forum acusticum 2026
8-12 September · Graz, Austria

LISTENER ACOUSTIC PERSONALISATION CHALLENGE LAP24: PERCEPTUAL EVALUATION OF HRTF UPSAMPLING

Fabian Brinkmann[1], Anton Hoyer[1], Aidan O. T. Hogg[2], Lorenzo Picinali[3], Michele Geronazzo[4], Stefan Weinzierl[1]

[1] *Audio Communication Group, Technische Universität Berlin, Germany*
[2] *Centre for Digital Music, Queen Mary University of London, UK*
[3] *Audio Experience Design, Dyson School of Design Engineering, Imperial College London, UK*
[4] *Department of Engineering and Management, University of Padova, Italy*

Abstract

The 2024 Listener Acoustic Personalisation (LAP24) challenge benchmarked the spatial upsampling (interpolation) of head-related transfer functions (HRTFs). Seven teams submitted HRTFs that were upsampled from four sparse grids containing between three and one hundred HRTF positions. The submissions included six deep learning-based approaches and one algorithmic approach. They were evaluated with respect to log-spectral distortion (LSD) and broadband interaural time and level differences (ITD, ILD). The learning-based upsampling methods often showed smaller differences to the reference HRTF than the algorithmic approach, especially for very sparse sampling grids. In the current study, we conducted a complementary perceptual evaluation of the LAP24 challenge algorithms with respect to colouration and source direction for a static sound source. This work confirmed the numerical results of the LAP Challenge at least for the top-ranked method. Beyond first place, however, the perceptual ranking differed from that obtained with physical error measures, due to the different aspects captured and their sensitivities. Moreover, an algorithm-specific diffuse field equalisation filter had little effect, suggesting that direction-dependent upsampling errors remain audible even in this case. The upsampled HRTFs, along with audio examples and ratings from the listening test, are publicly available to facilitate benchmarking of future upsampling algorithms.

Keywords: *head-related transfer function, upsampling, interpolation*

***Corresponding author**: fabian.brinkmann@tu-berlin.de.*

1. Introduction

HRTFs contain monaural and interaural cues that the human auditory system interprets for determining the position of a sound source. Based on these cues, listeners derive an internal representation of the surrounding three-dimensional auditory space, including information about the number and position of sources [1]. HRTFs can thus be used to render virtual sources and are fundamental for headphone-based audio in extended reality contexts. Some applications can highly benefit from individual HRTFs to convey the most natural sound rendering, including realistic sound colour and source position [2], [3, Fig. 5.15]. However, specialised equipment is needed to acquire individual HRTFs [4].

One way to reduce the acquisition effort is to measure HRTFs at only a few source positions and use interpolation—also termed spatial upsampling—to obtain HRTFs at arbitrary positions. However, this is a non-trivial task: interaural and monaural cues can change drastically even between neighbouring source positions, making HRTFs hard to interpolate. Furthermore, humans are highly sensitive to these cues, leaving little room for interpolation errors.

The LAP24 challenge comprised two tasks: HRTF dataset normalisation [5] and spatial upsampling [6]. In this work, we focus on the latter, where participants were invited to reconstruct high-spatial-resolution HRTFs from four sparse spatial grids. The submissions were benchmarked against densely sampled reference HRTFs with respect to LSD, ITD, and ILD. Seven teams submitted upsampling approaches, six of which used deep learning. The deep-learning approaches often outperformed the algorithmic submission, especially for very sparse grids containing 3 and 5 sources. However, this predominance was less clear for localisation errors computed with auditory models.

Building upon the LAP24 challenge, we conducted a

perceptual evaluation of the submitted HRTFs, which had been missing to date. This focused on a comparison of the various submissions based on two specific qualitative attributes. In addition, we published data of the LAP24 challenge submissions and evaluations, as well as of the current study, to provide a platform for benchmarking of future HRTF upsampling approaches.

2. Methods

The experiment assessed the perceived colouration and direction of upsampled HRTFs against the reference HRTF. The first part of the experiment tested different sparse sampling grids while the second part investigated the effect of HRTF equalisation.

2.1. Upsampled HRTFs

We considered all upsampled HRTFs that passed the validation step of the LAP24 challenge, and refer to them using the abbreviations from [6]. In ranked order and starting with the winning submission, these are: MERL_2 [7], IOA3D [8], SYT_FSP-AE [9], Kalimotxo [10], SUpDEq [11], and GEP_GAN [12]. All algorithms except SUpDEq used deep learning and were trained on the first 200 subjects from the SONICOM HRTF database [13]. Deep learning algorithms separately upsampled HRTF magnitude spectra and time of arrival (ToA), made interpolated magnitude spectra minimum phase and delayed them by the interpolated ToA. SUpDEq interpolated time-aligned complex HRTF spectra and reversed the alignment in post-processing. Note that MERL_1 was excluded from this study because it performed almost identically to MERL_2, which was submitted by the same team. Detailed descriptions of the algorithms are available from the references above.

2.2. Setup and conditions

Upsampled HRTFs were submitted on a discrete target sampling grid of 793 source positions with a resolution of 5° in azimuth and 10° in elevation for elevations above −45° [13]. HRTFs were upsampled from four sparse grids (SGs) containing subsets of 3, 5, 19, and 100 source positions. SG3 contained one source at 90° azimuth (left), one at 0° azimuth and elevation (front), and one at 90° elevation (above). SG5 comprised sources centred around the front of the listener: three in the horizontal plane at azimuths of ±45° and 0°, and two in the median plane at elevations of ±45°. SG19 had a resolution of 60° in azimuth at elevations of ±45° and 0°, plus one source directly above at 90° elevation. SG100 was created by taking every eighth position of the target grid.

In addition to the four SGs, we also tested two HRTF equalisations, i.e. we applied two different direction-independent correction filters to all tested HRTFs. The *reference* condition applied an identical diffuse-field equalisation to all methods. The equalisation filter was computed as the RMS average of the reference HRTF magnitude spectra across source positions separately for each ear, inverted with regularisation to avoid excessive gains above 16 kHz, and applied as a minimum-phase filter to the upsampled and reference HRTFs. In the *individual* condition, the equalisation filter was computed analogously, but using the upsampled HRTFs instead of the reference HRTF, resulting in a method-specific equalisation. The reference equalisation emphasises upsampling errors and represents the most critical evaluation, whereas individual equalisation is more common in consumer applications and thus better reflects practical use.

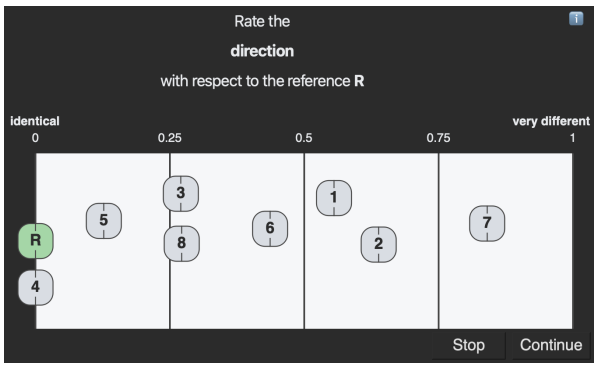

Figure 1: Drag & drop MUSHRA-like rating interface. In this example, condition **4** is rated to be identical to the reference **R** and condition **7** is rated to differ the most.

Upsampled HRTFs were rated against the reference HRTFs with respect to the attributes *colouration* (defined as the 'timbral impression which is determined by frequency-dependent loudness differences between two sounds') and *direction* (defined as the 'perceived direction of the sound source'). In both cases, a rating of 0 indicates that stimuli were perceived as identical to the reference, and a rating of 1 that they were perceived as very different. We selected the two attributes due to their importance for basic audio quality [14], with colouration denoting 'how' something sounds and direction describing from 'where' the sound is being perceived.

In addition to the upsampled HRTFs, we included a hidden reference and the average HRTF. The latter was obtained by taking the minimum-phase RMS average of the HRTF magnitude spectra across the first 200 subjects of the SONICOM HRTF database [13] separately for each source position. The minimum-phase spectra were delayed by the average ToA (computed using the *threshold* method [15] with a threshold of -20 dB) to achieve the desired ITDs. The average HRTF was already used in the LAP24 challenge as an additional intermediate benchmark, alongside the reference. This comparison shows whether the evaluated upsampling methods go beyond a direction-wise average over the dataset and effectively retain individual-specific features.

To facilitate comparison of the eight conditions (six upsampled HRTFs, SONICOM average, hidden reference), participants used a drag-and-drop MUSHRA-like rating interface (cf. [16] and Fig. 1). Clicking and holding a tile triggers audio playback and highlights

the active condition in green. Releasing the tile automatically plays the reference (labelled **R**). Ratings are assigned by dragging a tile into the rating area, where the horizontal position determines the rating on the displayed scale, while the vertical position has no effect. The interface was implemented with `PyQt` (v6.11.0) and required participants to listen to each condition at least once.

We computed static binaural stimuli for a source at 30° azimuth and 30° elevation, and frozen pink noise pulses, and used the Python `sounddevice` (v0.5.5) package for audio playback. The source position is particularly challenging, as it lies outside SG3 and SG5, at distances of 41° and 28° from the nearest source position, respectively. Pink noise pulses (200 ms noise, 100 ms silence, 2 ms fade-in/out) were chosen because their broadband spectrum and temporal transients are suitable to reveal subtle differences in colouration and direction. All signal processing was performed with `pyfar` (v0.8.0). Dynamic renderings involving head or source movements would have required additional interpolation and were therefore omitted.

2.3. Procedure

After downloading the test files, participants followed instructions in an interactive Python notebook to install the required packages and conduct the test in a self-administered manner. They were instructed to wear studio-quality headphones, conduct the test in a quiet environment, carefully read the attribute definitions, focus on one attribute at a time, and listen to the stimuli in any order and as often as desired. A training phase, consisting of a single rating screen with four conditions spanning the expected range of differences, allowed participants to adjust the playback volume to a comfortable level, explore the interface, and familiarise themselves with the conditions.

In the first part of the experiment, participants assessed the effect of sparsity and upsampling approach across eight rating screens (2 attributes × 4 SGs) using the reference HRTF equalisation. All upsampling approaches for a given SG were presented on a single screen in randomised order. The order of attributes and SGs was also randomised, but all four SGs were completed in one block per attribute. In the second part, participants rated HRTFs with individual equalisation across two rating screens (2 attributes × SG3). We used SG3 only, because we expected upsampling errors to be largest here, and hence the effect of the equalisation to be strongest. The equalisation was always assessed last, to ensure participants were accustomed to the range of differences and were able to put the differences for individual equalisation into perspective.

When the attribute to be rated changed, an intermediate screen displayed its name and definition to inform participants. A blue info button allowed participants to revisit the definition at any time. At the end, participants completed a questionnaire reporting, among other aspects, their headphones, self-assessed understanding of the attributes, and whether environmental noise disturbed their ratings.

2.4. Statistical analyses

Ratings were analysed using repeated-measures linear mixed models in GAMLj3 v3.6.5 [17], with method, sparsity, and HRTF equalisation as fixed effects, and random intercepts for participants. Variance was estimated using REML, and post-hoc comparisons were Holm-corrected. Normality of residuals was assessed visually via Q-Q plots.

3. Results

The following discusses the listening test results. Due to page restrictions, statistical details are provided in the digital appendix linked in Section 5.

3.1. Participants

Twenty-five participants took part in the test (mean age 29 years, SD 5.5; 21 male, 2 female and 2 gender-diverse). All were staff and students of the authors' institutes. All but one reported having participated in at least one listening experiment before and having actively listened to sound or music for 2 hours per day over the past month. No one reported a diagnosed hearing impairment, nor did they report that background noise disturbed the ratings. Twenty-four reported that the attributes were clear to them, while one was unsure. The test, including instructions and training, took an average of 27 minutes. Data from a rating screen were discarded if a participant rated the hidden reference at least 0.1 (10% of the rating scale) worse than a test condition, thereby discarding 4.8% of the total ratings.

3.2. Effect of sparsity

The model for colouration explains $R^2_{\text{cond.}} = 75.2\%$ ($p < .001$) of the variance if considering fixed and random effects, and $R^2_{\text{marg.}} = 64.5\%$ ($p < .001$) if considering fixed effects only. The method explains the highest amount of variance with an adjusted partial eta squared of $\eta^2_p = 0.7$, followed by the interaction method × sparsity $\eta^2_p = 0.24$ and sparsity $\eta^2_p = 0.08$ (all $p < .001$). The model for direction explains less variance ($R^2_{\text{cond.}} = 61.9\%$, $R^2_{\text{marg.}} = 27.7\%$, $p < .001$) and shows the same order of main effects and interactions (method: $\eta^2_p = 0.38$, method × sparsity $\eta^2_p = 0.11$, sparsity $\eta^2_p = 0.04$, all $p < .001$).

The estimated marginal means (EMMs) for method in Table 1 illustrate the main effect of method, with MERL_2 and Kalimotxo outperforming all other methods for both tested attributes. These are also the only methods that yield better overall ratings than the average HRTF. The remaining methods perform less consistently across colouration and direction, with SYT_FSP-AE showing the largest colouration and GEP_GAN the most inaccurate direction. Informal listening suggested that the high colouration ratings for SYT_FSP-AE were due to a pronounced low-frequency colouration that was almost independent of the SG.

Table 1 also shows overall EMMs averaged across at-

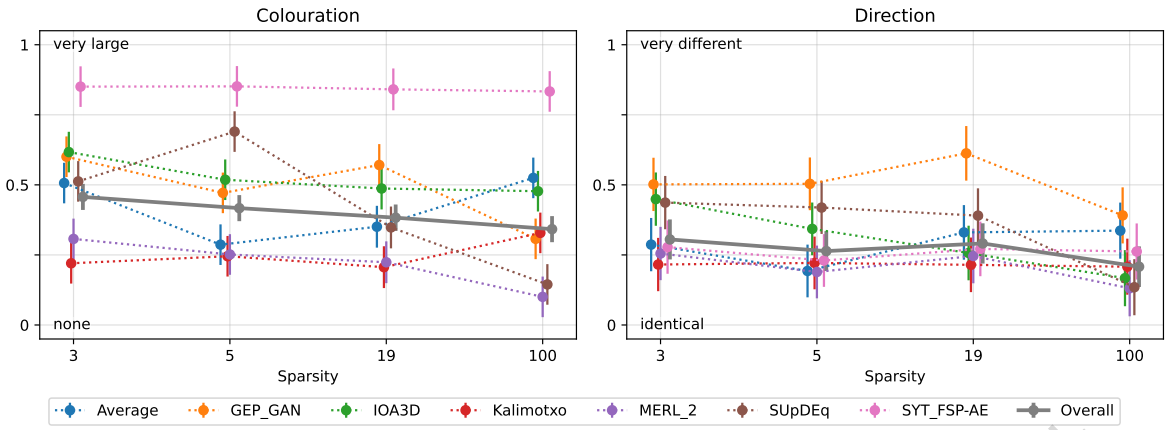

Figure 2: EMMs for method and sparsity with 95% confidence intervals of their interaction. The *overall* data refers to the EMMs for the main effect of sparsity pooled across methods.

Table 1: EMMs for method and the linear mixed model analysing the effect of sparsity. Overall means were averaged across colouration and direction. Numbers in parentheses denote how many ranks a method placed better (+) or worse (-) compared to the official LAP24 challenge ranking [6, Table 2].

Method	Overall	Colour.	Direct.
MERL_2 (±0)	0.21	0.22	0.21
Kalimotxo (+2)	0.23	0.25	0.21
Average HRTF	0.35	0.42	0.29
SUpDEq (+2)	0.38	0.42	0.35
IOA3D (−2)	0.41	0.52	0.30
GEP_GAN (+1)	0.49	0.49	0.50
SYT_FSP-AE (−3)	0.55	0.84	0.26

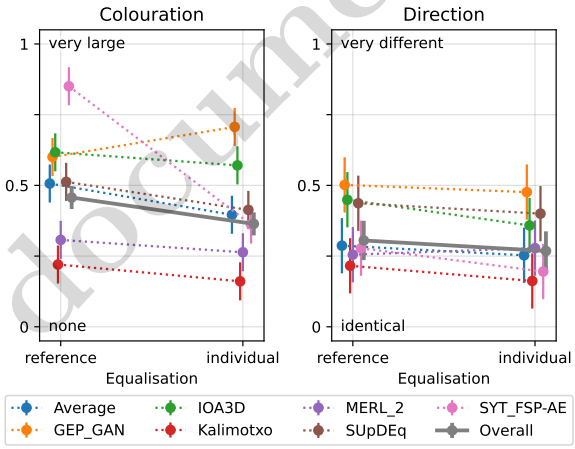

Figure 3: EMMs for method and equalisation with 95% confidence intervals of their interaction. The *overall* data refers to the EMMs for the main effect of equalisation pooled across methods.

tributes. Colouration dominates this measure, because EMMs for colouration are larger than for direction. Although we did not assess the relative importance of the two attributes, the greater observed colouration differences are consistent with the importance of timbral fidelity reported by Rumsey et al. [14]. Comparing the ranking from the LAP24 challenge [6, Table 2] with that established by the overall EMMs shows that only MERL_2 ranked identically in both cases, and that four of the remaining five methods climbed or fell by at least two ranks.

The estimated marginal means for sparsity are given in Fig. 2. For colouration, all differences between successive SGs (e.g. SG3 vs. SG5) are significant with $p < .05$, indicating a decrease in colouration with decreasing sparsity (increasing density). For direction, differences are only significant between SG19 and SG100 ($p < .001$), and marginally significant between SG3 and SG5 ($p = .054$).

Due to the significant interaction of method × sparsity, the main effects have limited validity and Fig. 2 shows the inconsistent behaviour of the methods across SGs. For colouration (Fig. 2, left), only MERL_2 and IOA3D have monotonically decreasing EMMs. For most other methods, EMMs tend to decrease as sparsity decreases, except for SYT_FSP-AE and Kalimotxo, where they are almost constant or increase. For direction (Fig. 2, right), IOA3D and SUpDEq show monotonically decreasing EMMs. For both attributes, the average HRTF does not exhibit a clear tendency of EMMs across SGs. This is to be expected, because the LAP24 challenge used HRTFs of different subjects for each SG to avoid providing additional data that could have been used for overfitting.

3.3. Effect of HRTF equalisation

More variance is explained by the model for colouration ($R^2_{\text{cond.}} = 72.7\%$, $R^2_{\text{marg.}} = 63.7\%$, $p < .001$) than by the model for direction ($R^2_{\text{cond.}} = 55.8\%$, $R^2_{\text{marg.}} = 25.5\%$, $p < .001$). In both cases, the method has the largest effect, with $\eta_p^2 = 0.67$ for colouration and $\eta_p^2 = 0.37$ for direction. The effect of equalisation is small in both cases (cf. grey lines in Fig. 3).

Colouration decreases by only 0.1 if applying method-specific equalisation, and errors in source direction decrease only marginally by 0.04.

For both attributes, we observed a significant method × equalisation interaction. However, disordinal interactions only appear for colouration, where GEP_GAN is the only method for which colouration increases due to method-specific equalisation. In addition, the effect is by far the most pronounced for SYT_FSP-AE, which shows a decrease in colouration of nearly 0.5. Informal listening suggested that the method-specific equalisation removed the low-frequency artefacts that presumably caused the high colouration ratings in the first place.

4. Discussion

We conducted the previously missing perceptual evaluation of HRTF upsampling approaches that took part in the LAP24 challenge by assessing the colouration and direction of upsampled HRTFs against the reference HRTF.

4.1. Comparison to LAP24

The first part of the experiment aimed to facilitate a comparison between the physical evaluation from the LAP24 challenge and the perceptual evaluation conducted in this study. Results in Table 1 (first column) show that rankings based on physical and perceptual measures differ for at least two reasons: First, the LAP24 challenge used LSD and broadband ITD and ILD to evaluate the algorithms, which are relatively simple physical error measures. While ITD and ILD may predict horizontal sound direction [15], [18], LSD was shown to be a rather poor predictor for colouration [19] and is usually not used to predict vertical sound direction. Hence, the error measures used in the LAP24 challenge might have limited perceptual validity. Second, our perceptual evaluation considered only a single source at a challenging position outside SG3 and SG5. This off-grid source position directly assesses the intended purpose of upsampling, although the results may vary for other source positions.

4.2. Directional colouration

We assessed the influence of HRTF equalisation in the second part of the experiment and found only a small influence on the tested attributes. While it might be encouraging that differences in source direction were almost unaffected by direction-independent equalisation, this suggests that direction-dependent colouration persists regardless of the equalisation. Because the listening test for this study was conducted in a self-administered manner using different headphones and various listening environments, it warrants further investigation in a fully controlled laboratory setting for clarification.

4.3. Benchmark for HRTF upsampling

We aimed to enable benchmarking of future HRTF upsampling algorithms by publishing data from the LAP24 challenge and the current study (cf. Section 5). If future studies want to make use of these data, we recommend including at least the best- and worst-performing methods for each attribute (cf. Table 1) in future listening tests to ensure comparability with respect to existing methods. Moreover, we welcome data contributions to make HRTFs that were upsampled with future approaches available to other researchers.

4.4. Limitations

Testing only a single source position constitutes a limitation of the study, and results may not generalise to other positions. Since we chose a challenging source position, it might be hypothesised that differences between algorithms decrease for other positions. We also did not assess dynamic rendering conditions, including moving sources and listeners, that are relevant to interactive real-time extended reality scenarios with three or six degrees of freedom. Future evaluations could, for example, include a moving source to assess whether hallucinations of deep learning-based upsampling induce audible artefacts.

To emphasise differences between upsampling approaches, we used anechoic stimuli and a comparison to an explicit reference. While it is known that adding reverberation and comparing to an implicit reference may decrease the range of perceived differences [20], [21], we considered these effects to be out of scope for the current study. Upsampling algorithms developed after the LAP24 challenge are also considered out of scope, but could now be evaluated using the data listed in Section 5.

Moreover, participants conducted the listening test unsupervised, using their own headphones and listening environment, which may have introduced additional between-subject variance and attenuated the observed differences between conditions compared with a fully controlled setting.

5. Data Availability

The upsampled and reference HRTFs, the audio examples from the listening test, the code to generate them, the perceptual ratings, and their detailed statistical analyses are available from `https://github.com/f-brinkmann/LAP24_task2_perceptual`.

References

[1] P. Majdak, R. Baumgartner, and C. Jenny, "Formation of Three-Dimensional Auditory Space," in *The Technology of Binaural Understanding*, ser. Modern Acoustics and Signal Processing, J. Blauert and J. Braasch, Eds., 1st ed., Cham, Switzerland: Springer / ASA Press, 2020, pp. 115–149.

[2] L. Picinali and B. F. Katz, "System-to-user and user-to-system adaptations in binaural audio," in *Sonic interactions in virtual environments*, Springer, 2022, pp. 115–143.

[3] F. Brinkmann and S. Weinzierl, "Audio Quality Assessment for Virtual Reality," in *Sonic Interactions in Virtual Environments*, ser. Human-

Computer Interaction Series, M. Geronazzo and S. Serafin, Eds., Cham, Switzerland: Springer, 2022, pp. 145–178. DOI: 10.1007/978-3-031-04021-4.

[4] 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., IntechOpen, Apr. 2022. DOI: 10.5772/intechopen.102908.

[5] R. Daugintis, R. Barumerli, M. Geronazzo, J. Pauwels, L. Picinali, and K. C. Poole, "Listener Acoustic Personalization Challenge—LAP24: Head-Related Transfer Function Dataset Harmonization," *IEEE Open J. Signal Process.*, vol. 6, pp. 950–964, 2025, ISSN: 2644-1322. DOI: 10.1109/OJSP.2025.3592601.

[6] A. O. T. Hogg et al., "Listener Acoustic Personalisation Challenge - LAP24: Head-Related Transfer Function Upsampling," *IEEE Open J. Signal Process.*, pp. 926–941, Jul. 2025. DOI: 10.1109/OJSP.2025.3588776.

[7] Y. Masuyama, G. Wichern, F. G. Germain, C. Ick, and J. Le Roux, "RANF: Neural Field-Based HRTF Spatial Upsampling With Retrieval Augmentation and Parameter Efficient Fine-Tuning," *IEEE Open J. Signal Process.*, vol. 7, pp. 32–41, 2026. DOI: 10.1109/OJSP.2025.3640517.

[8] J. Zhao, D. Yao, and J. Li, "Head-Related Transfer Function Upsampling With Spatial Extrapolation Features," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 33, pp. 1034–1048, 2025. DOI: 10.1109/TASLPRO.2025.3544080.

[9] Y. Ito, T. Nakamura, S. Koyama, and H. Saruwatari, "Head-Related Transfer Function Interpolation From Spatially Sparse Measurements Using Autoencoder With Source Position Conditioning," in *Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022, pp. 1–5. DOI: 10.1109/IWAENC53105.2022.9914751.

[10] C. Arevalo and J. Villegas, "Spatial Upsampling of Head-Related Impulse Responses via Elevation-Wise Encoder-Decoder Networks," *IEEE Open J. Signal Process.*, vol. 6, pp. 1086–1093, 2025. DOI: 10.1109/OJSP.2025.3613209.

[11] J. M. Arend, C. Pörschmann, S. Weinzierl, and F. Brinkmann, "Magnitude-Corrected and Time-Aligned Interpolation of Head-Related Transfer Functions," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3783–3799, Nov. 2023. DOI: 10.1109/TASLP.2023.3313908.

[12] 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.

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

[14] F. Rumsey, S. Zieliński, R. Kassier, and S. Bech, "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 968–976, Aug. 2005. DOI: 10.1121/1.1945368.

[15] A. Andreopoulou and B. F. G. Katz, "Identification of perceptually relevant methods of interaural time difference estimation," *J. Acoust. Soc. Am.*, vol. 142, no. 2, pp. 588–598, Aug. 2017. DOI: 10.1121/1.4996457.

[16] C. Völker, T. Bisitz, R. Huber, B. Kollmeier, and S. M. A. Ernst, "Modifications of the Multi stimulus test with Hidden Reference and Anchor (MUSHRA) for use in audiology," *Int. J. Audiology*, 2016. DOI: 10.1080/14992027.2016.1220680.

[17] M. Gallucci, *GAMLj3: GAMLj Suite for Linear Models (R package v.3.6.5)*, https://github.com/gamlj/gamlj, (accessed 2026-06-19), 2025.

[18] 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 Acust.*, vol. 7, p. 12, 2023. DOI: 10.1051/aacus/2023006.

[19] 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, Mar. 2025. DOI: 10.17743/jaes.2022.0192.

[20] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the authenticity of individual dynamic binaural synthesis," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1784–1795, Oct. 2017. DOI: 10.1121/1.5005606.

[21] N. Meyer-Kahlen, S. J. Schlecht, S. Amengual Garí, and T. Lokki, "Testing Auditory Illusions in Augmented Reality: Plausibility, Transfer-Plausibility, and Authenticity," *J. Audio Eng. Soc.*, vol. 72, no. 11, pp. 797–812, Nov. 2024. DOI: 10.17743/jaes.2022.0178.