Audio Engineering Society

Conference Paper 14

Presented at the AES International Conference on Headphone Technology
2025 August 27–29, Espoo, Finland

This conference paper was selected based on a submitted abstract and 750-word precis that have been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This conference paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

One listener, sixteen HRTFs: a perceptual benchmark of individualisation methods

David Poirier-Quinot[1], Pablo Gutierrez-Parera[2,3], and Brian F. G. Katz[1]

[1]*Sorbonne Université, CNRS, Jean le Rond ∂'Alembert Institute, UMR 7190, 4 Place Jussieu, 75005 Paris, France*
[2]*Institute of Telecommunications and Multimedia Applications (ITEAM), Universitat Politècnica de València, Spain*
[3]*Telecommunication Research Institute (TELMA), Universidad de Málaga, Málaga, Spain*

Correspondence should be addressed to Brian F.G. Katz (brian.katz@sorbonne-universite.fr)

ABSTRACT

This paper presents initial results from a perceptual comparison of thirteen individual HRTF datasets, each obtained using different acquisition methods on a single subject. The evaluation, conducted by the same subject, employed two commonly used perceptual assessment approaches: a localization test and a trajectory quality ranking task. Results show that conventionally measured HRTFs yield the highest localization accuracy, outperforming simplified measurement techniques which produced the least accurate results. Numerical simulations using the boundary element method achieved performance comparable to that of measured HRTFs. In contrast, the commercial black-box solutions tested exhibited lower accuracy, on par with the three additional non-individual HRTFs included for reference. Findings also suggest that spectral distortion may serve as a rough predictor of the perceptual quality of HRTFs.

1 Introduction

Individualised Head-Related Transfer Functions (HRTFs) have long been shown to improve spatial auditory localisation performance in perceptual tasks [1]. Most existing studies have compared laboratory-measured individual HRTFs to generic, non-individual alternatives [2]. In contrast, relatively few have investigated the perceptual differences between individual HRTFs obtained under varying acquisition conditions, and even fewer have performed direct comparisons across multiple individualised datasets. With the growing availability of HRTF measurement facilities and the emergence of alternative acquisition methods – such as numerical simulations and commercial solutions from mesh scans or photographic data – there remains a lack of perceptual evaluations comparing a large set of individual HRTFs.

In this study, a single participant conducted a perceptual comparison of sixteen HRTFs, thirteen of which were either measured on them directly or generated through numerical simulations based on scans or video-derived morphological data. Two established perceptual assessment methods were employed: a localisation test [3] and a trajectory quality ranking task [4].

The structure of this paper is as follows. Section 2 describes the HRTF datasets and the two perceptual evaluation tasks. Section 3 presents the results, comparing HRTF performance across both tasks and analysing the influence of acquisition method. Section 4 explores the use of spectral distortion as an objective predictor of HRTF perceptual quality.

2 Materials and methods

2.1 Participant

A single participant took part in the experiment. The participant, one of the authors, is a researcher with over three decades of experience in spatial hearing and binaural rendering. Over this period, they have participated in numerous listening experiments, ranging from localisation tasks to authenticity evaluations. Hearing tests prior to the experiment confirmed the participant had no hearing loss in the audible frequency range.

2.2 HRTF datasets

Sixteen HRTFs were included in the evaluation. Of these, nine were obtained through direct acoustic measurements and four were generated via simulations based on images, videos, or 3D scans of the participant's morphology. The remaining three were non-individual HRTFs, included to provide additional context for interpreting the results. A summary of the HRTF datasets is provided in Table 1.

Table 1: HRTF datasets included in the evaluation.

Name	Type	Details
$H_{1\ \mathrm{meas}}$	meas.	-
$H_{2\ \mathrm{meas}}$	meas.	-
$H_{3\ \mathrm{meas}}$	meas.	-
$H_{4\ \mathrm{meas}}$	meas.	-
$H_{5\ \mathrm{meas}}$	meas.	-
$H_{6\ \mathrm{psim}}$	sim.	images + undisclosed algo.
$H_{7\ \mathrm{pmeas}}$	meas.	earmolds on mannikin
$H_{8\ \mathrm{pmeas}}$	meas.	MEMS + near field cylinder
$H_{9\ \mathrm{psim}}$	sim.	images + undisclosed algo.
$H_{10\ \mathrm{meas}}$	meas.	non-anechoic + cylinder
$H_{11\ \mathrm{sim}}$	sim.	scans + BEM
$H_{12\ \mathrm{meas}}$	meas.	-
$H_{13\ \mathrm{sim}}$	sim.	scans + BEM
$H_{14\ \mathrm{ni\text{-}ku100}}$	non ind.	KU100
$H_{15\ \mathrm{ni\text{-}1}}$	non ind.	-
$H_{16\ \mathrm{ni\text{-}2}}$	non ind.	-

Of the nine individual HRTFs, six were acquired using standard measurement procedures similar to those described by Engel et al. [5], in which the participant is typically seated in an anechoic chamber while test stimuli are presented from multiple positions on a spherical loudspeaker array. $H_{10\ \mathrm{meas}}$ was measured in a non-anechoic environment using a cylindrical loudspeaker distribution positioned 0.6 m from the participant, with time-windowing applied to suppress room reflections. $H_{7\ \mathrm{pmeas}}$ was obtained using a dummy head equipped with pinnae replicas moulded from the participant's anatomy. $H_{8\ \mathrm{pmeas}}$ was recorded using near-field MEMS loudspeakers in a cylinder, with responses subsequently extrapolated to far-field spherical positions.

Of the four simulated HRTFs, $H_{11\ \mathrm{sim}}$ and $H_{13\ \mathrm{sim}}$ were computed using a boundary element method (BEM) applied to high-resolution meshes of the participant's head and shoulders [6, 7]. Each was based on a different mesh obtained via a handheld light scanner. In contrast, $H_{6\ \mathrm{psim}}$ and $H_{9\ \mathrm{psim}}$ were generated using commercial undisclosed HRTF individualisation solutions, relying on images or videos captured with a smartphone camera rotated around the participant's head.

The three additional non-individual HRTFs – $H_{14\ \mathrm{ni\text{-}ku100}}$, $H_{15\ \mathrm{ni\text{-}1}}$, and $H_{16\ \mathrm{ni\text{-}2}}$ – correspond to a KU100 dummy head and two entries from the BiLi database [8].

2.3 Homogenisation of the HRTF datasets

Minor adjustments were applied to the HRTFs to ensure that none would perceptually stand out during evaluation. The first step in this homogenisation process addressed differences in head rotation across datasets. Since the notion of a "*neutral*" head orientation is difficult to define – and may vary depending on the measurement or simulation procedure – horizontal rotation offsets (*i.e.* rotation around the vertical axis through the participant's head) were corrected using an ITD alignment procedure.

Interaural time differences (ITDs) were estimated on the horizontal plane using the first onset detection method applied to HRTFs bandpass filtered between 300 Hz and 3000 Hz [9]. The resulting ITD patterns

were then aligned by applying a rotation that maximised correlation across datasets. The largest pre-alignment deviation observed was 4.7°. This correction step helped isolate perceptually relevant spectral and temporal differences by removing incidental misalignments between the participant and the measurement systems.

A critical requirement for the perceptual evaluation was that none of the HRTF sets should sound noticeably different from the others. Such discrepancies could allow the participant to recognise specific HRTFs, potentially compromising the blind evaluation protocol. They could also bias localisation performance. For example, Algazi et al. [10] reported that low spectral energy content is often associated with rearward sound sources. Presenting HRTFs with substantially different spectral energy distributions could therefore lead to systematic biases in perceived positions.

To minimise spectral differences introduced by the audio measurement chain (*e.g.* microphones or loudspeakers), Direct Transfer Functions (DTFs) were computed for each HRTF set and used to generate the final test stimuli. DTFs can be seen as "*equalised*" HRTFs, in which the frequency content common to all spatial directions – estimated via a spatially weighted average – has been removed [11].

A careful inspection of the impulse responses led to time-domain cropping of certain HRTFs to remove extraneous reflections. This step further ensured that no HRTF stood out perceptually, reducing the risk of the participant identifying individual datasets due to differences unrelated to HRTF characteristics.

Finally, a two-stage gain normalisation procedure was applied to ensure that none of the HRTF sets would be perceived as louder than the others. In the first stage, each set was normalised such that the energy of the frontal position measurement corresponded to unitary gain. The second gain adjustment, applied after the generation of the localisation and trajectory rating task stimuli, is described in Sections 2.4 and 2.5.

2.4 Localisation task description

The localisation task was conducted using external reference headphones (Sennheiser HD600) worn over a head-mounted display (HMD, Meta Quest 3). The virtual environment was developed in Unity and presented within the HMD. The experiment took place

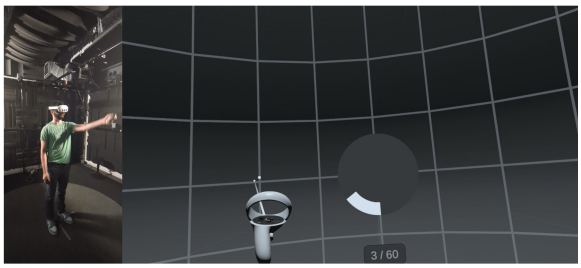


Fig. 1: (left) Example participant in the acoustically treated room during the localisation task. (right) Participant viewpoint in the virtual environment.

in an acoustically treated room with a mid-frequency reverberation time of $T_{30,1000\,\mathrm{Hz}} = 0.12\,\mathrm{s}$ and an ambient noise level below 30 dBA. The participant was immersed in a virtual scene, facing a visual anchor and positioned at the centre of a semi-transparent sphere with a radius of 1.95 m. The sphere featured a subtle texture to aid spatial memory and provide a stable visual frame of reference during the localisation task [3]. As shown in Figure 1, maintaining head orientation toward the anchor initiated a circular progress bar, which filled over approximately 1 s before triggering the spatialised audio stimulus.

The stimuli were sequences of three bursts of white noise, 40 ms each with a 4 ms cosine-squared onset/offset ramp separated by a 30 ms silence, for a total duration of 180 ms. Full spectrum burst sounds were chosen to optimize localization ability [1]. To avoid the participant becoming familiar with the noise signal, 8 different pre-generated random noise bursts were used, as well as random gains in a ± 3 dB range applied to each stimulus. Paired with the visual anchor mechanism of fixated gaze/head-orientation, the short audio stimulus prevents the participant initiating head movement during presentation of the target stimulus.

Upon hearing the stimulus, the participant used one of the HMD hand-held controllers to report the perceived location of the sound on the sphere. To facilitate aiming, each controller was equipped with a laser pointer, and a marker was displayed at the reported location upon validation. The reporting method was inspired from the "manual pointing" reporting method evaluated by Bahu et al. [12]. The reporting process was conditional on successful tracking of the hand-held controllers,

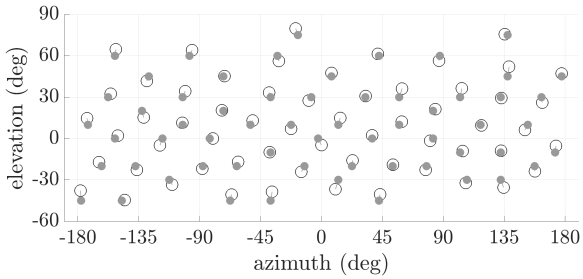


Fig. 2: Target positions used in the localisation task (white circles) and typical selected positions (gray circles) for HRTF $H_{5\ meas}$. No target positions were defined below 45° elevation, as most HRTF datasets did not include measurements in that region.

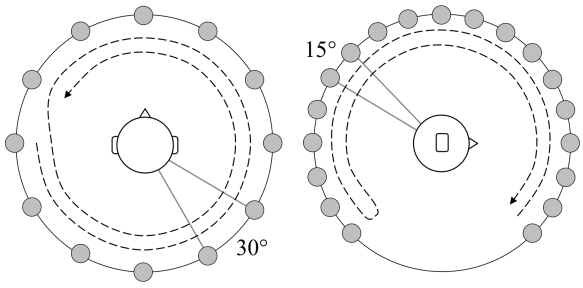


Fig. 3: Trajectories used in the rating task. (left) Horizontal and (right) median plane. Arrows indicate the start and end positions, as well as the direction of motion ($- \rightarrow$). Auditory positions ranged from 0° to 330° in the horizontal plane, and from $-45°$ to 225° in the median plane.

preventing responses in case of temporary tracking loss. After validation, the visual anchor reappeared, and the participant proceeded to the next localisation trial.

Stimuli were randomly spatialised at one of 60 predefined positions, evenly distributed on a sphere. Each position was presented once per HRTF, resulting in a total of $16 \times 60 = 960$ trials. This design was preferred over repeating a smaller set of source positions to prevent the participant from memorising target locations over the course of the experiment. The localisation task was completed in approximately 35 min, corresponding to a mean of 2.2 s per trial.

No interpolation was used in the selection of HRTFs for auralisation. Instead, the nearest available measurement to each target position was selected, as illustrated in Figure 2. For all datasets except $H_{8\ pmeas}$ and $H_{10\ meas}$, the selected positions were within 15° of the 60 predefined target positions. In the case of $H_{8\ pmeas}$ and $H_{10\ meas}$, the largest deviations were 18° and 38°, respectively, due to their cylindrical measurement distributions, which did not include directions above the participant's head.

All 960 stimuli were pre-rendered prior to being loaded into the Unity scene. To ensure consistent playback level across HRTFs, the gain of each HRTF was adjusted such that the 16 *sequences* – each consisting of the 60 aggregated stimuli – exhibited equal loudness. These sequences were assembled solely for the purpose of gain normalisation and were not used in the actual experiment.

2.5 Trajectory rating task description

The trajectory rating task was conducted using a Matlab-based user interface running on a laptop. The participant was seated in the same acoustically treated room and wore the same reference headphones as during the localisation task. Each trial consisted of rating 16 auditory trajectories – one per HRTF – on a fixed 9-point scale. Two types of trajectories were used, as illustrated in Figure 3: one rotating around the participant on the horizontal plane, and the other on the median plane. Before the task began, the participant was presented with both trajectory types. They were instructed to rate each HRTF according to how closely the perceived trajectory matched the intended one.

The participant was encouraged to use the full range of the rating scale and was required to assign at least one best-match (rank = 9) and one worst-match (rank = 1) per session. Both the horizontal and median plane trajectories were rated three times to assess consistency in HRTF evaluation [13]. HRTFs were anonymised and their presentation order was randomised across repetitions. The task was completed in about 90 min.

The same eight variants of the triple-burst white noise stimulus used in the localisation task were employed to construct the trajectories. Each burst group was separated by a 180 ms silent interval. To reduce the risk of HRTF identification across repetitions, distinct trajectory stimuli were generated for each of the repetitions using random sequences of the triple-burst variants. A level roving of ± 3 dB was applied to each triple burst,

and overall gain was adjusted to ensure consistent perceived loudness across all trajectories.

2.6 Data Analysis

The analysis of the localisation task follows the methodology and metrics detailed in [14]. In addition, the term *summed* confusions is used to denote the total number of "front-back", "in-cone", and "off-cone" confusions, *i.e.* any response falling outside a 45° cone centred on the target direction.

The term *score* refers to a metric value, normalised between zero and one, obtained by an HRTF in either perceptual task. The corresponding *rank* is derived from the score by sorting HRTFs from highest to lowest performance, preserving only the ordinal relationship between them.

Ranking similarity is quantified using Kendall's coefficient of concordance W, which ranges from zero (no agreement) to one (identical rankings). Similarity between continuous score distributions is computed using Pearson's linear correlation coefficient, which exhibits similar behaviour.

Table 2: Great-circle error, local great-circle error, and summed confusion rates observed during the localisation task. Mean values are reported across all 16 HRTFs, as well as for the top three performers ($H_{1\ meas}$, $H_{2\ meas}$, and $H_{3\ meas}$) based on great-circle error. Results from *majdak* [3] and *stitt* [2] are included for comparison. These studies evaluated training effects on localisation using individual HRTFs with 10 and 16 participants, respectively. Training durations varied, with a minimum of 2 h distributed over several days. Pre- and post-training results are shown.

	gc (deg)	local gc (deg)	conf. (%)
All HRTFs	35.8 ± 32.1	21.7 ± 11.5	23.1
3 best	27.7 ± 27.5	18.8 ± 11.4	15.0
majdak (pre)	34.0	20.8	30.0
majdak (post)	24.2	17.3	15.0
stitt (pre)	38.1	20.8	25.0
stitt (post)	27.9	18.2	15.0

3 Results

3.1 Localisation

The participant's overall localisation performance is summarised in Table 2, alongside results from two prior studies on individual HRTF localisation for comparison. The data indicate that the participant's average performance across all 16 HRTFs is in line with that of untrained participants using their own individual HRTFs. For the three best-performing HRTFs, the participant's accuracy reaches levels similar to those typically reported after multiple training sessions. Figure 4 shows localisation results across HRTFs.

3.2 Trajectory ratings

The participant's overall repeatability scores for the trajectory rating task are presented in Table 3. Compared to previously reported results on ratings of 8 non-individual HRTFs [15], the participant exhibited relatively low repeatability scores for both horizontal and median plane trajectories. This outcome was expected, as rating consistency generally decreases with a larger number of HRTFs, especially when many are individualised and thus more similar in perceptual quality. When restricting the analysis to the 8 HRTFs with the highest intra-condition consistency, the average repeatability score rises to 0.67, which is comparable to the findings of [15]. Figure 5 shows the participant results for the trajectory rating task.

3.3 Analysis of HRTF performance

The HRTF scores obtained from the two perceptual tasks are shown in Figure 6. The results reveal a strong correlation between the localisation and trajectory rating scores. Given this consistency, the mean rank across both tasks will be used in subsequent analyses to evaluate overall HRTF performance.

Results suggest that high-quality measured HRTFs outperform both non-individual and simulated HRTFs. As expected, some measured HRTFs rank lower, likely due to inconsistencies in the measurement process. The poor overall performance of $H_{8\ pmeas}$ and $H_{7\ pmeas}$ indicates that poorly measured individual HRTFs can be outperformed by random or even worst-match non-individual sets. A detailed analysis of the causes behind

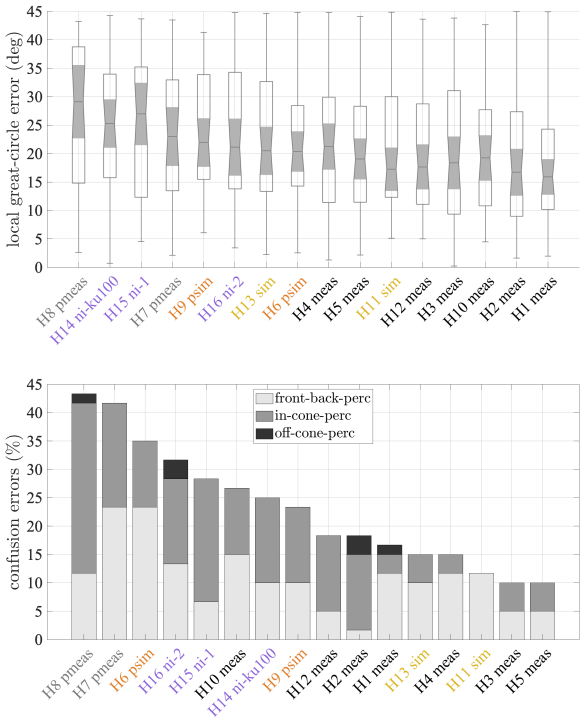


Fig. 4: (top) Local great-circle error and (bottom) summed confusion rates across HRTFs. HRTFs are ordered from left to right by decreasing localisation performance, based on (top) local great-circle error and (bottom) summed confusion rates, respectively.

these underperforming sets is beyond the scope of this preliminary study. Notably, the five top-ranking individual HRTFs were measured using procedures similar to that described by Pollack et al. [16].

Simulated HRTFs appear average at first glance, better than non-individual ones but not matching the best measured sets. However, in terms of localisation performance, $H_{11\ \mathrm{sim}}$ and $H_{13\ \mathrm{sim}}$ closely approach top-tier measured HRTFs (Figure 6). This is largely due to their excellent confusion rates (Figure 4), though slightly lower scores in local great-circle error and trajectory ratings kept them from matching the top performers.

4 Objective predictor of HRTF performance

An analysis was conducted using spectral distortion (SD) between HRTFs to investigate whether this objective metric could serve as a predictor of perceptual performance. SD was calculated following a method similar to that described by Andreopoulou and Katz [17], using DTFs without applying spectral smoothing.

Before computing SD, several pre-processing steps were applied to the DTFs. By design, DTFs remove common spectral content and ensure a consistent reference level across HRTF sets. To enable fair comparison, each DTF was interpolated to match the spatial sampling of the reference HRTF. Following the recommendation in Andreopoulou and Katz [17], only sources between elevations $-45°$ and $45°$ were retained to avoid interpolation artefacts in the polar regions, where measurements are typically sparse. SD was computed only for the ipsilateral ear (left ear for left hemisphere and vice versa) to reduce the influence of low signal-to-noise ratios often encountered in the contralateral ear. The frequency range considered for SD computation was limited to 3 to 18 kHz.

The analysis used the top-ranking HRTF $H_{3\ \mathrm{meas}}$ as the reference. As shown in Figure 6, the resulting SD values were highly correlated with the perceptual scores from both evaluation tasks. However, closer inspection reveals that this correlation is largely driven by poorly performing HRTFs. Among higher-ranking HRTFs, SD values tend to vary and offer less predictive power at the individual level. In this context, SD seems less

Table 3: Repeatability scores for the trajectory rating task, computed using Kendall's coefficient of concordance W. W_{hor} and W_{med} denote the repeatability scores for the horizontal and median plane trajectory rankings, respectively. $W_{\mathrm{hor+med}}$ represents the repeatability score computed across all rankings. Scores are reported for all 16 HRTFs, as well as for the subset of the eight most repeatable HRTFs selected based on $W_{\mathrm{hor+med}}$. Results from 28 participants in *zagala* [15], who performed the same task on 8 non-individual HRTFs, are included for comparison.

	W_{hor}	W_{med}	$W_{\mathrm{hor+med}}$
All HRTFs	0.35	0.46	0.41
8 best	0.95	0.93	0.67
zagala [15]	.	.	0.60

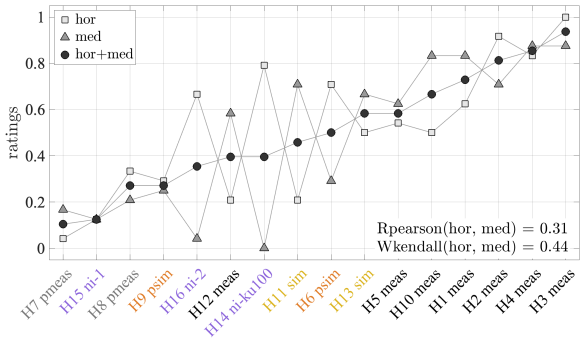


Fig. 5: Results of the trajectory rating task. Comparison of horizontal and median plane trajectory ratings across HRTF, ordered left to right by increasing combined rating (horizontal + median).

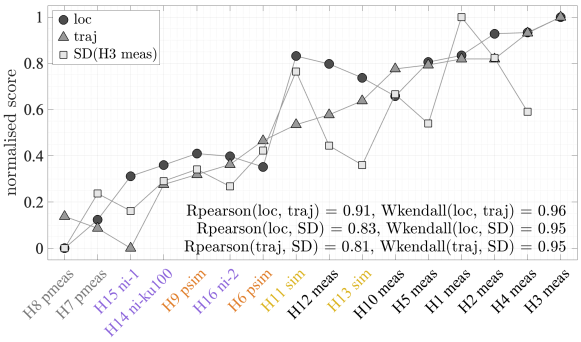


Fig. 6: Comparison of HRTF scores between the localisation (great-circle error) and trajectory (combined rating) tasks. HRTFs ordered from left to right by increasing average score across both tasks. SD($H_{3\ meas}$) represents the spectral distortion from the top-ranking HRTF $H_{3\ meas}$, discussed in Section 4. For both SD and great-circle error, the value used is $1 - \text{score}$, normalised to the $[0, 1]$ range, ensuring that higher scores consistently indicate better HRTF performance across metrics.

useful for discriminating between good and excellent HRTFs. While SD may not be practical for selecting optimal HRTFs – since doing so presupposes access to a reliable reference – it could still be valuable for validating the perceptual similarity of HRTFs obtained via alternative acquisition methods. Its utility would lie primarily in identifying large perceptual deviations (*e.g.* non-individual vs. individual), rather than making fine-grained distinctions among individualised sets.

5 Conclusion

This study reported the results of two perceptual evaluation tasks conducted by a single participant on sixteen HRTFs, thirteen of which were individual, either measured on or simulated for the participant across a span of thirty years. The two tasks – a localisation test and a trajectory rating – provided complementary assessments of HRTF quality. Their outcomes, which were strongly correlated, served to compare the relative performance of the HRTFs.

As somewhat anticipated, the best-performing HRTFs were those obtained through conventional measurement procedures. Nonetheless, two BEM-based simulations approached the top-performing measured HRTFs, particularly in localisation accuracy. The results also confirmed that a poorly measured individual HRTF could be outperformed by non-individual alternatives. Finally, SD computed against the best perceptually rated HRTF emerged as a useful objective predictor of performance in both tasks, though limited to distinguishing broadly divergent sets.

Future work will include a more detailed statistical analysis of the reported results. Further investigation will explore the use of the Auditory Modelling Toolbox [18] to predict HRTF performance with the same intention as SD was used here. A second testing phase will directly compare HRTF and DTF renderings to quantify their perceptual differences. Finally, a third phase will replicate the evaluation in a mildly noisy environment to assess whether performance differences persist under acoustic conditions more representative of an augmented reality context.

References

[1] Begault, D. R., Wenzel, E. M., and Anderson, M. R., "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," **Journal of the Audio Engineering Society**, 49(10), pp. 904–916, 2001.

[2] Stitt, P., Picinali, L., and Katz, B. F. G., "Auditory accommodation to poorly matched non-individual spectral localization cues through active learning," **Scientific Reports**, 9(1), pp. 1063:1–14, 2019, doi:10.1038/s41598-018-37873-0.

[3] Majdak, P., Goupell, M. J., and Laback, B., "3-D localization of virtual sound sources: effects of visual environment, pointing method, and training," **Attention, Perception, & Psychophysics**, 72(2), pp. 454–469, 2010, doi:10.3758/APP.72.2.454.

[4] Parseihian, G. and Katz, B. F. G., "Rapid head-related transfer function adaptation using a virtual auditory environment," **Journal of the Acoustical Society of America**, 131(4), pp. 2948–2957, 2012, doi:10.1121/1.3687448.

[5] Engel, I., Daugintis, R., Vicente, T., Hogg, A. O., Pauwels, J., Tournier, A. J., and Picinali, L., "The SONICOM HRTF dataset," **Journal of the Audio Engineering Society**, 71(5), pp. 241–253, 2023, doi:10.17743/jaes.2022.0066.

[6] Ziegelwanger, H., Kreuzer, W., and Majdak, P., "Mesh2HRTF: an open-source software package for the numerical calculation of head-related transfer functions," in **International Congress on Sound and Vibration (ICSV)**, pp. 1–8, 2015, doi:10.13140/RG.2.1.1707.1128.

[7] A. Bourachot, J.-B. D., C. Langrenne, "Évaluation des performances de localisations de sources binaurales issus d'HRTFs modélisées à partir de scans 3D," in **Congrès Français d'Acoustique**, pp. 1–1, 2025.

[8] Carpentier, T., Bahu, H., Noisternig, M., and Warusfel, O., "Measurement of a head-related transfer function database with high spatial resolution," in **Forum Acusticum**, 2014.

[9] Andreopoulou, A. and Katz, B. F. G., "Identification of perceptually relevant methods of inter-aural time difference estimation," **Journal of the Acoustical Society of America**, 142(2), pp. 588–598, 2017, doi:10.1121/1.4996457.

[10] Algazi, V. R., Avendano, C., and Duda, R. O., "Elevation localization and head-related transfer function analysis at low frequencies," **Journal of the Acoustical Society of America**, 109(3), pp. 1110–1122, 2001, doi:10.1121/1.1349185.

[11] Middlebrooks, J. C., "Individual differences in external-ear transfer functions reduced by scaling in frequency," **Journal of the Acoustical Society of America**, 106(3), pp. 1480–1492, 1999, doi:10.1121/1.427176.

[12] Bahu, H., Carpentier, T., Noisternig, M., and Warusfel, O., "Comparison of different egocentric pointing methods for 3D sound localization experiments," **Acta Acustica**, 102(1), pp. 107–118, 2016, doi:10.3813/AAA.918928.

[13] Andreopoulou, A. and Katz, B. F. G., "Investigation on subjective HRTF rating repeatability," in **Audio Engineering Society Convention 140**, pp. 9597:1–10, 2016.

[14] Poirier-Quinot, D., Lawless, M. S., Stitt, P., and Katz, B. F. G., "HRTF performance evaluation: methodology and metrics for localisation accuracy and learning assessment," in **Advances in Fundamental and Applied Research on Spatial Audio**, chapter 3, pp. 1–32, Intech Open, 2022, doi:10.5772/intechopen.104931.

[15] Zagala, F., Noisternig, M., and Katz, B. F. G., "Comparison of direct and indirect perceptual head-related transfer function selection methods," **Journal of the Acoustical Society of America**, 147(5), pp. 3376–3389, 2020, doi:10.1121/10.0001183.

[16] Pollack, K., Majdak, P., and Kreuzer, W., "Modern acquisition of personalised head-related transfer functions: an overview," in **Advances in Fundamental and Applied Research on Spatial Audio**, chapter 2, IntechOpen, 2022, doi:10.5772/intechopen.102908.

[17] Andreopoulou, A. and Katz, B. F. G., "Perceptual impact on localization quality evaluations of common pre-processing for non-individual head-related transfer functions," **Journal of the Audio Engineering Society**, 70(5), pp. 340–354, 2022, doi:10.17743/jaes.2022.0008.

[18] Majdak, P., Hollomey, C., and Baumgartner, R., "AMT 1.x: a toolbox for reproducible research in auditory modeling," **Acta Acustica**, 6(19), 2022, ISSN 2681-4617, doi:10.1051/aacus/2022011.