IMPROVING ELEVATION PERCEPTION WITH A TOOL FOR IMAGE-GUIDED HEAD-RELATED TRANSFER FUNCTION SELECTION

Michele Geronazzo

Dept. of Neurological, Biomedical and Movement Sciences,
University of Verona
Verona, Italy
michele.geronazzo@univr.it

Enrico Peruch, Fabio Prandoni, and Federico Avanzini

Dept. of Information Engineering
University of Padova
Padova, Italy
avanzini@dei.unipd.it

ABSTRACT

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 subject's pinna, the procedure selects from a database the HRTF set that best fits the anthropometry of that subject. 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 an auditory model for sound localization in the mid-sagittal plane available from previous literature. Using virtual subjects from a HRTF database, a virtual experiment is implemented to assess the vertical localization performance of the database subjects when they are provided with HRTF sets selected by the proposed procedure. Results report a statistically significant improvement in predictions of localization performance for selected HRTFs compared to KEMAR HRTF which is a commercial standard in many binaural audio solutions; moreover, the proposed analysis provides useful indications to refine the perceptually-motivated metrics that guides the selection.

1. 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 listener himself. Knowledge of such a complex process is needed in order to develop accurate and realistic artificial sound spatialization in several application domains, including music listening, entertainment (e.g. gaming), immersive virtual reality, sensory substitution devices (e.g. for visually-impaired users), teleoperation, tele-conferencing, and so on [1].

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 Transfers 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) [2]. This paper deals in particular with elevation localization cues, which are mainly provided by monaural spectral features of the HRTF. Specifically, the scattering of acoustic waves in the proximity of the pinna creates a complex and individual topography of pressure nodes which is not completely understood [3, 4], 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 constants with varying elevations.

Individual anthropometric features of the human body have a key role in shaping individual HRTFs (see the discussion in Sec. 2 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 [5]. 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 [5] 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 psychoacustic 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. The proposed HRTF selection procedure is here validated on subjects from the CIPIC subjects [6] for whom 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. Performances in elevation perception are evaluated by means of an auditory model for sound localization in the mid-sagittal plane [7] (i.e., the vertical plane dividing the listener's head in left and right halves) provided by the Auditory Modeling Toolbox[1]. Using virtual subjects from the CIPIC database, we present a virtual experiment that assesses the vertical localization performance of CIPIC subjects when they are provided with HRTF sets selected by the proposed procedure.

2. RELATED WORKS

One of the main limitations of binaural audio technologies for commercial use is the hard work behind the creation of the in-

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

dividual HRTFs that capture all of the physical effects creating a personal perception of immersive audio. The measurement of a listener's own HRTFs in all directions requires a special measuring apparatus and a long measurement time, often a too heavy task to perform for each subject involved in every-day application. That is the main reason why alternative ways are preferred to provide HRTFs giving to listeners a personalized, but not individually created, HRTF set: a trade off between quality and cost of the acoustic data for audio rendering [8, 9].

2.1. Individual / Own HRTFs

The standard setup for individual HRTF measurement is in 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. The listener, seated in the center of the sphere, has microphones in his/her ears. After subject preparation, HRIRs are measured playing analytic signals and recording responses collected at the ears for each loudspeaker position in space (see [9] for a systematic review on this topic).

The main goal is to extract the set of HRTFs for every listener thus providing him/her the individual/own transfer function. In addition to the above mentioned high demanding requirements (time and equipment), there are some other critical aspects in HRTF measurements; listener's pose is usually limited to a few positions (standing or sitting), relatively few specific locations around his/her body, and own intrinsic characterization without considering that pinna shape is one of the human body part that always grows during lifetime [10]. Moreover, repeatability of HRTF measurements are still a delicate issue [11].

2.2. Personalized / Generalized HRTFs

The personalized HRTFs are chosen among available HRTFs of a dataset instead of doing individual measurements. This procedure is based on a match between external subjects (the one without individual HRTFs) and internal, i.e. belonging to a database, subjects with already stored information (both acoustics and anthropometry) . The most interesting and important part, is the method of how is selected a specific set of HRTFs to an external subject. Researchers are finding different ways to deal with this issue and there are a variety of alternatives using common hardware and/or software tools. The main benefit of this approach is that a user can be guided to a self selection of their best HRTF set without needing a special equipment or knowledge. It has to be noted that the personalized HRTF can not guarantee the same performance as their own HRTF but they usually provide better performance than the generic dummy-head HRTFs such those of Knowles Electronic Manikin for Acoustic Research (KEMAR) [12].

In the following, we summarize three main approaches to HRTF selection.

- **DOMISO**[13][2]
  In this technique, subjects can choose their most suitable HRTFs from among many, taken from a database following tournament-style listening tests. The database (corpus) is built using different subjects, storing 120 sets of HRTFs, one set per listener.
  Performances of this technique were evaluated by Yukio Iwaya that proved that the personalized DOMISO HRTFs results were similar to individualized HRTFs ones but very different from the away condition (totally random HRTF, that could not win the tournament).
- **Two steps selection**[14, 15]
  This is a technique based on two different steps. Usually the first step selects one subset from a complete initial pool of HRTF sets, removing worse HRTFs from a perceptual point of view. The second step refines the selection in order to obtain the best match among generic HRTFs of a dataset which is reduced in size compared to the complete database.
- **Matching anthropometric ear parameters**[16, 17]
  This method is based on finding the best match HRTF in the anthropometric domain, matching the external ear shape of a subject using anthropometric measurements available in the database.[3]

[2] DOMISO: Determination method of OptimuM Impulse-response by Sound Orientation

3. IMAGE-GUIDED HRTF SELECTION

Another approach to HRTF selection problem consists in mapping anthropometric features into the HRTF domain, following a ray-tracing modeling of pinna acoustics [18, 19]. The main idea is to draw pinna contours on a 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 [5].

3.1. 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 π:

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

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 [19] on the CIPIC database [6] proved that almost 80% of the subjects in the database exhibit a clear negative reflection in their HRIRs. Under this assumption, notches are

[3] see section 3.2 for further details.

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 it has been shown [5] that the first and most prominent notch in the HRTF is typically associated to the most external pinna contour on the helix border (the "C_1" contour hereafter).

Now, assume that N estimates of the C_1 contour and K estimates of the focus point have been traces on a 2D picture of the pinna of a subject (the meaning of N and K is explained later in Sec. 3.2). We define the basic notch distance metric in the form of a mismatch function between the corresponding notch frequencies, and the notch frequencies of a HRTF:

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

where $f_0^{(k,n)}(\varphi) = c/[2d_c^{(k,n)}(\varphi)]$ are the frequencies extracted from the image and contours of the subject, and F_0 are the notch frequencies extracted from the HRTF with *ad-hoc* algorithms such as those developed in [18, 20, 17]; (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. Extracted notches need to be grouped into a single track evolving through elevation consistently, a labeling algorithm (e.g. [19]) performed such computation along subsequent HRTFs.

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 three 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).
- **Top-M appearance:** for a given integer M, for each HRTF, a similarity score is assigned according to the number of times (for all the (k, n) pairs) in which that HRTF ranks in the first M positions.

3.2. A HRTF selection tool

Based on the concepts outlined above, we propose a tool for selecting from a database a 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.

Graphical user interface. Figure 1 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 efficiently using the three 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.

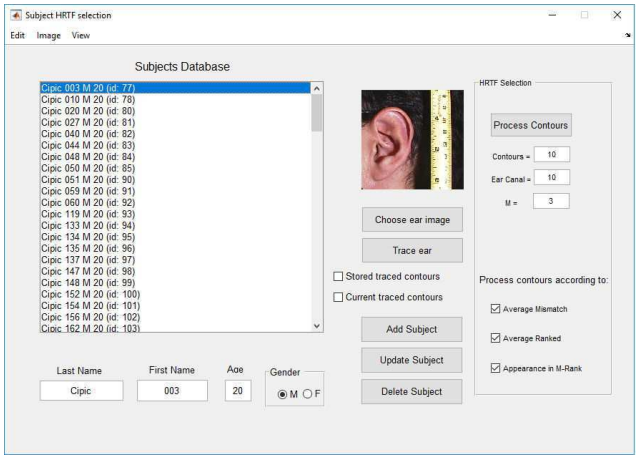


Figure 1: A tool for HRTF selection: main Graphical User Interface.

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.

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 pinna image the contours drawn in the current session. [4]

One last parameter to be set, M, refers to the top-M appearance metrics discussed above. By clicking on the "*Process Contours*", the application returns the ranked positions of the database HRTFs according to the three metrics.

Database of generic HRTFs and extracted F_0. The public database used for our purpose is the CIPIC [6]. The first release provided HRTFs for 43 subjects (plus two 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 the first prominent notch in each HRTF were extracted with the *structural decomposition algorithm* [20, 9] and F_0 tracks were labeled with the algorithm in [19] and then stored in a custom data structure;

Guidelines for contour tracing. In the trace-ear 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 point K positions of the

[4] 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.

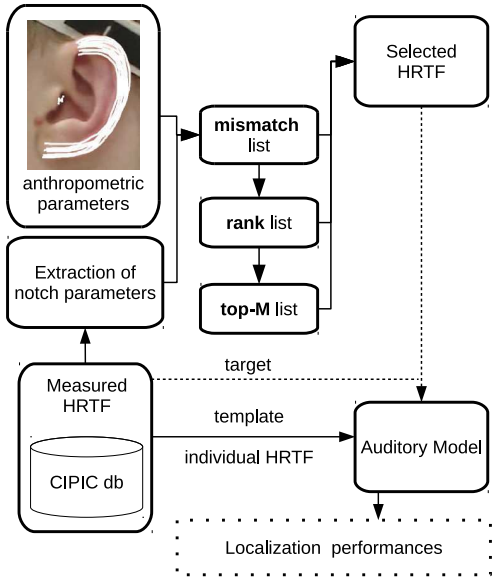


Figure 2: Schematic view of the proposed validation procedure with auditory model predictions.

ear canal entrance. 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 which will be very 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 [19] 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 with a physical human ear allow the operator to easily localize where the ear canal is, reducing also uncertainty for the estimation of external pinna contours.

4. VALIDATION

The main aim of the proposed validation procedure is to verify the effectiveness of our HRTF selection tool in providing a subject with a HRTF that is reasonably close to his/her individual HRTF, by only using a picture of his/her external ear. Strengths and limits of such an approach are discussed also with the support of an auditory model to predict performance in elevation perception. Figure 2 depicts a schematic view of the entire validation process.

4.1. Data acquisition and 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 5 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 $N \times M$ estimates), in increasing order of mismatch.
- **Average rank:** CIPIC subjects are sorted according to their rank in the previous ranking (averaged over the $N \times M$ estimates), in increasing order of rank.
- **Top-M appearance:** CIPIC subjects are sorted according to the number of their occurrences of the in the top-3 positions for each (n, k) pair of estimates, in decreasing order of occurrence count.

For each metrics, we defined three *best fitting HRTFs* by choosing the HRTFs ranking first in each ranking: 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 significances in performance between conditions.

4.2. Auditory model simulations

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. We adopted a recent model [7], 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 generic HRTF set) is converted into a *directional transfer function* (DTF). In the second phase, the new representation is compared with a *template*, i.e. individual DTFs computed from individual HRTFs, thus simulating the localization process of the auditory system (see previous works [21] 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 $U = 1.8$ which is similar to average human sensitivity [7]. We predicted elevation performance for every virtual subject when listening with his/her own 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):

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

where $L = \{i \in N : 1 \leq i \leq N_\phi, |\phi_i - \varphi_j| \, mod 180^\circ < 90^\circ\}$ 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 [22], thus providing a single number that quantifies localization performance [21].[5] In order to verify statistically significant differences between predicted average PEs, paired t-tests were performed between pairs of localization performances using different HRTFs.

5. RESULTS AND DISCUSSION

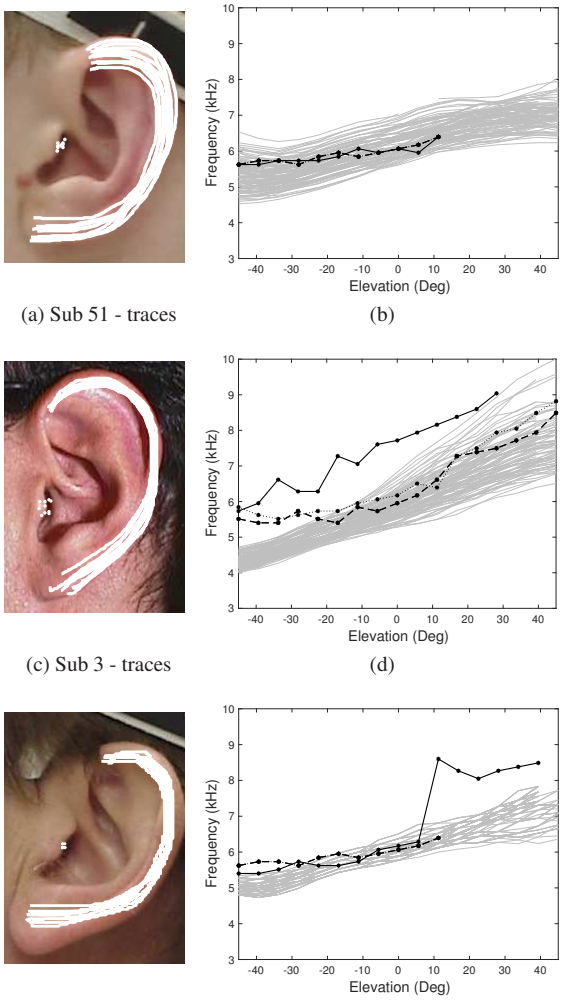
A preliminary analysis on data distribution of rank values derived from mismatches between $f_0^{(k,n)}(\varphi)$ and individual HRTF's $F_0(\varphi)$ (22×100 observations) was conducted in order to identify the existence of outliers for our metrics. Samples in the last quartile of this distribution 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 which can not 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.

Figure 3 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 2 of the 5 subjects which were previously removed). [6] On the other hand, the case where a systematic lowering in notch frequency of traces occurred (in 3 of the 5 subjects previously removed) deserves a more careful consideration: from one of our previous studies [19], we identified a 20% of CIPIC subjects for whom a positive reflection coefficient better models the acoustic contribution of the pinna. Accordingly, it is reasonable to think that those three ex-

[5] 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 [7]) derives from several concurrent factors, such as dynamic auditory cues, visual information, familiarity with sound sources and training [23], thus it was not considered in this study.

[6] Repeatability of HRTF measurements are still a delicate issue, suggesting a high variability in spectral details [11].

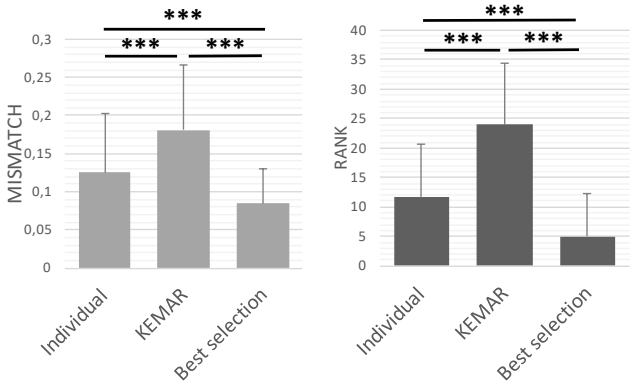


(a) Sub 51 - traces (b)

(c) Sub 3 - traces (d)

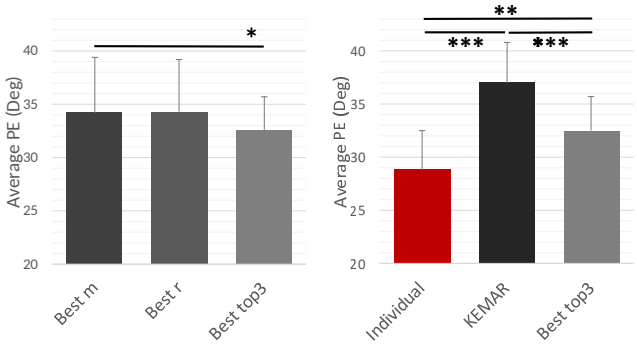
(e) Sub 44 - traces (f)

Figure 3: (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 this examples, best HRTF selection according to mismatch and rank metrics do not differ significantly.



(a) Mismatch (b) Rank

Figure 4: Global statistics (average + standard deviation) 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).



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

Figure 5: Global statistics (average + standard deviation) for localization prediction in elevation 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).

cluded subjects can be assigned to this special group. [7]

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. 4 clearly shows this aspect); Kruskal Wallis nonparametric one-way ANOVAs with three levels of feedback condition (individual HRTF, KEMAR, best m) provided a statistically significant result for mismatch [$\chi^2(2)$=1141.8, $p \ll 0.001$]; Pairwise post-hoc Wilcoxon tests for paired samples with Holm-Bonferroni correction revealed statistical differences among all pairs of conditions ($p \ll 0.001$). The same analysis with (individual HRTF, KEMAR, best r) as levels of feedback condition provided a statistically significant result for rank [$\chi^2(2)$=2362.3, $p \ll 0.001$] with statistical differences among all pairs of conditions ($p \ll 0.001$). At first glance, since we were trying to select the individual HRTFs from pinna contours, these results appear to be counter intuitive because we always selected a generic HRTF which differed from the individual HRTF in terms of both mismatch and rank. However, this evidence can not be misleading because we already know from our previous study in [5] that the notch associated to the pinna helix border is not enough to describe elevation cues for all listeners. Moreover, biometric recognition studies [24] 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 of notch frequency matches, providing a good average rank anyway (11), though preventing the individual HRTFs to be chosen as best selection.

Surprisingly, localization predictions from auditory model simulations provided average local polar RMS error (average PE) which has a statistically significant difference between best m and best top3 metrics, $t(16) = 2.134$, $p < 0.05$ (see Fig. 5.a for a graphical representation). This results suggest that best top3 yields better localization performances than best m, and with a similar compared to best r (not proven to be statistically significant in this study). Intuitively, best top3 metric is more robust to contour uncertainty because of the M-constraint in its definition, that allowed us to filter out variability due to HRTF set with sparse appearances in our rankings.

Finally, localization predictions were computed also for individual HRTFs and KEMAR virtual listening. Pairwise t-tests reveal significant differences in average PEs between individual listening condition and KEMAR ($t(16) = -7.79$, $p \ll 0.001$), and between individual listening condition and best top3 HRTF ($t(16) = -4.13$, $p < 0.01$), reporting a better performance with individual HRTFs. Moreover, pairwise t-test reports significant differences in average PEs between best top3 HRTF and KEMAR ($t(16) = 5.590$, $p \ll 0.001$), with a better performance of the selected HRTF compared to dummy-head listening condition. This final result further confirms the gap between individual HRTF listening and the proposed HRTF selection based on pinna image; on the other hand, best top 3 criteria selected generic HRTFs that outperformed KEMAR listening condition.

6. CONCLUSIONS

The proposed distance metrics considering 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. On the other hand, 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 $12th$ 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

[7] Unfortunately, we were not able to directly compare our current study with [19] because different CIPIC subjects were considered.

dataset.

- **better performance than KEMAR**: confirming our previous findings in psychoacoustic evaluation [5], 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.

Further research is still needed in order to increase the applicability of our notch distance metrics; CIPIC subjects can be also analyzed 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 [5]. 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.

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" which roughly acquires individual HRTFs using a smartphone loudspeaker as sound source and binaural microphones as receivers [25]. In this way, the frequencies $f_0^{(k,n)}(\varphi)$ could be directly computed in the acoustic domain, further reducing manual intervention.

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 phase [26] with the goal of overcoming common limitations in HRTF personalization studies, such as low number of participants (e.g. [17]), coherence in simplifications of localization experiment (e.g. [15]), and reliability of the predictions with computational auditory models [27].

7. REFERENCES

[1] J Blauert, *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing. Springer Berlin Heidelberg, 2013.

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

[3] 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," *The Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3832–3841, 2012.

[4] S Prepeliţă, M Geronazzo, F Avanzini, and L Savioja, "Influence of Voxelization on Finite Difference Time Domain Simulations of Head-Related Transfer Functions," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2489–2504, May 2016.

[5] M Geronazzo, S Spagnol, A Bedin, and F Avanzini, "Enhancing Vertical Localization with Image-guided Selection of Non-individual Head-Related Transfer Functions," in *IEEE Int. Conf. on Acoust. Speech Signal Process. (ICASSP 2014)*, Florence, Italy, May 2014, pp. 4496–4500.

[6] V. R Algazi, R. O Duda, D. M Thompson, and C Avendano, "The CIPIC HRTF Database," in *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, New Paltz, New York, USA, Oct. 2001, pp. 1–4.

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

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

[9] M Geronazzo, *Mixed structural models for 3D audio in virtual environments*, Ph.D. Thesis, University of Padova, Padova, Italy, Apr. 2014.

[10] 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 Science International*, vol. 187, no. 1–3, pp. 110.e1–110.e7, May 2009.

[11] A Andreopoulou, D Begault, and B Katz, "Inter-Laboratory Round Robin HRTF Measurement Comparison," *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, no. 99, pp. 1–1, 2015.

[12] W. G Gardner and K. D Martin, "HRTF Measurements of a KEMAR," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, June 1995.

[13] Y Iwaya, "Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears," *Acoustical science and technology*, vol. 27, no. 6, pp. 340–343, 2006.

[14] L Sarlat, O Warusfel, and I Viaud-Delmon, "Ventriloquism aftereffects occur in the rear hemisphere," *Neuroscience Letters*, vol. 404, no. 3, pp. 324–329, Sept. 2006.

[15] B. F. G Katz and G Parseihian, "Perceptually based Head-Related Transfer Function Database Optimization," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. EL99–EL105, Feb. 2012.

[16] D Zotkin, R Duraiswami, and L Davis, "Rendering localized spatial audio in a virtual auditory space," *IEEE Transactions on Multimedia*, vol. 6, no. 4, pp. 553 – 564, Aug. 2004.

[17] 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," *The Journal of the Acoustical Society of America*, vol. 136, no. 1, pp. 317–333, July 2014.

[18] V. C Raykar, R Duraiswami, and B Yegnanarayana, "Extracting the Frequencies of the Pinna Spectral Notches in Measured Head Related Impulse Responses," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 364–374, July 2005.

[19] 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.

[20] M Geronazzo, S Spagnol, and F Avanzini, "Estimation and Modeling of Pinna-Related Transfer Functions," in *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, Sept. 2010, pp. 431–438.

[21] M Geronazzo, A Carraro, and F Avanzini, "Evaluating vertical localization performance of 3d sound rendering models with a perceptual metric," in *2015 IEEE 2nd VR Workshop on Sonic Interactions for Virtual Environments (SIVE)*, Arles, France, Mar. 2015, pp. 1–5, IEEE Computer Society.

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

[23] E. M Wenzel, M Arruda, D. J Kistler, and F. L Wightman, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993, 00940.

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

[25] 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 Symposium on Virtual Reality Software and Technology (VRST 2016)*, Munich, Germany, Nov. 2016, pp. 23–26.

[26] M Geronazzo, J Kleimola, and P Majdak, "Personalization Support for Binaural Headphone Reproduction in Web Browsers," in *Proc. 1st Web Audio Conference*, Paris, France, Jan. 2015.

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