the first rank positions.

Fig. 4 depicts the three typical tracing scenarios: (a) a consistent trace-notch correspondence, (b) a systematic lowering in notch frequency of traces, and (c) an irregular notch detection. In the first case, traced contours and individual HRTF notches are in the same range resulting in the ideal condition of applicability for the proposed metric. The latter situation occasionally occurred due to irregularities of HRTF measurements or erroneous track label assignment of $F_0(\varphi)$ evolving through elevation (in two of the five subjects that were previously removed).[4] On the other hand, the case where a systematic lowering in notch frequency of traces occurred (in three of the five subjects previously removed) deserves a more careful consideration: from one of our previous studies [29], we identified a 20% of CIPIC subjects for whom a positive reflection coefficient better models the acoustic contribution of the pinna. Accordingly,

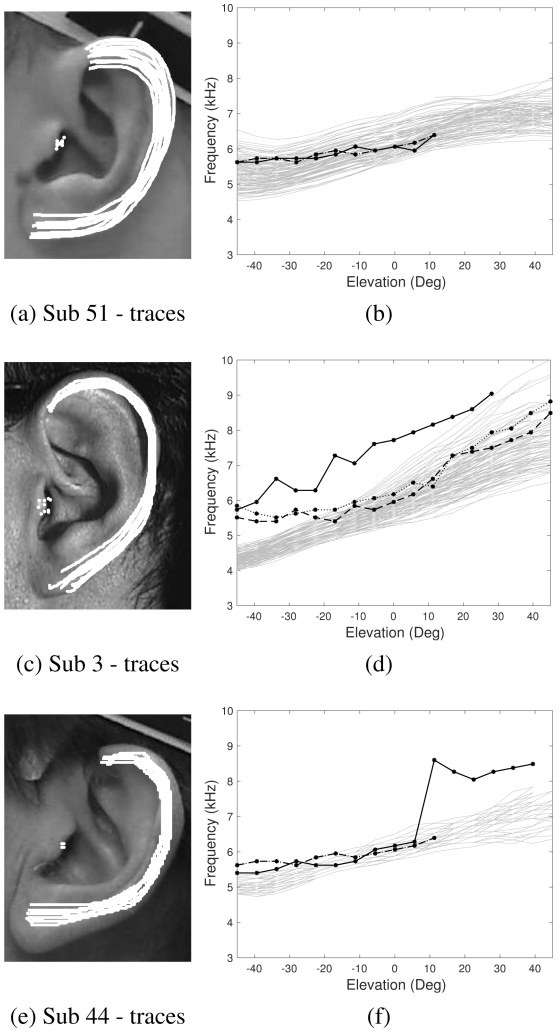


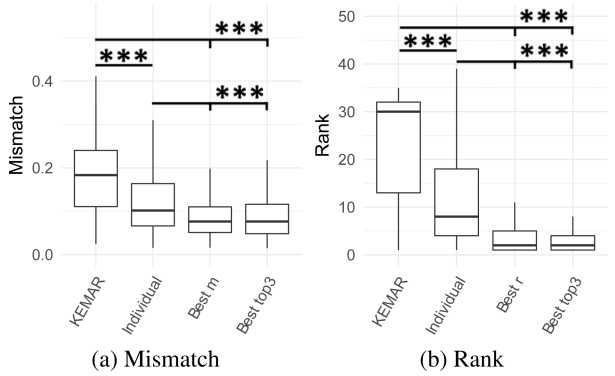
Fig. 4. (a, c, e) Examples of traced C_1/focus points for three CIPIC subjects; (b, d, f) corresponding $f_0^{(k,n)}(\varphi)$ (light gray lines) with $F_0(\varphi)$ values of individual HRTFs (black solid line), best selection according to mismatch/rank metric (black dotted line), and best selection according to Top-3 metric (black dash-dotted line). In these examples, best HRTF selection according to mismatch and rank metrics do not differ significantly.

it is reasonable to think that those three excluded subjects can be assigned to this special group.[5]

Our metrics based on notch distance clearly distinguish the three sets of HRTF, i.e., individual, KEMAR, and best selected, in terms of mismatch and rank (Fig. 5 shows this aspect); Kruskal Wallis nonparametric one-way ANOVAs with four levels of feedback condition (individual HRTF, KEMAR, best m, best top3) provided a statistically significant result for mismatch [$\chi^2(3) = 1460.6\$, \$p \ll 0.001$]; pairwise post-hoc Wilcoxon tests for paired samples with Holm-Bonferroni correction revealed statistical differences among all pairs of conditions ($p \ll 0.001$) except for the two best selection methods ($p = .69$). The same analysis

[4] Repeatability of HRTF measurements are still a delicate issue, suggesting a high variability in spectral details [20].

[5] Unfortunately, we were not able to directly compare our current study with [29] because different CIPIC subjects were considered.



(a) Mismatch (b) Rank

Fig. 5. Global statistics for metric assessment on (a) mismatch, (b) rank, grouped by HRTF condition. Asterisks and bars indicate, where present, a significant difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ at *post-hoc* test).

with (individual HRTF, KEMAR, best r, best top3) as levels of feedback condition provided a statistically significant result for rank [$\chi^2(3\$) = 3004.8$, $\$p \ll 0.001$] with statistical differences among all pairs of conditions ($p \ll 0.001$) except the two best selection methods (p = .22).

Expressing the best top3 selection in terms of mismatch and rank in the $N \times M$ estimates, no significant differences were found compared to the other best selection methods. However, one can identify the following trends by looking at data distributions: mismatch values of best top3 tended to be higher and more close to tendency of individual HRTFs, and rank distribution was even more compressed to first positions. Summarizing, the best top3 selection showed mismatch value more similar to individual values while keeping its rank more stable.

4 AUDITORY MODEL PREDICTIONS

Using the predictions of an auditory model, we simulated a virtual experiment where every CIPIC listener would be asked to provide an absolute localization judgment about spatialized auditory stimulus.

4.1 The Model

We adopted a recent model [17] that follows a *"template-based"* paradigm implementing a comparison between the internal representation of an incoming sound at the eardrum and a reference template. Spectral features from different HRTFs correlate with the direction of arrival, leading to a spectro-to-spatial mapping and a perceptual metric for elevation performances.

The model is based on two processing phases. During peripheral processing, an internal representation of the incoming sound is created and the *target* sound (e.g., a non-individual HRTF set) is separated into a directional independent transfer function, also knows as *common transfer function* (CTF) and a *directional transfer function* (DTF). According to [32], the CTF can be computed from the average magnitude spectrum among all HRTFs of a specific listener, thus it includes information regarding the pinna's omni-directional resonance and the transfer functions of the HRTF measurement setup. It is worthwhile to note that the simulation of headphone listening introduces an extra acoustic contribution, i.e., formally described by the individual headphone transfer function (HpTF), that could be approximately considered directional independent [33]. For each direction of arrival, the ratio between the corresponding HRTF and the listener's CTF results in the corresponding DTF.

In the second phase, the new representation is compared with a *template*, i.e., individual DTFs computed from individual HRTFs without considering any directional independent contributions, thus simulating the localization process of the auditory system (see previous works [14] for further details on this methodology). For each target angle, the probability that the virtual listener points to a specific angle defines the *similarity index* (SI). The index value results from the distance (in degrees) between the target angle and the response angle, which is the argument of a Gaussian distribution with zero-mean and standard deviation, called *uncertainty*, U. The lower the U value, the higher the sensitivity of the virtual listener in discriminating different spectral profiles resulting in a measure of probability rather than a deterministic value.

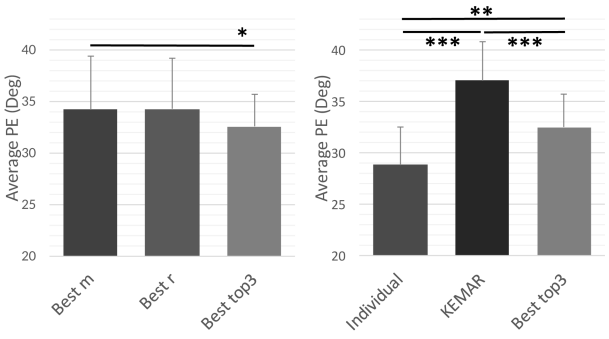
The virtual experiment was conducted simulating listeners with all analyzed CIPIC HRTFs, using an uncertainty value $U = 1.8$ that is similar to average human sensitivity [17]. We predicted elevation performance for every virtual subject when listening with their individual HRTFs, with those of CIPIC subject 165 (the KEMAR), and the best m / best r / best top3 selected HRTFs. The precision for the j-th elevation response close to the target position is defined in the *local polar RMS error* (*PE*):

$$PE_j = \sqrt{\frac{\sum_{i \in L} (\phi_i - \varphi_j)^2 p_j[\phi_i]}{\sum_{i \in L} p_j[\phi_i]}}, \tag{7}$$

where $L = \{i \in N\colon 1 \le i \le N_\phi,\ |\phi_i - \varphi_j| mod 180^\circ < 90^\circ\}$ defines local elevation responses within $\pm 90^\circ$ w.r.t. the local response ϕ_i and the target position φ_j, and $p_j[\phi_i]$ denotes the prediction, i.e., probability mass vector.

The average PE was computed considering only elevation responses φ_j between $[-45^\circ, +45^\circ]$, where inter-subject variability in human spatial hearing emerges [34], thus providing a single number that quantifies localization performance [14].[6] In order to verify statistically significant differences between predicted average PEs, paired t-tests were performed between pairs of localization performances using different HRTFs.

[6] We focused on local polar error in the frontal median plane, where individual elevation-dependent HRTF spectral features perceptually dominate; on the contrary, front-back confusion rate (similar to quadrant error rate *QE* in [17]) derives from several concurrent factors, such as dynamic auditory cues, visual information, familiarity with sound sources and training [35], thus it was not considered in this study.



(a) HRTF selection metrics (b) Individual vs. generic

Fig. 6. Global statistics (average + standard deviation) for prediction on average PE for (a) metrics based on notch distance mismatch, (b) individual vs. generic (KEMAR) vs. personalized (best top3). Asterisks and bars indicate, where present, a significant difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ at *post-hoc* test).

4.2 Results

Localization predictions from auditory model simulations provided average PEs with statistically significant differences between best m and best top3 metrics, $t(16) = 2.134$, $p < 0.05$ (see Fig. 6(a) for a graphical representation). These results suggest that best top3 yields better localization performances than best m and best r (although statistical significance was proven for best m only). Intuitively, the best top3 metric is more robust to contour uncertainty because of the M-constraint in its definition, that filters out variability due to HRTF sets with sporadic appearances in the top positions.

Finally, predictions were computed also for individual HRTFs and KEMAR virtual listening. Pairwise t-tests reveal significant differences in average PEs between individual and KEMAR ($t(16) = -7.79$, $p \ll 0.001$), and between individual and best top3 ($t(16) = -4.13$, $p < 0.01$), reporting a better performance with individual HRTFs. Moreover, pairwise t-test reports significant differences between best top3 and KEMAR ($t(16) = 5.590$, $p \ll 0.001$), with the former performing better than the latter.

5 LOCALIZATION TEST IN VIRTUAL REALITY

A localization test requires subjects to localize sounds coming from different directions. We evaluated localization performances of 15 normal-hearing subjects (13 males and 2 females, mean age 27 $\pm$ 4.7) with two HRTF sets (conditions). For each participant, the best personalized HRTF set was selected according to the top-M criterion with $M = 3$ following the same procedure of Sec. 3. The second HRTF set was that of a KEMAR, for all subjects. The aim of such an experiment was to further validate our HRTF selection procedure and to compare the overall results with the predictions of Sec. 4.

Ideally, the localization test should include a third condition, i.e., individual HRTF: this would provide a more solid ground truth for the evaluation. On the other hand, as discussed in Sec. 1.1, several sources of errors in measurements and rendering limit the applicability of individual binaural synthesis, which thus remains out of the scope of the present work. The comparison between the KEMAR HRTFs (which are a typical choice in real-world applications) and personalized HRTFs provides a viable approach to detect and assess improvements brought by our HRTF selection procedure. Other aspects, such as headphone equalization, were kept constant among listening conditions, in order to equally reduce coloration while remaining consistent with the reference auditory modeling of Sec. 4, which does not consider directional independent contribution (i.e., CTFs) in the processing.

Experimental sessions took place in a silent booth with a certified noise attenuation of 45 dB, placed inside a quiet laboratory, which ensured that no environmental sound was perceived by experimental subjects. Sessions had an overall average duration of 30 minutes per participant. We adopted a short localization test that was developed for screening purposes of [36] and for which a formal validation is currently included in a manuscript in preparation.

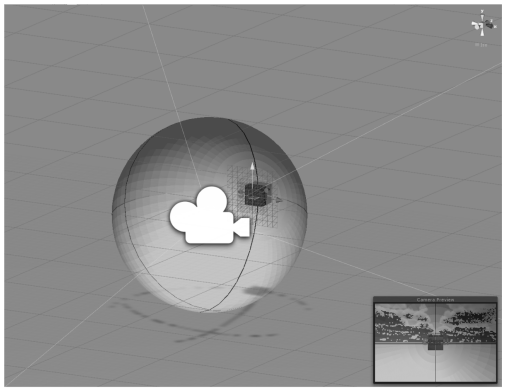
5.1 Apparatus

The experimental setup used a Samsung Galaxy S7 paired with the Samsung Gear VR for 3D video virtual reality rendering at 50 Hz. The Unity VR environment was used to simulate a minimalistic yet ecological outdoor scenario, which included a blue sky and a green grass garden (see Fig. 7(a) for a third-person view). Participants were placed inside a semi-transparent sphere with a 1 m radius, and were free to look around with unconstrained head movements. The sphere was also equipped with lines indicating the horizontal, median, and traversal planes in order to help orientation and the subsequent selection of response angles. Head orientation was visually rendered in the scene through a virtual laser pointer located at the center of the head, which produced a red dot on the 1 m sphere in front of the subject.

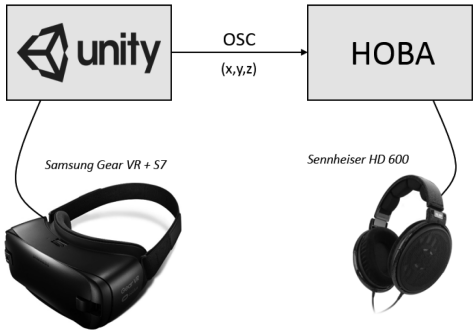
An external computer was used to run the audio environment: the computer received the coordinates of head movements from the Unity VR application via wireless communication through a router using the Open Sound Control (OSC) protocol[7]. The audio rendering was provided by the "HRTFs On-demand for Binaural Audio" (HOBA) web framework using Firefox and Node.js[8] (a schematic representation of the system can be found in Fig. 7(b)).

All sounds were played back through Sennheiser HD600 headphones. Equalization filters for these headphones were computed from their headphone impulse responses (HpIRs) measured over more than 100 human subjects at the Acoustic Research Institute of the Austrian Academy of Sciences;[9] data are available in SOFA format [37]. Such an equalization was needed to compensate for the average acoustic headphone contribution, thus reducing

[7] http://opensoundcontrol.org/
[8] https://github.com/hoba3d/
[9] http://sofacoustics.org/data/headphones/ari



(a)



(b)

Fig. 7. (a) The Unity experiment scene. (b) Block scheme of the application.

spectral coloration. The Equalizer APO software[10] was used to perform a low-latency convolution between the final spatial stimuli and the inverse average HpIR after being rendered in the HOBA framework. It has to be noted that the end-to-end audio latency of such system was 140 ms which did not impair a localization task of an auditory stimulus longer than 1 second [38, 39].[11]

5.2 Stimuli

Since the aim of the experiment was to validate the auditory model predictions discussed in Sec. 4, auditory stimuli were purely anechoic virtual sounds. The choice of a grass garden for the visual scenario, presented above, was meant to be consistent with this choice as it represents a reasonably anechoic environment.

The basic stimulus was composed by a train of three 40 ms gaussian noise bursts with 30 ms of silence between each burst, repeated six times (to a total duration of 3 s). Final stimuli were generated convolving noise bursts with CIPIC HRTFs corresponding to 24 directions resulting from the combinations of 6 azimuth values (target θ from -180° to 120° in 60° steps) and 4 elevation values (target ϕ from -28.125° to 56.250° in 28.125° steps). The inter-aural coordinate system [13] was used, which considers azimuth values $\theta \in [-180^\circ, 180^\circ]$ and elevation values $\phi \in [-90^\circ, 90^\circ]$.

The distance of sound sources was set to $r = 1$ m, which corresponds to the CIPIC measurement setup. The presentation level of the stimuli was 60 dBA.

For each experimental subject, two different HRTF sets were used to generate the stimuli: the KEMAR and the best top3 HRTF of the subject. In conclusion, two experimental blocks each made of 24 stimuli were generated for each experimental subject.

5.3 Procedure

The first step was to acquire pinna side pictures of experimental subjects in controlled conditions, in order to add them to the database of our HRTF selection tool and extract their anthropometric data as discussed in Sec. 2.2. Subjects were seated on a chair and briefed about the following procedure:

- Subjects had to center their head in a mirror placed in front of them, which had a straight vertical black line delimiting two halves;
- They had to close alternatively one eye and check with the other eye if the closed one was exactly on the vertical black line; this action was repeated several times until both eyes appeared on the black line, thus ensuring correct head orientation (perpendicular to the mirror);
- The Galaxy S7 was fixed to a tripod 50 cm away from the right ear and a measuring tape was placed near the pinna of the subject in order to allow subsequent pixel-to-meter conversion;
- The experimenter took a picture of the pinna with the Galaxy S7, which was then loaded into the database of the HRTF selection tool;
- Using the tool, the best top3 HRTF of the subject was determined.

The VR localization task for a single trial was structured as follows. Participants had to start the trial by looking straight ahead and pointing their head at a specific object in the virtual environment, namely a brown cube at position $\theta = \phi = 0$. After two seconds, a stimulus was played back. As soon as participants heard it, they had to rotate their head so as to point at the perceived stimulus direction. Finally, they had to stay still and tap on the Samsung Gear VR touch-pad to confirm the perceived stimulus direction.

Each subject had to complete two blocks of 24 stimuli (KEMAR and best top3, with no repetitions), with a 3 minute pause between blocks. The presentation order of the two blocks was alternated between subjects in order to compensate for learning effects. Within each block, the presentation order of stimuli was randomized.

5.4 Data Analysis

Data were processed to compute azimuth and elevation errors for each trial, focusing on average values and trends

[10] https://sourceforge.net/projects/equalizerapo/

[11] Network latency in the connection between the HOBA framework and the Samsung Galaxy S7 heavily influenced such measurements.

in order to allow a direct comparison with the predictions of the auditory model discussed in Sec. 4.2.

In order to account for front-back reversals, the azimuth error was defined as $E_\theta = \min\{E_{\theta,1}, E_{\theta,2}\}$, where

$$E_{\theta,1} = \begin{cases} |\theta - \hat{\theta}| - 360^\circ| \text{ if } |\theta - \hat{\theta}| > 180^\circ \\ |\theta - \hat{\theta}| \quad \text{otherwise} \end{cases}$$

$$E_{\theta,2} = \begin{cases} |\text{Cone}(\hat{\theta}) - \theta| - 360^\circ| \text{ if } |\text{Cone}(\hat{\theta}) - \theta| > 180^\circ \\ |\text{Cone}(\hat{\theta}) - \theta \quad \text{otherwise} \end{cases}$$

where θ is the stimulus direction of arrival, $\hat{\theta}$ is the perceived azimuth angle pointed by the subject, and Cone(θ) provides the mirrored azimuth value in the cone of confusion that has the same absolute angular difference as θ with respect to the origin of the front/back hemisphere.

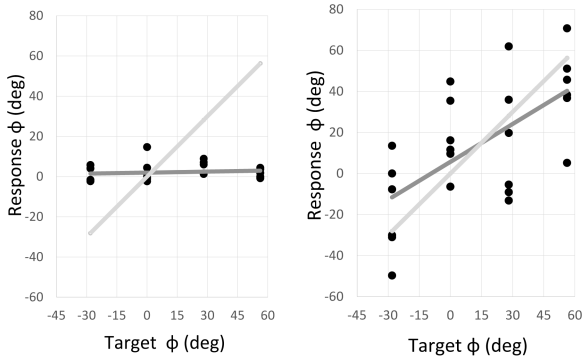
The elevation error E_ϕ was simply computed as the difference between the actual elevation angle of the stimulus, ϕ, and the perceived elevation angle, $\hat{\phi}$. For the sake of comparability, we defined also the average root mean square error following Eq. (7):

$$\text{Average } PE = \sqrt{\frac{\sum_i (\phi_i - \hat{\phi}_i)^2}{N}}, \tag{8}$$

where i spans all target angles considering elevation between -28.125° and 56.250°, leading to a $N = 24$.

It is known that performances in vertical localization vary remarkably among individuals more than horizontal localization [40]. In principle, one should find a reliable procedure for estimating the individual uncertainty of each participant in order to consider homogeneous groups of participants. However, obtaining such information is not trivial and requires an ad-hoc investigation with time- and resource-consuming localization experiments with loudspeakers or individual HRTFs that are beyond the scope of this paper. Since our previous analysis with auditory model simulations considers virtual subjects with an average localization uncertainty (the U parameter in the model), for a fair comparison we had to determine participants for which we could obtain unreliable responses.

Therefore, we proposed a practical post-screening criterion in order to identify participants who were not able to localize in elevation with any of the proposed HRTF sets. More precisely, a linear regression analysis was performed on ϕ and $\hat{\phi}$ of both HRTF sets separately. For each participant, such identification considered a threshold on the slope of the linear regression, with the following psycho-physical interpretation: a slope threshold value of .20 corresponds to a total angular variation of 9° along all the perceived elevations. This value is comparable to the localization blur for frontal sound events in the vertical plane [34].[12] This means that participants with slope values less than .20 were not able to perceive any change in sound source elevation, and consequently these subjects were not considered in our analysis. On the other hand, we could assume the formal possibility that exists at least a non-individual HRTF set able to provide elevation cues for every participant. However, this hypothesis is difficult to prove for an arbitrary listener in practice, leading to our main motivation in adopting auditory models for a systematic analysis [14]. Accordingly, the excluded subjects exhibited high uncertainties in their performances, which were not compatible with the choice of $U = 1.8$ in simulations from Sec. 4.



(a) KEMAR (b) Best top 3

Fig. 8. Elevation scatterplots of subject 15. Gray line: ideal response curve. Black line: linear fit of data.

A preliminary analysis on data distributions of E_θ and average PE data of valid subjects showed that normality assumption was not violated according to a Shapiro-Wilk test; thus, pairwise comparison on HRTF conditions was performed through t-tests for paired samples in order to asses differences in the overall localization performance between audio conditions. Such data analysis was performed on both the entire pool of participants and those with limited uncertainty.

5.5 Results

A summary of the localization data is presented in Table 1. The global statistics on average localization performances for all participants did not exhibit statistically significant differences between conditions in both azimuth as shown by paired t-test results on E_θ ($t(14) = 1.14$, $p = .28$), and elevation on PE ($t(14) = 1.47$, $p = .16$). In particular, Fig. 9(a) shows averages and standard deviations for PE grouped by conditions.

On the other hand, our criterion on perceived elevation led us to exclude six unreliable participants (IDs: 3, 5, 7, 9, 10, and 12) for whom the slope values in both KEMAR and best top3 selection are $\leq .2$. The remaining nine subjects were able to discriminate the vertical dimension of sound with different sensitivities depending on the conditions. Subjects 2, 6, 15 clearly perceived elevation with personalized HRTFs only (see Fig. 8 for an example of localization performance); subjects 1, 4, and 13 were able to localize in both conditions; subjects 8, 11, and 14 exhibited a weak variation with elevation suggesting their ability in localizing sound in the vertical dimension despite a non-optimal HRTF set for them. In particular, subjects

[12] The localization blur identifies the average margin of angular error in the human auditory system, and it can be expressed in terms of minimum audible angle (MAA).

Table 1. The mean-values, standard deviations for azimuth and elevation errors in degrees, together with slope-values, intercepts, and statistical significance on the linear regression obtained during the localization test. Bold IDs identify valid elevation localizers. Asterisks on r^2 indicate, where present, a significant relation between the predictor and elevation data (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ at *post-hoc* test).

	KEMAR					Personalized - best top3				
ID	E_θ	E_ϕ	*Sl.*	*Int.*	r^2 (p)	E_θ	E_ϕ	*Sl.*	*Int.*	r^2 (p)
1	7.17, ± 17.62	19.85, ± 19.49	.63	16.53	.45***	7.61, ± 7.23	19.96, ± 15.02	.49	17.59	.49***
2	11.02, ± 10.63	22.58, ± 22.96	.24	8.66	.10	9.14, ± 7.98	19.95, ± 14.52	.42	10.14	.36**
3	84.16, ± 56.05	33.08, ± 19.05	.21	15.46	.056	21.88, ± 20.91	41.39, ± 28.57	.27	48.67	.12
4	8.86, ± 6.66	34.98, ± 24.16	.36	41.72	.34**	5.76, ± 6.16	25.25, ± 22.89	.45	23.01	.25*
5	29.41, ± 22.53	28.69, ± 21.36	.00	– .43	.00	23.19, ± 20.78	36.54, ± 27.02	.12	–14.69	.04
6	16.53, ± 13.00	27.33, ± 25.88	0.33	7.09	.09	18.80, ± 21.19	18.71, ± 16.33	.56	16.60	.50***
7	14.71, ± 14.81	25.43, ± 24.14	0.18	–7.48	.17*	21.90, ± 15.70	33.97, ± 25.80	.02	–7.77	.00
8	16.69, ± 14.84	38.57, ± 27.50	.02	50.11	.00	19.83, ± 15.92	30.81, ± 24.34	.21	34.40	.01
9	16.07, ± 14.16	30.91, ± 21.78	.07	23.05	.01	12.17, ± 11.64	31.64, ± 19.92	.09	30.31	.02
10	25.76, ± 20.68	42.24, ± 31.73	–.09	14.48	.00	16.26, ± 14.09	34.95, ± 26.52	–.15	4.28	.04
11	12.80, ± 13.08	33.28, ± 29.61	.27	–.29	.05	12.80, ± 12.58	34.43, ± 27.15	.14	11.62	.01
12	7.97, ± 5.48	29.76, ± 21.17	.01	–5.25	.02	9.14, ± 9.62	26.29, ± 18.29	.12	2.77	.11
13	11.74, ± 14.67	25.17, ± 18.69	.33	31.70	.51***	10.19, ± 6.72	26.31, ± 16.02	.34	27.37	.43***
14	14.65, ± 15.79	42.00, ± 30.64	.25	40.13	.05	19.11, ± 13.36	31.60, ± 26.09	.17	34.36	.06
15	15.22, ± 17.77	27.78, ± 18.85	.02	1.98	.02	12.71, ± 12.25	21.73, ± 14.80	.61	5.69	.41***

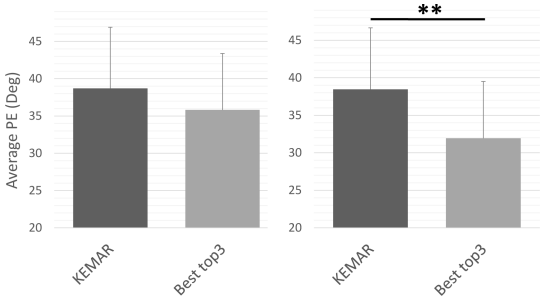


Fig. 9. Global statistics (average + standard deviation) of localization performances in elevation on average PE for generic (KEMAR) vs. personalized (best top3) which were computed for (a) all participants, and (b) participants with limited uncertainty. Asterisks and bars indicate, where present, a significant difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ at post-hoc test).

8 and 14 probably experienced an "elevation bias" identified by a high value of the intercepts, i.e., $\geq 40^\circ$, due to the systematically higher location of the spectral features in non-individual HRTFs compared to individual HRTFs [41].

In this homogeneous pool of participants, average localization performances in the horizontal plane did not exhibit statistically significant differences between conditions, as shown by t-test results on E_θ (t(8) = .16, p = .87). With regard to average *PE*, the paired t-test between KEMAR and best top3 conditions reported a statistically significant difference (t(8) = 4.51, p <0.01). Fig. 9(b) shows averages and standard deviations grouped by conditions, which are in good agreement with those provided by the predictions of the auditory model (see Fig. 6). Average values have a difference of $\pm 1^\circ$ and standard deviations of $\pm 4^\circ$ denoting less variability for predicted data due to difference in the available subject pool.

6 GENERAL DISCUSSION

The proposed distance metrics based on the C_1 contour provides insufficient features in order to unambiguously identify an individual HRTF from the corresponding side picture of the listener. Moreover, multiple tracing of C_1 and of the focus point adds further variability to the procedure resulting in extra uncertainty for the validation.

In particular, the mismatch analysis reported in Sec. 3.1 provided counter-intuitive results: the procedure tried to select the individual HRTFs from pinna contours, yet it always selected a generic HRTF that differed from the individual one in terms of both mismatch and rank. However, this evidence can be interpreted in light of previous studies [13], which showed that the notch associated to the pinna helix border is not able to describe elevation cues for all listeners. Moreover, biometric recognition studies [42] show that the pinna concha wall is also relevant in order to uniquely identify a person. Finally, multiple contours tracing highly contributes to the uncertainty in the matching of notch frequencies. As a result, the rank of individual HRTFs is reasonably good on average (the median is 8, as shown in Fig. 5(b), while the mean is 11.61), although individual HRTFs are generally not chosen as best selection. Despite the limitations of the proposed distance metrics, the predictions in localization performance provided by the auditory model (see Sec. 4) suggest that the personalized HRTFs selected by our procedure outperform dummy-head HRTFs. Fig. 6(b), in particular, shows that the local polar RMS error achieved with the best top3 personalized HRTF is significantly lower than the one achieved with the generic KEMAR HRTF, although it is still significantly higher than the one achieved with individual HRTFs.

The results provided by the auditory model are further corroborated by the outcomes of the localization test with 15 subjects, reported in Sec. 5. These outcomes provide the most significant original contribution of this paper. The

striking similarity between the performance predicted by the auditory model and the actual performances observed on experimental subjects with limited uncertainty confirms the validity of our approach: auditory models of spatial sound perception can be effectively used to quantify the perceptual similarity of HRTFs, and thus to assess the goodness of personalized HRTF selection procedures. On the other hand, the localization performances achieved by those participants confirm that personalized HRTFs selected with our distance metrics result in significantly lower polar errors than generic KEMAR HRTFs: this evidence is summarized in Fig. 9(b).

The take-home message of these results is that the distance metrics of Eq. (6) and the manual procedure described in Sec. 2 provide a practical approach to HRTF personalization that outperforms generic HRTFs. Despite the limitations discussed above, it has the advantage of requiring a minimal amount of subject information (a single 2D picture) and of using anthropometric data (the C_1 contour) that can be easily extracted even by a non-expert operator.

Further research is still needed in order to increase the applicability of our notch distance metrics; CIPIC subjects can be also analyzed by applying Eqs. (2) and (3) (notches caused by positive reflections) to Eq. (6), and localization predictions with both reflection signs can be compared. Contours associated to antihelix and concha reflections can be traced, and the mismatch definition can be modified accordingly by combining the contributions of each contour with different weights [13]. Furthermore, notch distance metrics, i.e., mismatch, rank, and top-M metrics, can be hierarchically applied in the HRTF selection process in order to refine the selection: as an example, starting from the top M metric one can disambiguate similar HRTF sets looking at mismatch and rank metrics. In particular, the influence of the M parameter on HRTF appearance in the rank metric has to be investigated in more detail.

It has to be noted that our methodology has a strong dependence on the chosen auditory model that established our ground truth within our research framework. The proposed listening evaluation is a key element for obtaining meaningful correspondences between simulations and reality, thus guiding the design of our experiments. In particular, the adopted auditory model provided its predictions based on DTF processing, not considering any non directional contributions such as CTFs and those from headphones. Accordingly, a generic headphone compensation was preferred to individual equalization, also providing direct applicability of our findings to typical setups for spatial audio rendering [33, 43].

It is worth emphasizing that the mismatch between listeners' individual CTFs (template) and CTFs from the selected HRTFs (target), and between individual and generic HpTFs are ignored in our predictions. From a methodological point of view, nothing prevents to replicate our study with a different auditory model allowing to take into account spectral variations due to both individual/generic headphone equalization and positioning [44] with their related impacts on localization, such as the recent model based on Bayesian framework [45].

Finally, it is indisputable that experimental validation with massive participation of human subjects will be highly relevant in terms of reliability of any HRTF selection procedure. A new research framework for binaural audio reproduction in web browsers is currently in development [46] with the goal of overcoming common limitations in HRTF personalization studies, such as low number of participants (e.g., [24]), coherence in simplifications of localization experiment (e.g., [27]), and reliability of the predictions with computational auditory models.

7 CONCLUSIONS

Our final result confirms that our image-guide HRTF selection procedure provides a useful tool in terms of:

- **Personalized dataset reduction**: Since individual HRTF rank is on average the 12*th* position, one can compute a personalized short list of $\approx$12 best candidate HRTFs for a given pinna picture in which finding with high probability a generic HRTF reasonably close to the individual one. Accordingly, a subsequent refinement of the HRTF selection procedure might be required through subjective selection procedures or additional data analysis on the reduced dataset.
- **Better performance than KEMAR**: Confirming our previous findings in psychoacoustic evaluation [13], auditory model predictions reported a statistically significant improvement in localization performance with generic HRTFs selected based on top3 metric compared to KEMAR; this result has important practical implications for binaural audio applications requiring HRTF personalization: our tool allows a user without specific expertise to choose a generic HRTF in a few minutes; this selection outperforms localization performance with KEMAR HRTFs, which are usually default solutions for commercial applications in VADs.

A fully automated contour-tracing procedure would be preferable with respect to the current manual approach. In past works, however, the use of computer vision techniques for the extraction of the main contours provided mixed results [47]. This is mainly due to (i) the insufficient information provided by a 2D picture (e.g., no depth information), and (ii) the inherent approximation in the HRTF mismatch function of Eq. (6), which uses a simple geometric law to estimate notches from contours. An attractive option for future works is to use the most effective contour-tracing strategies devised by expert operators with the current manual tool to train a neural network, which would learn such strategies and apply them on new pinna pictures.

An alternative approach, which is currently being investigated, amounts to estimating the first pinna notch directly via acoustic measurements, through a so-called "acoustic selfie" that roughly acquires individual HRTFs using a smartphone loudspeaker as sound source and binaural microphones as receivers [48]. In this way, the

frequencies $f_0^{(k,n)}(\varphi)$ could be directly computed in the acoustic domain, further reducing manual intervention.

8 ACKNOWLEDGMENTS

This study was supported by the 2016–2021 strategic program "Knowledge for the World" of Aalborg University with a grant awarded to Michele Geronazzo.

9 REFERENCES

[1] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, 1st ed. (Springer Publishing Company, (2007).

[2] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display* (J. Ross Publishing, Plantatation, FL, 2013).

[3] J. Blauert (ed.) *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing Series (Springer Berlin Heidelberg, 2013).

[4] M. Cohen and J. Villegas, "Applications of Audio Augmented Reality: Wearware, Everyware, Anyware, and Awareware," in W. Barfield (ed.), *Fundamentals of Wearable Computers and Augmented Reality, Second Edition*, pp. 309–330 (CRC Press, 2015).

[5] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A Spatial Audio Quality Inventory (SAQI)," *Acta Acust. united Ac.*, vol. 100, no. 5, pp. 984–994 (2014 Sep.).

[6] 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), doi:10.1121/1.399963.

[7] 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).

[8] S. Prepeliţă, M. Geronazzo, F. Avanzini, and L. Savioja, "Influence of Voxelization on Finite Difference Time Domain Simulations of Head-Related Transfer Functions," *J. Acoust. Soc. Amer.*, vol. 139, no. 5, pp. 2489–2504 (2016 May), doi:10.1121/1.4947546.

[9] E. A. Macpherson, and A. T. Sabin, "Binaural Weighting of Monaural Spectral Cues for Sound Localization," *J. Acoust. Soc. Amer.*, vol. 121, no. 6, pp. 3677–3688 (2007 Jun.), doi:10.1121/1.2722048.

[10] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation Localization and Head-Related Transfer Function Analysis at Low Frequencies," *J. Acoust. Soc. Amer.*, vol. 109, no. 3, pp. 1110–1122 (2001), doi:10.1121/1.1349185.

[11] C. Kim, R. Mason, and T. Brookes, "Head Movements Made by Listeners in Experimental and Real-Life Listening Activities," *J. Audio Eng. Soc.*, vol. 61, pp. 425–438 (2013 Jun.).

[12] F. Brinkmann, A. Lindau, S. Weinzerl, S. van de Par, M. Müller-Trapet, R. Opdam, and M. Vorländer, "A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations," *J. Audio Eng. Soc.*, vol. 65, pp. 841–848 (2017 Oct.), doi:10.17743/jaes.2017.0033.

[13] M. Geronazzo, S. Spagnol, A. Bedin, and F. Avanzini, "Enhancing Vertical Localization with Image-Guided Selection of Non-individual Head-Related Transfer Functions," *IEEE Int. Conf. on Acoust. Speech Signal Process. (ICASSP 2014)*, pp. 4496–4500 (2014 May), doi:10.1109/ICASSP.2014.6854446.

[14] M. Geronazzo, S. Spagnol, and F. Avanzini, "Do We Need Individual Head-Related Transfer Functions for Vertical Localization? The Case Study of a Spectral Notch Distance Metric," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 7, pp. 1243–1256 (2018 Jul.), doi:10.1109/TASLP.2018.2821846.

[15] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini, "Improving Elevation Perception with a Tool for Image-Guided Head-Related Transfer Function Selection," *Proc. of the 20th Int. Conf. Digital Audio Effects (DAFx-17)*, pp. 397–404 (2017 Sep.).

[16] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF Database," *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, pp. 1–4 (2001 Oct.).

[17] R. Baumgartner, P. Majdak, and B. Laback, "Assessment of Sagittal-Plane Sound Localization Performance in Spatial-Audio Applications," in J. Blauert (ed.), *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing Series, pp. 93–119 (Springer Berlin Heidelberg, 2013).

[18] M. Geronazzo, S. Spagnol, and F. Avanzini, "Mixed Structural Modeling of Head-Related Transfer Functions for Customized Binaural Audio Delivery," *Proc. 18th Int. Conf. Digital Signal Process. (DSP 2013)*, pp. 1–8 (2013 Jul.).

[19] C. Sforza, G. Grandi, M. Binelli, D. G. Tommasi, R. Rosati, and V. F. Ferrario, "Age- and Sex-Related Changes in the Normal Human Ear," *Forensic Sci. Int.*, vol. 187, no. 1–3, pp. 110.e1–110.e7 (2009 May).

[20] A. Andreopoulou, D. Begault, and B. Katz, "Inter-Laboratory Round Robin HRTF Measurement Comparison," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 895–906 (2015 Aug.), doi:10.1109/JSTSP.2015.2400417.

[21] R. Bomhardt, I. C. P. Mejía, A. Zell, and J. Fels, "Required Measurement Accuracy of Head Dimensions for Modeling the Interaural Time Difference," *J. Audio Eng. Soc.*, vol. 66, pp. 114–126 (2018 Mar.), doi:10.17743/jaes.2018.0005.

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

[23] C. Guezenoc and R. Seguier, "HRTF Individualization: A Survey," presented at the *145th Convention of the Audio Engineering Society* (2018 Oct.), convention paper 10129.

[24] 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 (2014 Jul.).

[25] 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).

[26] I. Nambu, M. Washizu, S. Morioka, Y. Hasegawa, W. Sakuma, S. Yano, H. Hokari, and Y. Wada, "Reinforcement-Learning-Based Personalization of Head-Related Transfer Functions," *J. Audio Eng. Soc.*, vol. 66, pp. 317–328 (2018 May), doi:10.17743/jaes.2018.0014.

[27] 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 (2012 Feb.).

[28] 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 (2005 Jul.).

[29] 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 (2013 Mar.), doi:10.1109/TASL.2012.2227730.

[30] B. Zonooz, E. Arani, K. P. Körding, P. A. T. R. Aalbers, T. Celikel, and A. J. V. Opstal, "Spectral Weighting Underlies Perceived Sound Elevation," *Scientific Reports*, vol. 9, no. 1, p. 1642 (2019 Feb.), doi:10.1038/s41598-018-37537-z.

[31] M. Geronazzo, S. Spagnol, and F. Avanzini, "Estimation and Modeling of Pinna-Related Transfer Functions," *Proc. of the 13th Int. Conf. Digital Audio Effects (DAFx-10)*, pp. 431–438 (2010 Sep.).

[32] J. C. Middlebrooks and D. M. Green, "Directional Dependence of Interaural Envelope Delays," *J. Acoust. Soc. Amer.*, vol. 87, no. 5, pp. 2149–2162 (1990 May), doi:10.1121/1.399183.

[33] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Srensen, "Transfer Characteristics of Headphones Measured on Human Ears," *J. Audio Eng. Soc.*, vol. 43, pp. 203–217 (1995 Apr.).

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

[35] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization Using Nonindividualized Head-RelatedTtransfer Functions," *J. Acoust. Soc. Amer.*, vol. 94, no. 1, pp. 111–123 (1993).

[36] M. Geronazzo, E. Sikström, J. Kleimola, F. Avanzini, A. De Götzen, and S. Serafin, "The Impact of a Good Vertical Localization with HRTFs in Short Immersive Virtual Reality Explorations," *Proc. 17th IEEE Int. Symp. on Mixed and Augmented Reality (ISMAR)*, pp. 1–8 (2018 Oct.).

[37] B. B. Boren, M. Geronazzo, P. Majdak, and E. Choueiri, "PHOnA: A Public Dataset of Measured Headphone Transfer Functions," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), convention paper 9126.

[38] D. S. Brungart, B. D. Simpson, R. L. McKinley, A. J. Kordik, R. C. Dallman, and D. A. Ovenshire, "The Interaction between Head-Tracker Latency, Source Duration, and Response Time in the Localization of Virtual Sound Sources," in *Proc. Int. Conf. Auditory Display 2004* (2004 Jul.).

[39] S. Yairi, Y. Iwaya, and Y. Suzuki, "Influence of Large System Latency of Virtual Auditory Display on Behavior of Head Movement in Sound Localization Task," *Acta Acust. united Ac.*, vol. 94, no. 6, pp. 1016–1023 (2008 Nov.), doi:10.3813/AAA.918117.

[40] P. Majdak, R. Baumgartner, and B. Laback, "Acoustic and Non-Acoustic Factors in Modeling Listener-Specific Performance of Sagittal-Plane Sound Localization," *Front Psychol.*, vol. 5, pp. 1–10 (2014 Apr.), doi:10.3389/fpsyg.2014.00319.

[41] 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).

[42] E. González, L. Alvarez, and L. Mazorra, "Normalization and Feature Extraction on Ear Images," *Proc. IEEE 46th Int. Carnahan Conf. Security Tech.*, pp. 97–104 (2012 Oct.).

[43] H. Møller, M. Sørensen, J. Friis, B. Clemen, and D. Hammershøi, "Binaural Technique: Do We Need Individual Recordings?" *J. Audio Eng. Soc.*, vol. 44, pp. 451–469 (1996 Jun.).

[44] M. Paquier and V. Koehl, "Discriminability of the Placement of Supra-Aural and Circumaural Headphones," *Applied Acoustics*, vol. 93, pp. 130–139 (2015 Jun.), doi:10.1016/j.apacoust.2015.01.023.

[45] R. Ege, A. J. V. Opstal, and M. M. V. Wanrooij, "Accuracy-Precision Trade-off in Human Sound Localisation," *Scientific Reports*, vol. 8, no. 1, p. 16399 (2018 Nov.), doi:10.1038/s41598-018-34512-6.

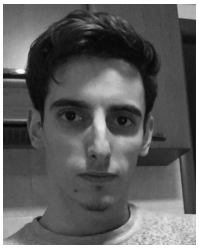
[46] M. Geronazzo, J. Kleimola, and P. Majdak, "Personalization Support for Binaural Headphone Reproduction in Web Browsers," *Proc. 1st Web Audio Conf.* (2015 Jan.).

[47] S. Spagnol, M. Geronazzo, D. Rocchesso, and F. Avanzini, "Synthetic Individual Binaural Audio Delivery by Pinna Image Processing," *Int. J. of Pervasive Computing and Communications*, vol. 10, no. 3, pp. 239–254 (2014), doi:10.1108/IJPCC-06-2014-0035.

[48] M. Geronazzo, J. Fantin, G. Sorato, G. Baldovino, and F. Avanzini, "Acoustic Selfies for Extraction of External Ear Features in Mobile Audio Augmented Reality," *Proc. 22nd ACM Symp. on Virtual Reality Software and Technology (VRST 2016)*, pp. 23–26 (2016 Nov.), doi:10.1145/2993369.2993376.

THE AUTHORS


Michele Geronazzo


Enrico Peruch


Fabio Prandoni


Federico Avanzini

Michele Geronazzo received his M.S. degree in computer engineering and his Ph.D. degree in information and communication technology from the University of Padova, in 2009 and 2014, respectively. Between 2014 and 2017, he worked as a postdoctoral researcher at University of Padova and University of Verona in the fields of ICT and neurosciences. He is currently postdoctoral researcher at Aalborg University Copenhagen, where he is with the "Multisensory Experience Lab" developing his research project "Acoustically-trained 3D audio models for virtual reality applications" (main topics: virtual acoustics, headphones, and binaural hearing). His main research interests involve 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 (AIA) with the "G. Sarcedote" award for best Ph.D. thesis in acoustics. He is also a member of the organizing committee of the IEEE VR Workshop on Sonic Interactions for Virtual Environments since 2015 (chair of the 2018 edition). Dr. Geronazzo served as guest editor for *Wireless Communications and Mobile Computing* (John Wiley & Sons and Hindawi publishers, 2018). He is a co-recipient of four best paper/poster awards and co-author of more than fifty scientific publications.

•

Enrico Peruch took his master degree in computer engineering in July 2017 from the University of Padua. His interest in VR environment and 3D audio application derives from his video game career, winning twice, the Italian championship. He develop a master thesis on validating custom HRTF selection methods, based on multiple pinna contours powered by Matlab, with an auditory model and in a VR Unity environment in collaboration with Fabio. Now he works at an IT consulting company as a web and application developer for an insurance company.

•

From an early age, Fabio Prandoni was attracted to math and music. Guitarist and piano player, then orchestral and trailer composer, he took his master degree in computer engineering from the University of Padova in April 2017, having completed his master thesis about a Unity virtual reality environment and personalized audio powering a Matlab application based on pinna contours. Now he works in a web agency as R&D supervisor and web developer.

•

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