IEEE Open Journal of Signal Processing

Received 26 February 2025; revised 5 July 2025; accepted 17 July 2025. Date of publication 28 July 2025; date of current version 8 August 2025. The review of this article was arranged by Associate Editor S. Koyama.

Digital Object Identifier 10.1109/OJSP.2025.3592601

Listener Acoustic Personalization Challenge—LAP24: Head-Related Transfer Function Dataset Harmonization

RAPOLAS DAUGINTIS [1], ROBERTO BARUMERLI [2], MICHELE GERONAZZO [1,3] (Senior Member, IEEE), JOHAN PAUWELS [4], LORENZO PICINALI [1], AND KATARINA C. POOLE [1]

[1]Audio Experience Design, Dyson School of Design Engineering, Imperial College London, SW7 2DB London, U.K.
[2]Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, 37129 Verona, Italy
[3]Department of Engineering and Management, University of Padova, 35122 Padova, Italy
[4]Centre for Digital Music, Queen Mary University of London, E1 4NS London, U.K.

CORRESPONDING AUTHOR: RAPOLAS DAUGINTIS (e-mail: r.daugintis21@imperial.ac.uk).

This work was supported by the SONICOM Project, funded by the Horizon 2020 Research and Innovation funding Programme under Grant Agreement 101017743.

ABSTRACT Big data analysis and collation for data-driven head-related transfer function (HRTF) personalization methods are often hindered by systematic differences between HRTF datasets. To address this issue, we designed Task 1 of the inaugural listener acoustic personalisation (LAP) challenge. Researchers were invited to propose strategies for harmonizing HRTFs from a collection of eight different datasets so that dataset-specific artifacts were mitigated while preserving the perceptually relevant attributes of the original HRTFs. Defining the two-sided task required a deeper assessment of the acoustic and perceptual HRTF descriptions to find an evaluation framework that encompassed the two domains. Consequently, a two-stage evaluation was devised to assess the submissions. First, an auditory sound localization model was used to test the perceptual validity of the harmonized HRTFs by estimating the difference in sound localization performance between the original and the harmonized versions. Then, a machine learning classifier was employed to differentiate harmonized HRTF datasets, and its accuracy was used to rank submissions. Three submissions were received, and one was declared a winner according to the evaluation criteria. Further analysis of the submissions revealed some limitations of the evaluation system, prompting a comprehensive review of the task's inherent complexities. This paper serves as a systematic account of the challenge and relevant considerations, intended to guide future advancements in the field of HRTF personalization research.

INDEX TERMS Computational auditory modeling, head-related transfer function harmonization, HRTF dataset, machine learning classification.

I. INTRODUCTION

Immersive audio is becoming increasingly popular in virtual and augmented reality (VR/AR) applications, teleconferencing, gaming, music, and film production. In a headphone-based scenario, the term typically refers to binaural audio rendering based on convolution with head-related transfer functions (HRTFs) to virtually emulate realistic spatial audio scenes for the listener. While individual HRTFs from detailed acoustic measurements or simulations using high-quality 3D head scans have been shown to provide the most realistic spatial audio experience [1], they are currently unfeasible for the mass market due to equipment and time constraints [2]. Instead, various indirect personalization techniques have been proposed [3], [4], some of which use data-driven approaches (a detailed review of existing literature is provided in [5]). However, multiple challenges still need to be addressed before machine learning (ML) methods can show their full potential in personalizing HRTFs to each user.

The Listener Acoustic Personalisation (LAP) challenge was designed as a platform to explore some of these open challenges, advance the state of the art, and contribute to the development of standardized metrics for personalized spatial

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

audio.[1] Previous challenges in spatial acoustics and hearing science have tackled related topics, for example, the LOCATA challenge for machine sound source localization [6], or the Clarity challenge for enhancing speech on hearing aids [7]. However, the LAP Challenge is, to our knowledge, the first of its kind in the immersive audio domain, dealing explicitly with both numerical and perceptual aspects of spatial hearing and spatial acoustic rendering. The first edition of the challenge ran between March and August 2024 and included two tasks. The focus of Task 1 was HRTF harmonization for the purpose of merging different HRTF datasets, while Task 2 dealt with HRTF spatial upsampling from a low number of directions. This paper reports and discusses outcomes of Task 1, while information about Task 2 can be found in an accompanying paper [8].

Task 1 of the challenge was motivated by the limited availability of high-quality HRTF measurements, restricting the generalization of ML algorithms for HRTF personalization. Various attempts have been made to use ML methods to, for example, synthesize HRTFs from anthropometric pinna measurements [9], [10] or upsample a small set of measurements to a dense HRTF grid [11]. However, the quality of ML solutions can depend on the quantity of training data [12], [13]. Due to the laborious and costly HRTF acquisition process, typical measured HRTF datasets contain data from fewer than 100 subjects, but few ML methods attempted to combine data from different datasets [5]. Reluctance to merge multiple datasets for ML training can be attributed to measurable dissimilarities between HRTFs from different databases [14]. Detectable dataset-specific artifacts in HRTFs [15] can cause overfitting and reduce generalizability across ML methods [16], thereby limiting the effectiveness of cross-dataset domain transfer approaches [13]. Furthermore, it remains uncertain to what extent these cross-database variations affect perceptual aspects of HRTF-based spatial audio rendering.

To address the HRTF dataset incompatibility issue, the Task 1 was dedicated to the matter of HRTF dataset harmonization. We defined an ideal harmonized HRTF as an HRTF that contains all the perceptually relevant components of its original counterpart but with dataset-specific artifacts removed or mitigated so that its origin could not be detected. This paper presents an overview of the challenge and the resulting submissions, providing the reader with a theoretical and practical review of the differences in HRTF acquisition methodologies (Section II), as well as defining the task and its evaluation methodology (Section III). Then it provides a short summary of the three submissions received (Section IV) and their performance (Section V). Since this was the first attempt at organizing a challenge in the immersive audio domain, its design required the development of several evaluation components from the ground up, with limited opportunities to draw on established methodologies or prior published work. As a result, several shortcomings became apparent during the submission analysis phase. This prompted a more in-depth analysis of the submitted data, presented in Section VI, along with a discussion of the challenge's limitations and potential improvements to the setup in Section VII.

[1] [Online]. Available: https://www.sonicom.eu/lap-challenge/

II. BACKGROUND

A. HRTF: ACOUSTIC DEFINITION AND PRACTICAL LIMITATIONS

Acoustically, the HRTF describes how a sound wave originating from a source in the free field is transformed at the entrance to the ear canal, accounting for the acoustic effects of the listener's head, torso, and pinnae [17]. Mathematically, it can be defined as follows (assuming far-field condition):

$$\mathrm{HRTF}_{L,R}(\theta, \phi, f) = \frac{P_{L,R}(\theta, \phi, f)}{P_{\mathrm{ref}}(f)} \tag{1}$$

where $P_{L,R}(\theta, \phi, f)$ represents the complex sound pressure at the left or the right ear canal as a function of frequency f and incident direction (θ, ϕ) of the sound wave, while $P_{\mathrm{ref}}(f)$ denotes the reference sound pressure at the center of the head without the listener present.[2]

Although this theoretical definition provides a precise formulation of HRTFs, practical implementations rely on acoustic measurements conducted in specialized facilities. These measurements introduce variations due to factors such as internal equipment noise, deviations from an ideal free-field environment, uncertainty of microphone placement, and other measurement constraints [18]. To enhance the quality and consistency of measured HRTFs, various post-processing methods have been identified. For example, some HRTF datasets are diffuse-field compensated by removing the average transfer function (i.e. the common transfer function) from the HRTF [19]. The compensated transfer functions are also known as directional transfer functions (DTFs) [20]. Some may also use methods to compensate for the poor low-frequency response of the loudspeakers used in the measurements [21], [22]. However, procedural differences across measurement facilities [23] further contribute to dataset-specific artifacts, complicating cross-dataset HRTF comparisons and underscoring the challenge of ensuring consistency across different datasets.

Existing compensation methods are limited in their ability to remove all setup-specific artifacts due to factors such as unaccounted systemic measurement biases, individual measurement variability, and mathematical limitations of traditional signal processing algorithms. This results in numerically significant differences in spectral magnitudes and variations in estimated interaural time differences between measurements of the same mannequin head at different facilities [14]. The heterogeneity in HRTF databases becomes even more obvious for ML algorithms, tasked with classifying HRTFs based

[2] Here, we refer to a singular HRTF as a set of direction-dependent filters, corresponding to one subject and an HRTF dataset as a collection of HRTFs measured in the same facility.

on their signals in the time or magnitude domain: Analysis of classification performance using gradient boosted trees reveals that low (below 5 kHz) and high (above 15 kHz) frequency regions are particularly informative to the classifiers [15].

Nonetheless, identifying and extracting dataset-specific artefacts remains challenging due to the complex multidimensional nature of HRTFs across direction, frequency, and listener morphology. Previous proposals on mitigating some of the numerical discrepancies between HRTF datasets include normalizing the original HRTFs by the global average and standard deviation [24] or by directional averages [16] of each HRTF database, as well as using Bayesian fusion [25] or extreme frequency smoothing [26]. However, the perceptual impact of these methods on HRTF quality has not been investigated. Therefore, more extensive perceptual evaluations are needed to understand how the numerical differences between HRTFs relate to perceived differences by human listeners.

B. HRTF: PERCEPTUAL CONSIDERATIONS

Numerical differences between datasets suggest the need for some level of harmonization beyond standard compensation methods when merging data. This is even more relevant when measured HRTF data are augmented by numerically modeled HRTFs for ML-based HRTF personalization [12]. Simulations can also vary due to mesh acquisition and numerical modeling errors [27], [28] and can be perceivably different from their measured counterparts [22]. However, the harmonization procedure must retain at least the perceptual *validity* of each HRTF. Therefore, finding a perceptually meaningful description of HRTFs is a key first step in developing a successful harmonization process.

When characterizing the HRTF, the conceptual distinction between *strong* and *narrow* definitions may be useful. These mirror broader debates, such as those in artificial intelligence (AI), where *strong/general AI* aims to replicate human cognition entirely and *narrow/weak AI* focuses on specific task-oriented capabilities [29]. Similarly, a *strong* HRTF aspires to encapsulate the full physical reality of acoustic transformations through rigorous and highly detailed measurements, emphasizing fidelity to physical properties. It characterizes the HRTF as an exhaustive acoustic measurement of the individual's body, accounting for every anatomical detail that affects sound propagation, such as the shape and position of the ears, head, and torso. This perspective emphasizes the *in vivo* measurement of HRTFs, aiming at a representation as close to physical reality as possible. However, it inherently assumes a static model, disregarding dynamic variables such as transient movements, skin elasticity, or even hair growth. The *strong* definition aims for the ideal but practically unattainable goal of achieving zero dissimilarity with negligible validation uncertainty.

Although, in general, measurements are considered a better indicator of reality [30], precise HRTF measurements are still impractical and physically unvalidated at any standardized level, despite their proven perceptual validity [31], [32], [33] and their potentially short acquisition times [34], [35], [36]. Moreover, significant discrepancies revealed by previously discussed cross-evaluation studies [14], [22] reflect a common challenge in establishing a universally accepted *ground truth* [37], [38]. These challenges illustrate the gap between theoretical aspirations and practical limitations, which requires complementary strategies such as incorporating perceptual validation metrics alongside physical measurements.

In contrast, *narrow* HRTFs can have broader applicability due to their adaptability and lower computational demands. By focusing on perceptual metrics, such as just-noticeable differences (JNDs), *narrow* HRTFs can be effectively validated within specific auditory dimensions. For example, left-right localization tasks rely on minimal interaural level and time differences (ILDs and ITDs), which can be approximated without exhaustive physical accuracy. Studies have shown that JNDs for ILD and ITD are higher in realistic reverberant environments compared to anechoic ones [39] and that JNDs for ITD increase for more lateralized positions [40], suggesting context-driven approximation strategies. Similarly, research on spectral peak and notch detection thresholds provides an indication of the spectral resolution requirements for pinna-related spectral cues [41] and highlights the importance of specific peaks and notches for sagittal plane localization [42]. In virtual auditory displays, approximate HRTF data might suffice to create convincing spatial auditory experiences without the exhaustive detail required in the *strong* definition [43]. The *narrow* definition acknowledges the inherent trade-offs in scalability, reproducibility, and context-specific validation of HRTF measurements and simulations.

The dichotomy between *strong* and *narrow* HRTF definitions underscores the interplay between theoretical rigor and practical utility, the former an aspirational benchmark and the latter a viable framework for advancing HRTF research and applications within the constraints of current technology.

C. METRIC SELECTION: MOTIVATION AND GOALS

To assess the HRTF harmonization task, a reasonable *narrow* HRTF definition must be found to ensure that harmonized HRTFs retain their relevant perceptual qualities while diminishing the idiosyncrasies introduced by specific acquisition methods. From various perceptual attributes that could be used to describe the quality (and validity) of a harmonized HRTF [44], we chose to focus on the sound source localizability as the most important aspect of HRTF-based binaural audio rendering. We expect that a measured individual HRTF should contain all the necessary acoustic cues for the listener to successfully associate the rendered sound with the correct direction of the sound source, despite the minor levels of noise introduced by the measurement system. Therefore, listener performance should be comparable when localizing binaural stimuli spatialized with the original and the harmonized HRTF. We acknowledge that ensuring comparable sound

source localisation performance might not prevent degradation of other qualitative aspects of binaural rendering, such as the perception of sound coloration or externalization. However, the objective formulation of sound source localizability via localization error metrics and its strong connection to other subjective qualities [45] provide a reliable framework for evaluating perceptual HRTF validity.

Although the *ground truth* for every perceptual assessment are human responses, designing and performing a robust listening test is time-consuming and labor-intensive. It becomes especially difficult when a large number of qualitative comparisons must be made, leading to problems with the reliability of the results [46], [47]. To partly address these issues, computational auditory models can be used to estimate human responses and reduce the need for physical tests [48]. Auditory models that aim to predict human sound localization performance can provide a useful metric to assess whether the harmonized HRTF has retained its essential perceptual components. These models work on the principle of template comparison (e.g., [49], [50]): They compare auditory cues (ITD, ILD, and/or spectral cues) contained in the template (e.g. original) HRTF and the target (e.g. harmonized) HRTF. Their output is an estimate of the sound source direction that is on par with human subject responses when performing a sound localization test with sounds rendered using a target HRTF. The average localization error across all simulated positions for each HRTF pair provides an interpretable metric framework [51] that can be compared to human localization performance [52].

While retaining the perceptual components of the original HRTF that provide accurate spatial cues to the listener, the harmonized version should ideally not contain any identifiable *signature*, specific to its acquisition system, in its signal. Differences between datasets can be perceptual (that is, identifiable by listeners) and/or numerical (measurable by means of numerical analysis). Although the elimination of perceptual differences is usually the main goal of any HRTF processing method, defining its success is cumbersome due to the aforementioned challenges with performing listening experiments. Compared to human hearing, numerical techniques may exhibit greater sensitivity to HRTF variations and hence provide an efficient and reproducible assessment of harmonization performance. This, together with the desire to allow participants to check their own solutions, served as motivation to use numerical ML methods tasked with identifying HRTF datasets [15] as evaluation criteria for HRTF harmonization. Although numerical evaluation methods should ideally align with perceptual models to avoid minimizing negligible HRTF differences, the development of such methods requires further research into task-specific perceptual sensitivity.

The two constraints — HRTF perceptual validity and its identifiability — can act in opposition to each other: removing database-specific HRTF artifacts often risks degrading HRTF's perceptual quality. This polarity poses a challenge when defining an optimal evaluation metric that takes both sides into account. Similar considerations could be found in previously organized challenges in the audio processing domain, such as the ICASSP 2023 Clarity Challenge [53] and the second Cadenza Challenge [54]. In the LAP challenge, a two-stage evaluation process was designed to first assess whether the submitted HRTFs were still perceptually valid (that is, they retained a *narrow* HRTF definition) and then assess how distinguishable each collection of the harmonized HRTFs is. This procedure is summarized in Fig. 1 and is described in more detail in the following section.

III. TASK OVERVIEW

In Task 1 of the 2024 LAP challenge, participants were asked to harmonize HRTFs from eight datasets measured or modeled in different research facilities. They were provided with a collection of 80 original HRTFs and asked to submit their harmonized versions. In addition, the participants had to provide a brief accompanying report explaining their methodology. They were encouraged to make their method public, but closed-source proprietary methods were allowed.

A. TASK DATASETS

Eight HRTF datasets were chosen based on the diversity of acquisition methods and the density of sampling grids (so that a common set of HRTF positions could be used for evaluation) while balancing the overall submission size. The details of each dataset are given in Table 1. Ten HRTFs were selected from each dataset, and the table lists their IDs, as found in the SOFA repository [61]. The 80 HRTFs were given to participants in their original form in the SOFA format [62] with all submissions expected in the same format.

B. TASK EVALUATION

As presented in Fig. 1, submissions were evaluated in two stages to account for the two constraints: perceptual HRTF validity (Stage 1) and difficulty in identifying its origin (Stage 2). Stage 1 acted as a validation stage, so submissions that were below defined thresholds were included in a further evaluation. Then, Stage 2 evaluation provided a classification accuracy that was used to rank the submissions and determine the winner.

1) STAGE 1

Stage 1 evaluation was performed using an auditory sound source localization model [50]. The auditory model uses a Bayesian framework to estimate the direction of the given sound source. It extracts auditory features from the binaural input (i.e., the target HRTF), including band-limited ITDs computed from binaural signals passed through a 10th-order low-pass Butterworth filter at 3000 Hz, ILDs from the broadband signals, and band-limited spectral cues via an equally weighted gammatone filter bank with equivalent rectangular bandwidth (ERB) bands between 700 Hz and 18 kHz. From the features computed from the target HRTF, the model compares them to stored templates (i.e. precomputed features from the original HRTF for specific directions) to generate a likelihood function.

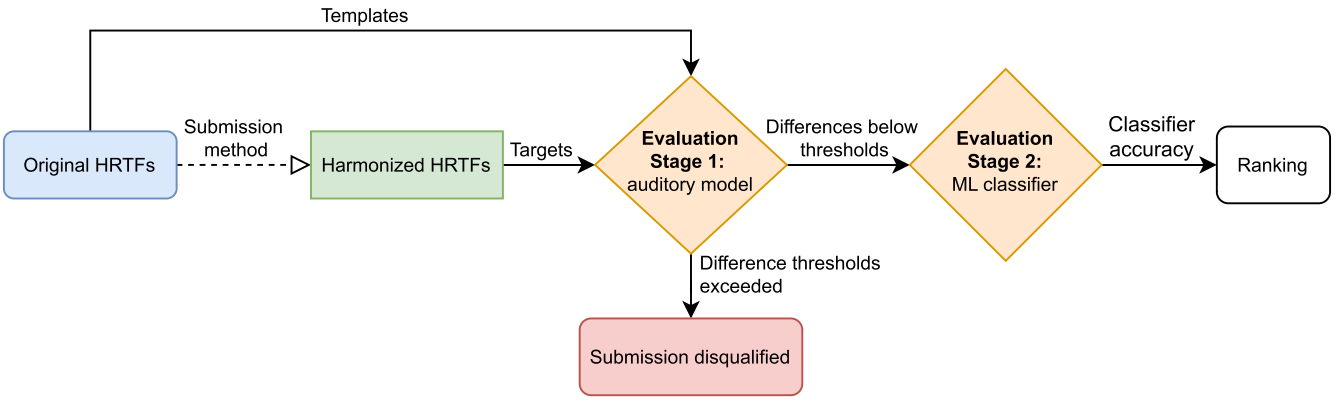


FIGURE 1. The challenge evaluation structure.

TABLE 1. Details of the HRTF Dataset Collection Selected for the Challenge

Dataset name	Sample rate (kHz)	HRIR duration (ms)	No. directions	Source distance (m)	Acquisition method	Subject IDs
3D3A [55]	96	21.333	648	0.76	measured	Subject[1, 2, 4, 10–16]
CHEDAR [56]	48	10	2522	2	simulated (head & torso)	00[01–10]
HUTUBS [22]	44.1	5.805	440	1.47	measured	pp[2–11]
RIEC [57]	48	10.666	865	1.5	measured	subject_[1–10]
SADIE II [21]	96	5.333	2114	1.2	measured	H[10–19]
SCUT [58]	96	5.333	864	1[1]	measured	subject00[01–10]
SONICOM [59]	96	5.333	828	1.5	measured	P00[01–10]
WiDESPREaD [60]	48	10	2522	2	simulated (pinnae only)	000[01–07, 09–13]

[1] The files contain HRIRs at different distances but only the ones at 1 metre were considered.

The model combines the extracted auditory features with a prior distribution that models human localization bias toward the horizontal plane [63]. Finally, it computes a posterior probability and identifies the maximum a posteriori (MAP) estimate of the source direction. The model has a set of free parameters to control its sensitivity to different features. For this evaluation, they were calibrated according to [64]. The computed localization metrics used for the evaluation were based on the aggregated differences between the original (target) positions and the estimated (response) positions along the interaural-polar coordinate axis [2]. They included metrics defined in [20]: lateral bias (a.k.a. accuracy [65]) and root mean square (RMS) error (a.k.a. precision; both calculated for positions within 60° lateral angle from the median plane to reduce the ceiling effect), local polar bias and RMS error (calculated for responses that were within 90° polar angle from the target and within 30° lateral angle from the median plane to avoid the high compression of the polar angle at the poles), and the quadrant error (a percentage rate of responses that were more than 90° polar angle from the target within 30° lateral angle from the median plane). In addition, a polar gain (the slope of the regression line for response vs. target polar angle for positions within 30° lateral angle from the median plane [66]) was used. Since each metric targets a specific localization phenomenon (e.g., local polar RMS error for fine-scale errors, quadrant error for gross confusions, and accuracy and polar gain for systematic bias), we employed this set to comprehensively evaluate performance in both horizontal and vertical dimensions.

To assess the distortion introduced by the harmonization process, thresholds on the maximum allowable difference between the error metrics generated by the model when localizing with the original HRTF and its harmonized version were defined (using the original HRTF features as a template in both cases). The thresholds were set by estimating the localization metrics using the so-called *Club Fritz* HRTF collection [14]. This collection comprises nine HRTFs,[3] each measured with the same mannequin head (Neumann KU100) but in different research facilities. Previous analyses had already revealed substantial inter-laboratory deviations in perceptually relevant features (e.g., up to 235 μs ITD and over 12.5 dB spectral amplitude differences [14]) and for simulated localization metrics in the median plane during a round-robin comparison [51]. To establish the thresholds, localization metrics were computed with the chosen auditory model in a round-robin fashion across the *Club Fritz* HRTFs using 72 common source directions (Fig. 2(a)). Thresholds were then set to the maximum differences between the localization metrics with the same target HRTF as the template and all other HRTFs as targets. These thresholds are presented in Table 2.

[3]The HRTF dataset measured at the Helsinki University of Technology was excluded due to a sparse measurement grid leading to poor interpolation of the templates.

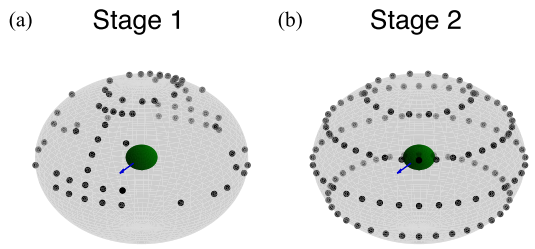


FIGURE 2. HRTF directions used for Stage 1 (a) and Stage 2 (b) evaluation. The central sphere with an arrow represents the position of the head, pointing forwards.

TABLE 2. Threshold Values for Stage 1 Evaluation Metrics

Metric		Difference threshold
Bias	lateral	5.86°
	local polar	12.67°
RMS error	lateral	20.71°
	local polar	5.90°
Quadrant error		34.56 pp[a]
Polar gain		0.33

[a]Percentage points.

To minimize metric variability between different grids in the task dataset, the error metrics were calculated from the model estimates for the 72 positions, closest to the *Club Fritz* common grid positions within each HRTF grid (Fig. 2(a)). Participants had to guarantee that at least 64 (80%) of their 80 harmonized HRTFs had localization metric differences below the thresholds for their submission to be passed to the second evaluation stage. Otherwise, the submission would be disqualified from the challenge.

Stage 1 evaluation was performed using MATLAB scripts, available to participants.[4] In this way, the participants were able to independently perform this evaluation for the original and harmonized HRTFs and assess their solutions.

2) STAGE 2

The second stage evaluated the effectiveness of the harmonization based on the principle outlined in [15]: If HRTFs are perfectly harmonized, an ML classifier should not be able to find any dataset-specific differences between them. To assign a score to the harmonization, the classifiers were trained by trying to maximize the separability between classes (dataset names in this case). However, the interpretation of the results was reversed: a failing classifier meant a successful harmonization. Therefore, the chance level of 0.125 (due to the eight different classes) represented perfect harmonization, while a perfect classification accuracy of 1 represented no harmonization at all.

To obtain scientifically interesting results, care must be taken not to let a classifier produce a degenerate solution, for example, by distinguishing datasets based on the number of directions or any other obvious difference listed in Table 1. Hence, minimal harmonization was applied to the input HRTFs as a preprocessing step. This included resampling to the minimum rate (44.1 kHz) and truncating to the minimum HRIR duration (5.333 ms), as in [15]. The common grid of source directions was established between all datasets such that the corresponding HRIRs could be directly compared without interpolation. Only these 126 source positions, displayed in Fig. 2(b), were kept as input to the classifiers.[5]

The classification performance was determined by an ensemble of models consisting of support vector machine classifiers (SVC) with linear and radial basis function (RBF) kernels, logistic regression, and a decision tree classifier. Each of these classifiers was run on HRIRs in the time domain and on the corresponding HRTFs in the magnitude domain, with the latter expressed both logarithmically (in decibels) and in linear units. These three input representations of the HRTF were provided to give the classifiers more opportunity to discover class-identifying features, since these classic ML algorithms are not able to learn feature representations (unlike more sophisticated deep learning techniques). Although a gradient-boosted tree classifier has been shown to be more performant than these simple classifiers [15], it was not used in this configuration due to its slower runtime. Its improved performance was deemed unnecessary, because the described configuration already achieved a perfect classification accuracy of 1 for the original HRTF files.

The 12 combinations of 4 classifiers and 3 input representations were trained and evaluated using five-fold cross-validation, resulting in an 80% /20% training/testing split of HRTFs in each fold. The combination with the highest accuracy in the training split was selected and its accuracy in the validation split used for the evaluation metric computation. The average accuracy over the five folds provided the output of a single run of the evaluator. To reduce inherent stochasticity due to random initialization of the classifiers, the evaluation process was run 10 times for each submission and the mean accuracy returned as their final score.

The Stage 2 evaluation code was provided to participants as a Python package for individual testing. Its source code is available online.[6]

IV. OVERVIEW OF SUBMISSION METHODOLOGIES

We received three submissions addressing the HRTF harmonization challenge. Here is a brief summary of the methods submitted, based on the information from the accompanying reports:

- *CoWiDeq:* The proposed approach is based on signal processing techniques and consists of three steps. Firstly,

[4][Online]. Available: https://github.com/Audio-Experience-Design/LAPChallenge

[5]The number of common source positions and their regular distribution over the sphere was the main selection criterion for the eight HRTF datasets. Note that it differs from the 10 datasets and 12 common source positions in [15] to ensure a better coverage of the sphere.

[6][Online]. Available: https://github.com/jpauwels/lap-task1

light dynamic compression is applied to reduce timbral variations. Then, each HRIR is windowed to 128 samples. Finally, HRTFs are diffuse-field equalized by spatially weighted RMS average HRTF, smoothed by modeling it as a second-order IIR filter series.

- *EqPCA-UoA:* The solution is based on the Principal Component Analysis (PCA). Firstly, the HRIRs are preprocessed by resampling, normalizing the RMS levels, finding the minimum-phase representations, zero-padding the filters to equal lengths, and retaining phase information for reconstruction. A matrix containing all HRTFs is then decomposed using PCA, and each HRTF is individually reconstructed by incrementally adding PCA components until the Stage 1 evaluation thresholds or a 95% reconstruction cap is met.
- *IOA3D:* The method generates harmonized HRTFs from binaural random samples using a neural network module and employs a perceptually relevant loss function to constrain both localization performance and classification accuracy. The network consists of three fully connected layers, each followed by batch normalization and ReLU activation. The perceptually relevant loss function is defined as a weighted sum of separate components: RMS difference to constrain ILD, spectral distortion within equivalent rectangular banwidth (ERB) frequency bands to constrain localization perception, and classification error to mitigate cross-dataset differences.

V. SUBMISION EVALUATION RESULTS

This section presents the results of the two-stage evaluation procedure as it was performed for the submission ranking and the challenge awards ceremony. Further analysis of the submissions is presented in the following sections.

Fig. 3 shows the differences in the estimated localization metrics between the original and harmonized HRTFs for each submission. The Stage 1 validation thresholds are also presented as dashed lines. While most estimated metric differences are similar across the three methods, the IOA3D HRTFs appear to have larger differences in lateral RMS errors (panel (c)) and quadrant errors (panel (e)). On the other hand, CoWiDeq method (and EqPCA-UoA to some extent) produces a few outliers with much higher differences in lateral bias (panel (a)), lateral RMS error (panel (c)), polar RMS error (panel (d)), and polar gain (panel (f)), especially within the RIEC dataset. Despite these minor variations, all three solutions passed the Stage 1 validation criteria by keeping at least 80% of the HRTFs below the thresholds across all six metrics. In total, 76 IOA3D, 72 CoWiDeq, and 69 EqPCA-UoA HRTFs (out of 80) fell below the thresholds in all modeled error metrics.

The second stage of evaluation is visualized in Fig. 4. For each submission, the figure shows the classification accuracy across all combinations of input and classifier types. The data from all ten iterations are presented, and each data point represents the mean accuracy of the individual five-fold

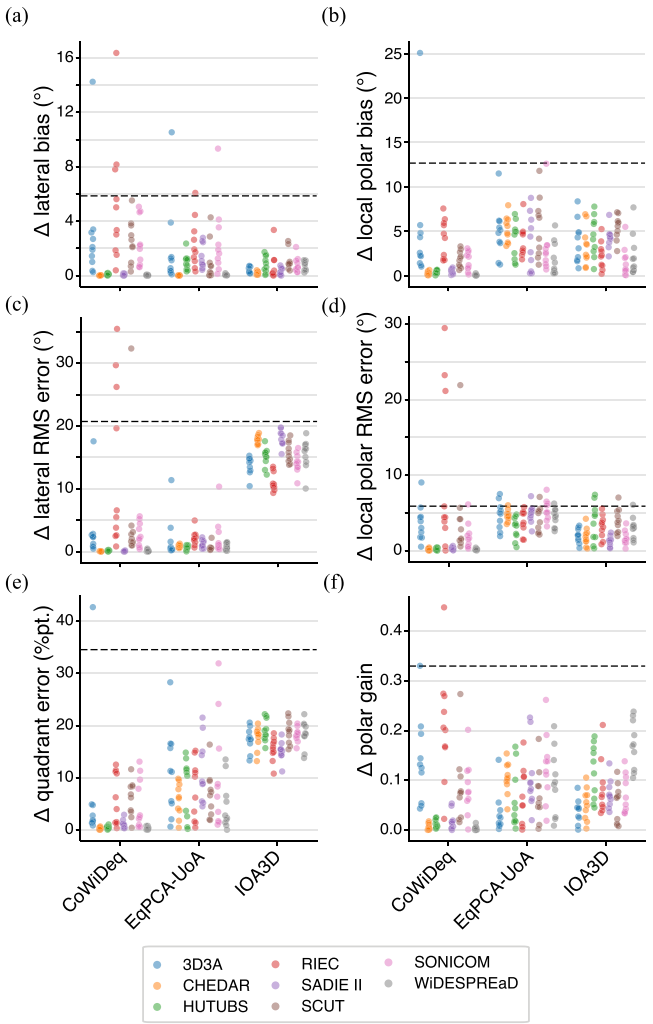


FIGURE 3. Estimated differences in lateral bias (a), local polar bias (b), lateral RMS error (c), local polar RMS error (d), quadrant error (e), and polar gain (f) between original and harmonized HRTFs using Stage 1 auditory model. Dashed lines represent metric thresholds for Stage 1 validation.

cross-validated grid search run. The best model from each iteration is highlighted with larger circular markers. The final accuracy score, which is calculated as the mean accuracy across those ten accuracy scores, is presented as a dashed horizontal line. The figure shows a diverse range of classifier accuracy scores that depend on the solution and the input features. The plot suggests that no one combination was universally robust against the varied range of modifications applied to the HRTFs. Panel (a) of the figure presents the classifier results with the original HRTFs. In that case, most of the results approach a perfect accuracy score of 1. The CoWiDeq solution (panel (b)) leads to a more varied classifier performance. All classifiers failed to distinguish between HRTF datasets when using the logarithmic magnitude response input, leading to chance-level performances. This might have been caused by a few faulty CoWiDeq HRTFs, causing convergence issues for some of the classifiers. However, linear SVC and decision tree classifiers were able to successfully identify most

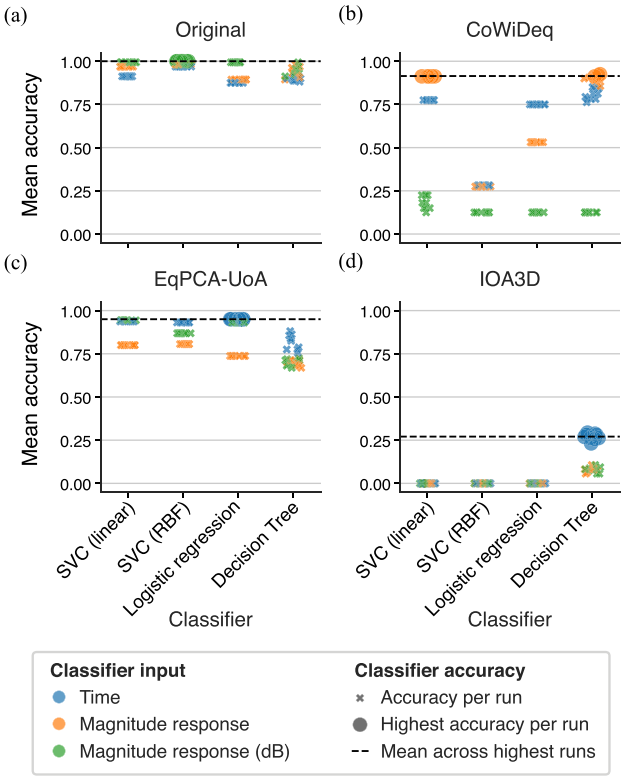


FIGURE 4. Accuracy scores of Stage 2 classifiers from 10 iterations of the five-fold cross-validated grid search for original HRTFs (a), and HRTFs harmonized using CoWiDeq (b), EqPCA-UoA (c), and IOA3D (d) methods. Combinations of classifier and input types that returned the highest accuracy score in each iteration (shown in circles) were averaged to obtain the final Stage 2 evaluation score (dashed line).

HRTF classes from linear magnitude responses, with a mean classification accuracy of 0.92. EqPCA-UoA solution (panel (c)) was less successful in tricking the classifiers, so most of them achieved accuracy above 0.75. The leading accuracy score was reached by logistic regression using time-domain input. Its mean accuracy score over the ten iterations was 0.95. Finally, the IOA3D solution (panel (d)) resulted in the lowest accuracy scores. Generally, the classifiers could not distinguish between different HRTF classes. The highest scorer was the decision tree classifier with time-domain input. It reached a mean accuracy of 0.27, a significantly lower score than the other two submissions.

All three methods managed to reduce the classifier accuracy to some extent while maintaining some of the HRTF information, according to the auditory model. The IOA3D solution was deemed the winner of the challenge according to the evaluation criteria. However, a more detailed analysis revealed flaws in the winning solution that are discussed in the following sections.

VI. FURTHER RESULT ANALYSIS

Despite the positive Stage 1 perceptual validation results for all three submitted solutions, informal listening tests with some of the harmonized HRTFs revealed a significant qualitative degradation of binaural rendering. This effect of the proposed methods on HRTFs was further numerically analyzed to understand the limitations and shortcomings of the current challenge setup.

Fig. 5(a) and (b) present example magnitude spectra along the median and horizontal planes of the original left ear HRTFs and the corresponding harmonized HRTFs from the three submissions. The examples include two acoustically measured HRTFs (SONICOM and HUTUBS) and two simulated ones (CHEDAR and WiDESPREaD). The figure shows a heterogeneous effect that each of the three proposed methods has on the spectral content of the HRTFs. A clear degradation of HRTF features is visible in two of the three submissions (EqPCA-UoA and IOA3D). The EqPCA-UoA method appears to blur sharp frequency-dependent notches in the HRTF magnitude spectra, while the magnitude spectra of the IOA3D solutions contain noise at the positions used by the Stage 2 evaluation (Fig. 2(b)). This is shown more clearly in Fig. 5(c), which presents the magnitude spectra of the example IOA3D HRTF (corresponding to the original SONICOM HRTF) at the Stage 2 grid positions only. The figure shows the logarithmic magnitude response input to the classifiers, where each panel represents one of the four horizontal planes at different elevations, comprising the grid shown in Fig. 2(b). The figure reveals that when the positions not used by the classifier are removed, the harmonized HRTF spectra lose all the typical HRTF characteristics, such as the frequency-dependent notches, visible in the original HRTF spectra (Fig. 5(b)).

Finally, Fig. 6 illustrates absolute ITDs and ILDs for the original and harmonized HRTF examples along the horizontal plane. These binaural features were derived from the HRIRs using the same methodology as the Stage 1 auditory model. The ITD was determined from the delay of the peak in the interaural cross-correlation of HRIR energy envelopes [67]. ILD was calculated as the difference in the RMS level between the HRIRs of the left and right ears. The ITD plots (panel (a)) reveal minimal differences between the approaches. This is to be expected, since the CoWiDeq procedure did not directly modify the ITDs while both the EqPCA-UoA and IOA3D methods extracted them before other modifications and reintroduced them subsequently. In contrast, the ILD graphs (panel (b)) demonstrated more pronounced variations. Some anomalies can be observed in EqPCA-UoA HRTFs, notably with the SONICOM HRTF. However, the IOA3D method significantly alters the original ILD pattern, especially nullifying the ILDs at positions associated with the Stage 2 evaluation grid. This is particularly evident with the HUTUBS HRTF whose ILD information is completely removed, because its horizontal plane only includes positions from the Stage 2 grid positions.

The IOA3D solution appears to take advantage of the flaw in the challenge setup that arises from the use of two evaluation grids. To confirm this hypothesis and better understand the effect of the chosen grid on Stage 1 evaluation, the localization metrics were also estimated using the Stage 2 grid, as well as a grid of 25 positions, common between the two grids. In these scenarios, the CoWiDeq method passed the validation

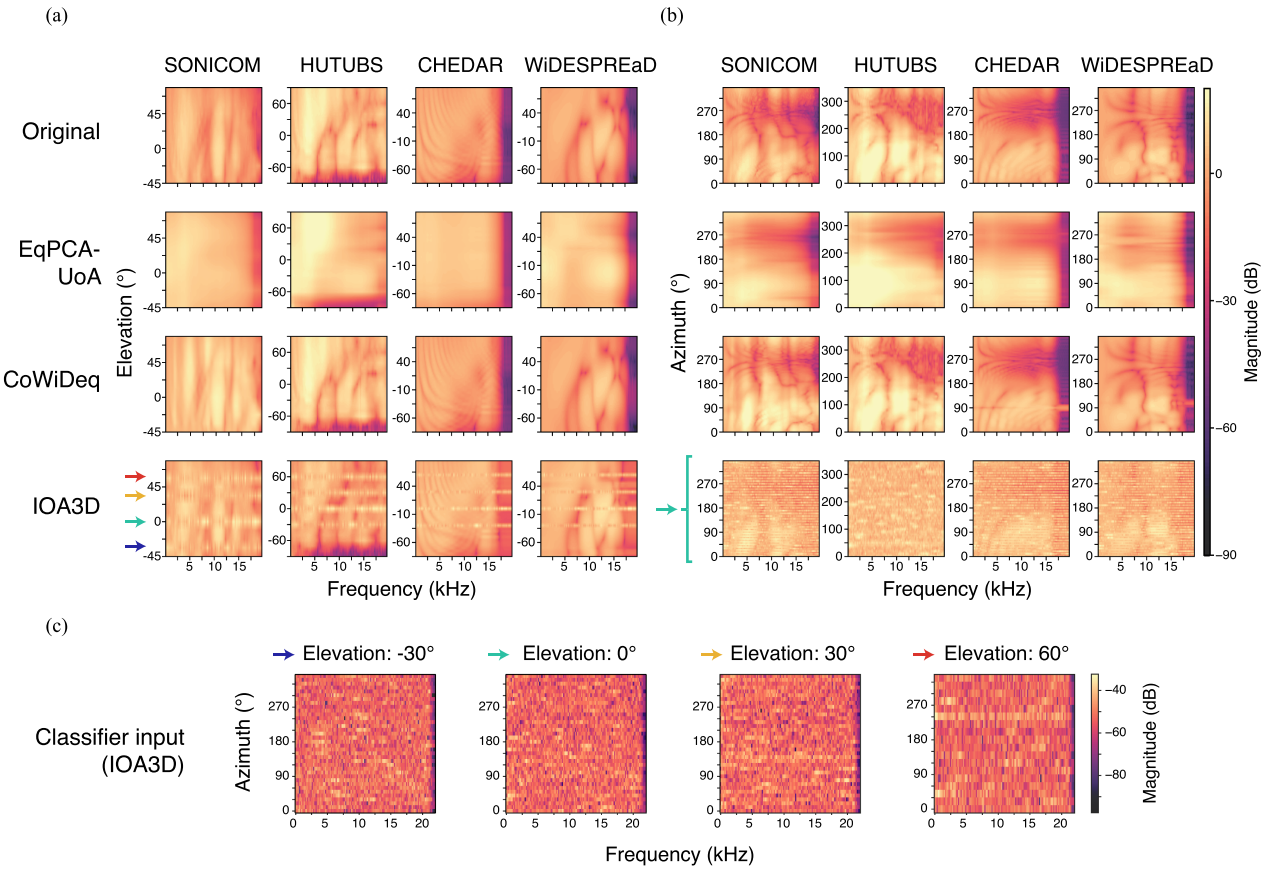


FIGURE 5. (a-b) Example original and harmonized left ear HRTF magnitude spectra for positions on the median (a) and horizontal (b) plane from four of the challenge HRTFs. (c) Example logarithmic magnitude response input to the Stage 2 classifier of a harmonized SONICOM HRTF using IOA3D method. Four panels contain HRTF positions across four horizontal planes comprising the full Stage 2 grid (Fig. 2(b)), indicated by the arrows in the full SONICOM HRTF spectra (a, bottom left). The 'Elevation: 0° ' panel in (c) is also a subset of the full horizontal plane spectra (b, bottom left).

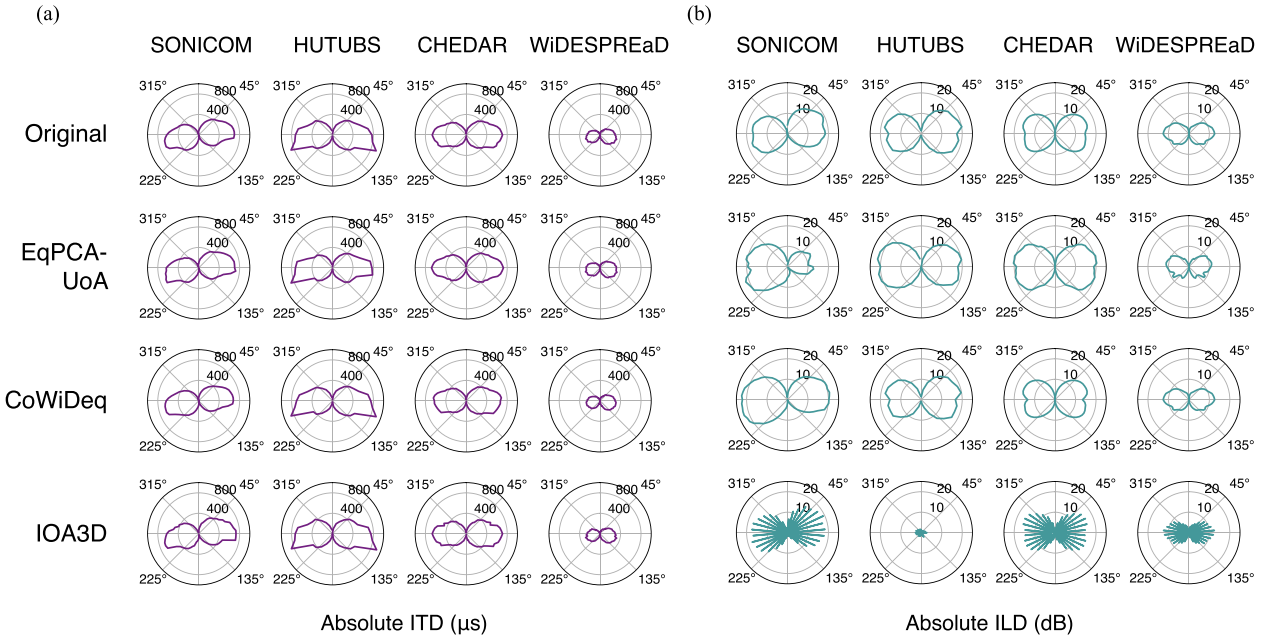


FIGURE 6. Absolute ITDs (a) and ILDs (b) of example original and harmonized HRTFs for azimuth angles along the horizontal plane.

thresholds, regardless of the grid used: 71 of 80 CoWiDeq HRTFs fell below the thresholds when using the Stage 2 grid and 64 were below when the common grid was used. However, the other two solutions did not meet the Stage 1 criteria when a different grid was used for the evaluation. For the EqPCA-UoA solution, only 19 HRTFs passed the criteria with the Stage 2 grid and 48 with the common grid, while 0 IOA3D HRTFs passed the thresholds using the Stage 2 grid and 1 with the common grid.

VII. DISCUSSION

The challenge evaluation integrated the numerical and perceptual aspects of HRTF research, with the aim of a robust assessment of the submissions that reflected the multifaceted definition of the HRTF harmonization procedure. Nonetheless, the plotted magnitude responses and additional result analysis demonstrate limitations of the current challenge structure. Our work highlights the complex nature of the HRTF harmonization process that can be summarized by the following polarities:

- acoustic vs. perceptual HRTF definitions;
- interpretable vs. closed-box evaluation methods;
- global (averaged) vs. local (direction-dependent) HRTF evaluation strategies.

These dualities must be addressed in order to improve the LAP challenge framework. They thus underlie further discussion on the challenge evaluation outcomes.

A. STAGE 1 EVALUATION

The use of auditory models marked an initial step toward automating the evaluation of harmonized HRTFs, but the challenge results demonstrated that precisely engineered solutions can exploit model limitations. The magnitude response figures suggest that the Stage 1 evaluation is not sufficiently sensitive to the degradation of HRTF magnitude spectra. The blurring of the sharp frequency-dependent notch pattern observed in the EqPCA-UoA and IOA3D HRTFs, although potentially imperceptible in some contexts of horizontal localization [68], is likely to impair the perception of sound source elevation, as it removes the main notches associated with vertical sound localization [69]. Yet, the averaging of the error metrics over the full sphere and large thresholds made it impossible for our model-based evaluation to detect such frequency blurring. Although the thresholds were based on the maximum estimated variability across real measurements, their values for some of the metric differences appear to be very high. For comparison, [20] reports an increase of less than 30 percentage points (pp) in quadrant errors and less than 10° increase in lateral RMS error for most subjects, localizing with non-individual HRTFs compared to individual ones. The loose Stage 1 thresholds (Table 2) suggest that a numerically engineered but perceptually invalid solution (e.g. replacing all HRTFs with one non-individual HRTF) could still pass the Stage 1 thresholds. Furthermore, considering that an average quadrant error in a real localization experiment with individual HRTFs can be around 20% [65], the quadrant error difference threshold above 30 pp allows for a harmonized HRTF with an estimated quadrant error above the chance rate (50%) to pass the criterion for this metric.

Initially, a more strict set of thresholds was imposed for the Stage 1 assessment, corresponding to the third quartiles of the estimated metric differences within the *Club Fritz* collection. However, the thresholds were increased from the third quartile of the metrics differences obtained from the round-robin comparison to the maximum metrics difference after participant feedback, suggesting that perceptually insignificant processing (such as low-passing the HRTFs at 18 kHz) would make harmonized HRTFs fail the Stage 1 assessment. This evidence highlights the limitations of using a model as a surrogate for perceptual tests when applied outside of its intended conditions. Designed for broadband noise sources and realistic individual HRTFs, the model accurately reproduces listener localization patterns by applying equal weighting to all spectral cues between 700 Hz and 18 kHz [50]. Deviations from these conditions expose the limitations of the model, which could be mitigated by incorporating spectral weighting functions to better reflect vertical localization patterns [70], [71]. While the model provides a useful tool for automating HRTF harmonization validation, the results reveal its limited view of auditory perception processes, which means that it can be easily misled by engineered modifications.

A potential solution to improve Stage 1 threshold sensitivity involves rethinking the two-step evaluation process. Rather than relying on a binary role of the Stage 1 validation, the perceptual quality of the HRTF could contribute to a combined final metric. This metric would synthesize the results from both stages. For each stage, a normalized representative metric could be derived and then aggregated, either by applying equal weights, akin to the ICASSP 2023 Clarity Challenge, that balanced the importance of speech intelligibility and audio quality criteria [53], or using a more intricate weighting strategy that aligns with the priorities of the challenge. This could result in methods such as IOA3D being penalized due to predicted increases in localization errors, as seen in Fig. 3(c) and (e), despite their good performance according to Stage 2 evaluation.

B. STAGE 2 EVALUATION

While the Stage 1 evaluation was designed to offer an interpretable metric, Stage 2 evaluation offered a straightforward numerical evaluation that served the purposes of the challenge, yet lacked interpretability of its results. The traditional ML techniques used in the evaluation process are fairly simple, which allows for some degree of insight when examining the fitted models, particularly for decision tree classifiers [15]. Nonetheless, transforming this analysis into a format relevant to the perceptual context and extracting meaningful dataset-specific features that classifiers use to distinguish between HRTF datasets remains a challenge. A future research avenue would be to find a more interpretable model for evaluating HRTF harmonization. Besides providing a sanity check for

the classifier performance, its results would aid in understanding the perceptual significance of the disparities among the datasets. Additionally, it could function as a reverse engineering tool, supporting the development of dataset-specific adjustments based on the model's outputs.

In addition to improving interpretability, future enhancements to the classifier should ensure its resilience against deliberately corruptive HRTF alterations. The authors of the CoWiDeq solution observed that the classifier's current accuracy could be compromised by introducing pure tone signatures to random HRTFs. They recommended exploring metrics based on perceptual evaluation. Similarly, the IOA3D solution was more tailored towards deceiving the classifier than aligning HRTFs in a perceptual way. Nevertheless, identifying a suitable perceptual metric or input features for ML classifiers requires further investigation, as any proposed data processing might mask undesirable artifacts and could be exploited by proposed harmonization solutions. Generally, the Stage 2 metric should be considered alongside the Stage 1 evaluation, which specifically targets perceptual HRTF quality. A comprehensive set of evaluation metrics would thus be more effective in identifying intentional perceptual HRTF degradation for the purpose of reducing its detectability.

C. EVALUATION GRIDS

A crucial aspect in designing Task 1 was determining which spatial grid to use for evaluating the submitted solutions. Stage 1 relied on a grid aligned with the *Club Fritz* collection, targeting perceptually relevant angular sectors (e.g. near the median plane) where human behavioral responses are especially sensitive to spatial cues [20]. In contrast, Stage 2 employed a uniform sampling of the sphere to promote geometric consistency and avoid spatial interpolation across datasets for the classifier. As a result, each approach prioritized different factors (i.e. perceptual relevance versus uniform acoustic coverage) leading to an inconsistency in how HRTFs were evaluated across the two stages.

With the evaluation of the submissions, the grid discrepancy exposed two main issues. First, there was no unified requirement to maintain spatial consistency in HRTF harmonization across the entire sphere. Second, evaluating harmonized HRTFs over the sphere proved inadequate for capturing local anomalies (e.g., if the distribution of errors is highly variable or skewed, a simple mean can mask substantial inaccuracies in specific regions). Even with a unified grid, behavioral localization metrics often cover a limited range of positions, typically within 30° of the median plane for polar-based metrics, leaving many directions unexamined [20]. This gap allowed ML models to exploit unused portions of the grid, as demonstrated by the IOA3D solution: By injecting random noise specifically at the Stage 2 positions and retaining perceptually relevant features where Stage 1 metrics were measured, IOA3D passed Stage 1 despite producing larger errors in other directions. Meanwhile, globally consistent approaches such as CoWiDeq and EqPCA-UoA appeared to be more robust, but still performed inconsistently when Stage 1 evaluation was conducted using different grids. Consequently, a supposedly *global* solution does not necessarily guarantee local coherence and can be exploited by methods optimized for particular subsets of directions.

In the future, caution is needed to balance perceptual and acoustic considerations on spatial grids in the harmonization process, reflecting the chosen HRTF definition. If the goal is to preserve perceptual relevance, metrics might focus on directions where human responses are most accurate or behaviorally critical, while still accounting for broader coverage. For instance, a more detailed analysis of quadrant confusion types [72], compensation for polar angle compression [73], or the probabilistic representation of behavioral responses [74] could improve the robustness of the evaluation. In parallel, ML-based evaluation might consider weighting errors, depending on spatial cue variability in the region [75] and adopt methods to ensure both local and global consistency. By unifying these perspectives, perceptual and acoustic, future versions of the challenge could effectively address the discrepancies that allowed methods such as IOA3D to selectively optimize for one stage at the expense of the other.

D. FUTURE AVENUES FOR HRTF HARMONIZATION

The LAP24 Task 1 served as an important step in systematizing the HRTF harmonization problem and merging perceptually informed HRTF evaluation approaches with ML methods. The work generated valuable discussions and highlighted key limitations that will help refine future approaches and advance the field. The difficulties associated with defining a good evaluation metric for this challenge reflect a broader challenge with the reliability of the metrics used as targets [76]. In this context, solutions tend to be optimized for higher metric scores, rather than addressing the challenge as a whole. This is particularly true for ML models, which, driven by numeric loss functions, can exploit evaluation loopholes. Nevertheless, we believe that organizing multiple iterations of the challenge with refined evaluation could lead to an improved HRTF harmonization method, stress-testing its metrics during each run. A cyclic procedure, similar to the series of "Clarity enhancement and prediction challenges" [7], could be considered, in which a challenge to find the best metric to assess harmonization would be organized in turn with a challenge to find the best harmonization procedure.

While listening tests should ultimately be included in challenge evaluation, their design requires further consideration. A fully crossed study, where each participant undergoes HRTF measurements at multiple facilities and evaluates all their harmonized versions, would be financially unfeasible, especially considering a high inter- and intra-subject result variability, requiring a large sample size. Such evaluation would also lack the repeatability needed for rapid, iterative solution testing by participating teams. As an alternative, an initial screening phase using listening tests could be added to the current evaluation setup to ensure a minimum qualitative

standard without requiring individual HRTF measurements. However, defining the minimum perceptual validity criteria based on such data remains a challenge.

Generally, several concepts should be clarified before proceeding with the next challenge. Firstly, a clearer task-oriented *narrow* definition must be established before defining the improved set of metrics. The current iteration focused on localization accuracy; however, more qualitative aspects, such as externalization or timbral coloration, could also be accounted for, potentially by using already existing models [77], [78]. In addition, the choice of the task dataset needs to be examined. The current selection included both measured and modeled HRTF datasets, featuring WiDESPREaD, which only simulated pinnae transfer functions. The challenge aimed to incorporate a diverse range of HRTFs to explore the possibility of consolidating them into a unified HRTF dataset. However, pinnae-only responses fundamentally differ from full HRIRs, which is evident from considerably smaller ITDs and ILDs, shown in Fig. 6. Therefore, aligning them with other HRTFs while maintaining relative perceptual similarity to their original versions might be an ill-posed task and fall outside the scope of what would typically be considered harmonization in this context. Overall, given the scarcity of research in the HRTF harmonization domain to date, considerations like this present an opportunity to establish HRTF harmonization baselines that would be useful for a wider scientific community.

VIII. CONCLUSION

In the LAP24 Challenge, the HRTF dataset harmonization task provided an opportunity for valuable theoretical and practical discussions on the definition of HRTF and the extent of its perceptual validity, addressing a gap within current HRTF personalization research. Aiming to identify and resolve problems when consolidating data for data-driven HRTF personalization methods, the challenge asked participants to harmonize HRTFs from different datasets in such a way that a trained ML classifier would not be able to distinguish their origin. The challenge also imposed a constraint of perceptual validity on harmonized HRTFs to ensure the *narrow* HRTF definition is retained. However, the proposed solutions of the three participating teams revealed weaknesses in the evaluation, prompting a thorough review of the evaluation setup and identifying key areas for future research to address and improve. The next iteration of the challenge should consider the use of the common evaluation grid for both the perceptual evaluation and the classification stage, and a new set of perceptual evaluation metrics to account for that grid, with more robust validation thresholds. Furthermore, enhancing the interpretability of the HRTF classifier output is necessary to strengthen its reliability and our understanding of systematic HRTF differences. Establishing strong evaluation criteria for HRTF harmonization may require multiple iterations, as ML methods can take advantage of their weaknesses, but each step would contribute to the advancement of our understanding of the HRTF personalization domain.

ACKNOWLEDGMENT

The authors would like to express sincere gratitude to the members of the teams who participated in the challenge and submitted their solutions: Hélène Bahu, Thibaut Carpentier, Markus Noisternig, Olivier Warusfel, and Jean-Marc Jot (CoWiDeq); Camilo Arevalo and Julián Villegas (EqPCA-UoA); and Jiale Zhao, Dingding Yao, Zelin Qiu, Chengzhong Wang, and Junfeng Li (IOA3D). We also thank Fabian Brinkmann, Glen McLachlan, Aidan O. T. Hogg, and Stavros Ntalampiras who helped us organize the challenge. Finally, we thank U-Sound and Dreamwaves, our industrial partners who sponsored the award ceremony, held during the 32nd European Signal Processing Conference (EUSIPCO 24). The LAP Challenge was sponsored by the IEEE Signal Processing Society.

REFERENCES

[1] M. Geronazzo and S. Serafin, Eds., *Sonic Interactions in Virtual Environments (Human-Computer Interaction Series)*. Cham, Switzerland: Springer, 2023, doi: 10.1007/978-3-031-04021-4.

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

[3] C. Guezenoc and R. Seguier, "HRTF individualization: A survey," in *Proc. Audio Eng. Soc. Conv.*, New York, NY, USA, 2018, Art. no. 10129. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=19855

[4] L. Picinali and B. F. G. Katz, "System-to-user and user-to-system adaptations in binaural audio," in *Sonic Interactions in Virtual Environments*, M. Geronazzo and S. Serafin, Eds. Berlin, Germany: Springer, 2023, pp. 115–143, doi: 10.1007/978-3-031-04021-4_4.

[5] D. Fantini, M. Geronazzo, F. Avanzini, and S. Ntalampiras, "A survey on machine learning techniques for head-related transfer function individualization," *IEEE Open J. Signal Process.*, vol. 6, pp. 30–56, 2025, doi: 10.1109/OJSP.2025.3528330.

[6] H. W. Löllmann et al., "The LOCATA challenge data corpus for acoustic source localization and tracking," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop*, 2018, pp. 410–414, doi: 10.1109/SAM.2018.8448644.

[7] S. Graetzer et al., "Clarity: Machine learning challenges to revolutionise hearing device processing," 2020, *arXiv:2006.11140*.

[8] A. O. T. Hogg et al., "Listener acoustic personalisation challenge - LAP24: Head-related transfer function upsampling," *IEEE Open J. Signal Process.*, early access, Jul. 14, 2025, doi: 10.1109/OJSP.2025.3588776.

[9] B. Zhi, D. N. Zotkin, and R. Duraiswami, "Towards fast and convenient end-to-end HRTF personalization," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 441–445, doi: 10.1109/ICASSP43922.2022.9746315.

[10] D. Yao et al., "An individualization approach for head-related transfer function in arbitrary directions based on deep learning," *JASA Exp. Lett.*, vol. 2, no. 6, 2022, Art. no. 064401, doi: 10.1121/10.0011575.

[11] A. O. T. Hogg, M. Jenkins, H. Liu, I. Squires, S. J. Cooper, and L. Picinali, "HRTF upsampling with a generative adversarial network using a gnomonic equiangular projection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2085–2099, 2024, doi: 10.1109/TASLP.2024.3375635.

[12] A. Hogg, H. Liu, M. Jenkins, and L. Picinali, "Exploring the impact of transfer learning on GAN-based HRTF upsampling," in *Proc. Conv. EAA Forum Acusticum*, Turin, Italy, 2023, pp. 2323–2328, doi: 10.61782/fa.2023.0266.

[13] L. Arbel, I. Ananthabhotla, Z. Ben-Hur, D. L. Alon, and B. Rafaely, "On HRTF notch frequency prediction using anthropometric features and neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Seoul, South Korea, 2024, pp. 816–820, doi: 10.1109/ICASSP48485.2024.10447610.

[14] A. Andreopoulou, D. R. Begault, and B. F. G. Katz, "Inter-laboratory round robin HRTF measurement comparison," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 895–906, Aug. 2015, doi: 10.1109/JSTSP.2015.2400417.

[15] J. Pauwels and L. Picinali, "On the relevance of the differences between HRTF measurement setups for machine learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Rhodes, Greece, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10096689.

[16] Y. Wen, Y. Zhang, and Z. Duan, "Mitigating cross-database differences for learning unified HRTF representation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, 2023, pp. 1–5, doi: 10.1109/WASPAA58266.2023.10248178.

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

[18] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *J. Audio Eng. Soc.*, vol. 43, no. 5, pp. 300–321, 1995.

[19] H. Møller, "Fundamentals of binaural technology," *Appl. Acoust.*, vol. 36, no. 3, pp. 171–218, 1992, doi: 10.1016/0003-682X(92)90046-U.

[20] 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, doi: 10.1121/1.427147.

[21] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database," *Appl. Sci.*, vol. 8, no. 11, 2018, Art. no. 2029, doi: 10.3390/app8112029.

[22] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses," *J. Audio Eng. Soc.*, vol. 67, no. 9, pp. 705–718, 2019, doi: 10.17743/jaes.2019.0024.

[23] S. Li and J. Peissig, "Measurement of head-related transfer functions: A review," *Appl. Sci.*, vol. 10, no. 14, 2020, Art. no. 5014, doi: 10.3390/app10145014.

[24] A. Andreopoulou and A. Roginska, "Towards the creation of a standardized HRTF repository," in *Proc. Audio Eng. Soc. Conv.*, New York, NY, USA, 2011, Art. no. 8571. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=16096

[25] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Gaussian process data fusion for heterogeneous HRTF datasets," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, 2013, pp. 1–4, doi: 10.1109/WASPAA.2013.6701842.

[26] X. Lu, Y. Wang, J. Sang, and C. Zheng, "BiCG: Binaural cue generation from unified HRTF datasets," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Hyderabad, India, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10890756.

[27] S. T. Prepeliă, J. G. Bolaños, M. Geronazzo, R. Mehra, and L. Savioja, "Pinna-related transfer functions and lossless wave equation using finite-difference methods: Verification and asymptotic solution," *J. Acoust. Soc. Amer.*, vol. 146, no. 5, pp. 3629–3645, 2019, doi: 10.1121/1.5131245.

[28] S. T. Prepeliă, J. G. Bolaños, M. Geronazzo, R. Mehra, and L. Savioja, "Pinna-related transfer functions and lossless wave equation using finite-difference methods: Validation with measurements," *J. Acoust. Soc. Amer.*, vol. 147, no. 5, pp. 3631–3645, 2020, doi: 10.1121/10.0001230.

[29] B. Goertzel, "Artificial general intelligence: Concept, state of the art, and future prospects," *J. Artif. Gen. Intell.*, vol. 5, no. 1, pp. 1–48, 2014, doi: 10.2478/jagi-2014-0001.

[30] W. L. Oberkampf and T. G. Trucano, "Verification and validation in computational fluid dynamics," *Prog. Aerosp. Sci.*, vol. 38, no. 3, pp. 209–272, 2002, doi: 10.1016/S0376-0421(02)00005-2.

[31] M. Morimoto and Y. Ando, "On the simulation of sound localization," *J. Acoust. Soc. Jpn.*, vol. 1, no. 3, pp. 167–174, 1980, doi: 10.1250/ast.1.167.

[32] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. II: Psychophysical validation," *J. Acoust. Soc. Amer.*, vol. 85, no. 2, pp. 868–878, 1989, doi: 10.1121/1.397558.

[33] E. H. A. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 528–537, 1999, doi: 10.1121/1.428321.

[34] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," *J. Audio Eng. Soc.*, vol. 55, no. 7/8, pp. 623–637, 2007.

[35] G. Enzner, "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 325–328, doi: 10.1109/ASPAA.2009.5346532.

[36] P. Dietrich, B. Masiero, and M. Vorländer, "On the optimization of the multiple exponential sweep method," *J. Audio Eng. Soc.*, vol. 61, no. 3, pp. 113–124, 2013.

[37] P. J. Roache, *Verification and Validation in Computational Science and Engineering*. Albuquerque, NM, USA: Hermosa, 1998.

[38] W. L. Oberkampf and C. J. Roy, *Verification and Validation in Scientific Computing*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[39] S. Klockgether and S. Van De Par, "Just noticeable differences of spatial cues in echoic and anechoic acoustical environments," *J. Acoust. Soc. Amer.*, vol. 140, no. 4, pp. EL352–EL357, 2016, doi: 10.1121/1.4964844.

[40] J. E. Mossop and J. F. Culling, "Lateralization of large interaural delays," *J. Acoust. Soc. Amer.*, vol. 104, no. 3, pp. 1574–1579, 1998, doi: 10.1121/1.424369.

[41] B. C. J. Moore, S. R. Oldfield, and G. J. Dooley, "Detection and discrimination of spectral peaks and notches at 1 and 8 kHz," *J. Acoust. Soc. Amer.*, vol. 85, no. 2, pp. 820–836, 1989, doi: 10.1121/1.397554.

[42] 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, doi: 10.1109/TASL.2012.2227730.

[43] C. C. Berger, M. Gonzalez-Franco, A. Tajadura-Jiménez, D. Florencio, and Z. Zhang, "Generic HRTFs may be good enough in virtual reality. Improving source localization through cross-modal plasticity," *Front. Neurosci.*, vol. 12, 2018, Art. no. 21, doi: 10.3389/fnins.2018.00021.

[44] L. S. R. Simon, N. Zacharov, and B. F. G. Katz, "Perceptual attributes for the comparison of head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 140, no. 5, pp. 3623–3632, 2016, doi: 10.1121/1.4966115.

[45] C. Jenny and C. Reuter, "Usability of individualized head-related transfer functions in virtual reality: Empirical study with perceptual attributes in sagittal plane sound localization," *JMIR Serious Games*, vol. 8, no. 3, 2020, Art. no. e17576, doi: 10.2196/17576.

[46] A. Andreopoulou and B. F. G. Katz, "Investigation on subjective HRTF rating repeatability," in *Proc. Audio Eng. Soc. Conv.*, Paris, France, 2016, Art. no. 9597. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=18295

[47] C. Kim, V. Lim, and L. Picinali, "Investigation into consistency of subjective and objective perceptual selection of non-individual head-related transfer functions," *J. Audio Eng. Soc.*, vol. 68, no. 11, pp. 819–831, 2020, doi: 10.17743/jaes.2020.0053.

[48] 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. 1247–1260, Jul. 2018, doi: 10.1109/TASLP.2018.2821846.

[49] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *J. Acoust. Soc. Amer.*, vol. 136, no. 2, pp. 791–802, 2014, doi: 10.1121/1.4887447.

[50] R. Barumerli, P. Majdak, M. Geronazzo, D. Meijer, F. Avanzini, and R. Baumgartner, "A Bayesian model for human directional localization of broadband static sound sources," *Acta Acustica*, vol. 7, 2023, Art. no. 12, doi: 10.1051/aacus/2023006.

[51] R. Barumerli, M. Geronazzo, and F. Avanzini, "Round robin comparison of inter-laboratory HRTF measurements–assessment with an auditory model for elevation," in *Proc. IEEE 4th VR Workshop Sonic Interact. Virtual Environ.*, 2018, pp. 1–5, doi: 10.1109/SIVE.2018.8577091.

[52] 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, 2014, Art. no. Art. no. 319, doi: 10.3389/fpsyg.2014.00319.

[53] T. J. Cox et al., "Overview of the 2023 ICASSP SP clarity challenge: Speech enhancement for hearing aids," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Rhodes, Greece, 2023, pp. 1–2, doi: 10.1109/ICASSP49357.2023.10433922.

[54] G. R. Dabike et al., "The first Cadenza challenges: Using machine learning competitions to improve music for listeners with a hearing loss," in *IEEE Open J. of Signal Process.*, vol. 6, pp. 722–734, 2025, doi: 10.1109/OJSP.2025.3578299.

[55] R. Sridhar, J. G. Tylka, and E. Y. Choueiri, "A database of head-related transfer function and morphological measurements," in *Proc. Audio Eng. Soc. Conv.*, New York, NY, USA, 2017, Art. no. 357. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=19308

[56] S. Ghorbal, X. Bonjour, and R. Séguier, "Computed hrirs and ears database for acoustic research," in *Proc. Audio Eng. Soc. Conv.*, 2020, Art. no. 13061. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=20778

[57] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, "Dataset of head-related transfer functions measured with a circular loudspeaker array," *Acoust. Sci. Technol.*, vol. 35, no. 3, pp. 159–165, 2014, doi: 10.1250/ast.35.159.

[58] G. Yu, R. Wu, Y. Liu, and B. Xie, "Near-field head-related transfer-function measurement and database of human subjects," *J. Acoust. Soc. Amer.*, vol. 143, no. 3, pp. EL194–EL198, 2018, doi: 10.1121/1.5027019.

[59] I. Engel et al., "The SONICOM HRTF dataset," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 241–253, 2023, doi: 10.17743/jaes.2022.0066.

[60] C. Guezenoc and R. Séguier, "A wide dataset of ear shapes and pinna-related transfer functions generated by random ear drawings," *J. Acoust. Soc. Amer.*, vol. 147, no. 6, pp. 4087–4096, 2020, doi: 10.1121/10.0001461.

[61] Acoustics Research Institute, "SOFA (spatially oriented format for acoustics)," 2025. Accessed: Jan. 8, 2025. [Online]. Available: https://www.sofaconventions.org/mediawiki/index.php?title=Files&oldid=2624

[62] P. Majdak, F. Zotter, F. Brinkmann, J. D. Muynke, M. Mihocic, and M. Noisternig, "Spatially oriented format for acoustics 2.1: Introduction and recent advances," *J. Audio Eng. Soc.*, vol. 70, no. 7/8, pp. 565–584, 2022, doi: 10.17743/jaes.2022.0026.

[63] R. Ege, A. J. V. Opstal, and M. M. Van Wanrooij, "Accuracy-precision trade-off in human sound localisation," *Sci. Rep.*, vol. 8, 2018, Art. no. 16399, doi: 10.1038/s41598-018-34512-6.

[64] R. Daugintis, R. Barumerli, L. Picinali, and M. Geronazzo, "Classifying non-individual head-related transfer functions with a computational auditory model: Calibration and metrics," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Rhodes, Greece, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10095152.

[65] P. Majdak, M. J. Goupell, and B. Laback, "3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Attention Percep. Psychophys.*, vol. 72, no. 2, pp. 454–469, 2010, doi: 10.3758/APP.72.2.454.

[66] E. A. Macpherson and J. C. Middlebrooks, "Localization of brief sounds: Effects of level and background noise," *J. Acoust. Soc. Amer.*, vol. 108, no. 4, pp. 1834–1849, 2000, doi: 10.1121/1.1310196.

[67] A. Andreopoulou and B. F. G. Katz, "Identification of perceptually relevant methods of inter-aural time difference estimation," *J. Acoust. Soc. Amer.*, vol. 142, no. 2, pp. 588–598, 2017, doi: 10.1121/1.4996457.

[68] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, no. 6713, pp. 747–749, 1998, doi: 10.1038/25526.

[69] E. H. A. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *J. Acoust. Soc. Amer.*, vol. 112, no. 4, pp. 1583–1596, 2002, doi: 10.1121/1.1501901.

[70] B. Zonooz, E. Arani, K. P. Körding, P. A. T. R. Aalbers, T. Celikel, and A. J. Van Opstal, "Spectral weighting underlies perceived sound elevation," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 1642, doi: 10.1038/s41598-018-37537-z.

[71] P. Lladó, P. Majdak, R. Barumerli, and R. Baumgartner, "Spectral weighting of monaural cues for auditory localization in sagittal planes," *Trends Hear.*, vol. 29, pp. 1–14, 2025, doi: 10.1177/23312165251317027.

[72] D. Poirier-Quinot, M. S. Lawless, P. Stitt, and B. F. G. Katz, "HRTF performance evaluation: Methodology and metrics for localisation accuracy and learning assessment," in *Advances in Fundamental and Applied Research on Spatial Audio*, B. F.G. Katz and P. Majdak, Eds. London, U.K.: IntechOpen, 2022, doi: 10.5772/intechopen.104931.

[73] S. Carlile, K. Balachandar, and H. Kelly, "Accommodating to new ears: The effects of sensory and sensory-motor feedback," *J. Acoust. Soc. Amer.*, vol. 135, no. 4, pp. 2002–2011, 2014, doi: 10.1121/1.4868369.

[74] R. Barumerli and P. Majdak, "FrAMBI: A software framework for auditory modeling based on Bayesian inference," *Neuroinformatics*, vol. 23, no. 2, Feb. 2025, Art. no. 20, doi: 10.1007/s12021-024-09702-5.

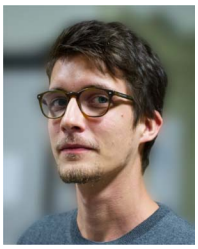
[75] J. C. Middlebrooks, "Narrow-band sound localization related to external ear acoustics," *J. Acoust. Soc. Amer.*, vol. 92, no. 5, pp. 2607–2624, 1992, doi: 10.1121/1.404400.

[76] J. Rodamar, "There ought to be a law! Campbell versus Goodhart," *Significance*, vol. 15, no. 6, p. 9, 2018, doi: 10.1111/j.1740-9713.2018.01205.x.

[77] R. Baumgartner and P. Majdak, "Decision making in auditory externalization perception: Model predictions for static conditions," *Acta Acustica*, vol. 5, 2021, Art. no. 59, doi: 10.1051/aacus/2021053.

[78] T. McKenzie and F. Brinkmann, "Toward an improved auditory model for predicting binaural coloration," *J. Audio Eng. Soc.*, vol. 73, no. 3, pp. 115–126, 2025, doi: 10.17743/jaes.2022.0192.

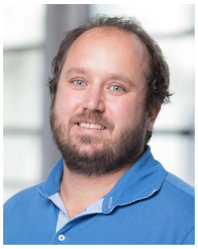
RAPOLAS DAUGINTIS received the M.Sc. degree in acoustics and audio technology from Aalto University, Espoo, Finland, in 2021. He is currently working toward the Ph.D. degree with Audio Experience Design team at the Dyson School of Design Engineering, Imperial College London, London, U.K. He is investigating perceptual differences of head-related transfer functions and the relevance of binaural spatial audio personalization in various extended reality scenarios. His work includes collaborations with members of the EU-funded SONICOM Project, which aims to improve immersive audio experiences for extended reality applications. His research interests include psychoacoustics, spatial audio, auditory and acoustic modeling, and extended reality.



ROBERTO BARUMERLI received the M.Sc. degree in 2016, and the Ph.D. degree in information and communication technology from the University of Padova, Padova, Italy, in 2021, focusing on computational auditory models for sound localization. He was a Postdoctoral Researcher at the Acoustics Research Institute, Vienna, Austria. He is currently a Postdoctoral Researcher at the University of Verona, Verona, Italy. His research interests include dynamics of auditory spatial perception through Bayesian inference and probabilistic modeling, integrating signal processing with behavioral and physiological data analysis. He specializes in experimental design, computational statistics, and auditory model development, applying these techniques to neuroscience, acoustics, and human-computer interaction. Up-to-date information on his research is available here: robertobarumerli.com.

MICHELE GERONAZZO (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in 2009 and 2014, respectively. Between 2014 and 2021, he was an Assistant Professor at the University of Udine, Udine, Italy and a Postdoctoral Researcher at Aalborg University, Copenhagen, Denmark. He is currently an Associate Professor at the University of Padova, Padova, Italy, and part of the coordination unit of the EU-H2020 Project SONICOM with Imperial College London, London, U.K. He is a co-author of more than 90 scientific publications. His main research interests involve modeling and simulation of complex human-machine sonic interactions. He has been part of the organizing committee of the IEEE VR Workshop on Sonic Interactions for Virtual Environments since 2015 (chair of the 2018, 2020, and 2022 editions). He is an Associate Editor for the IEEE OPEN JOURNAL OF SIGNAL PROCESSING, *ACM Transactions on Applied Perception*, and Editor of *Sonic Interactions in Virtual Environments* (Springer-Nature). He is a co-recipient of six Best Paper/Poster Awards.

JOHAN PAUWELS received the M.Sc. degree in electrical engineering and a second M.Sc. degree in artificial intelligence from KU Leuven, Leuven, Belgium, in 2006 and 2007, respectively. He received the Ph.D. degree in automatic harmony recognition from audio, from Ghent University, Ghent, Belgium, in 2016. He has held research positions at IRCAM, Queen Mary University of London, and Imperial College London. He is currently a Lecturer with the Centre for Digital Music at Queen Mary University of London, London, U.K. His main research focuses on AI and signal processing applied to music and immersive audio.



LORENZO PICINALI received the M.Sc. and Ph.D. degrees in 2013 and 2010, respectively. He is currently a Professor of spatial acoustics and immersive audio at Imperial College London, London, U.K., where he leads the Audio Experience Design team (www.axdesign.co.uk). His research explores both the perceptual and computational aspects of acoustics, alongside their practical applications. The aim is to understand how humans perceive spatial sound features, such as source locations and reverberation, and use this knowledge to design algorithms and tools for creating virtual acoustic simulations. His main research goal is to seamlessly blend real and virtual acoustic environments, ultimately enhancing hearing. He also works on eco-acoustic monitoring, designing autonomous recorders, and using audio to study the impact of human activity on remote ecosystems.

KATARINA C. POOLE received the M.Sc. degree in 2017 and the Ph.D. degree in auditory neuroscience from University College London (UCL), London, U.K., in 2023, where she investigated how the brain extracts auditory patterns using methodologies such as psychophysics, electrophysiology, and optogenetics. She is currently a Research Associate in immersive audio with Audio Experience Design team at the Dyson School of Design Engineering, Imperial College London, London, U.K. Her work includes collaboration with members of the EU-funded SONICOM Project to improve immersive audio experiences and collaboration with UCL and the Eriksholm Research Centre to understand non-target spatial hearing in normal and hearing-impaired listeners. Her current research focuses on assessing numerical and perceptual differences between measured and synthetic head-related transfer functions.