OPEN ACCESS Freely available online

I. Engel, R. Daugintis, T. Vicente, A. O. T. Hogg,
J. Pauwels, A. J. Tournier and L. Picinali,
"The SONICOM HRTF Dataset"
J. Audio Eng. Soc., vol. 71, no. 5, pp. 241–253, (2023 May).
DOI: https://doi.org/10.17743/jaes.2022.0066

The SONICOM HRTF Dataset

ISAAC ENGEL, RAPOLAS DAUGINTIS, THIBAULT VICENTE, AIDAN O. T. HOGG,
(isaac.engel@huawei.com) (r.daugintis21@imperial.ac.uk) (t.vicente@imperial.ac.uk) (a.hogg@imperial.ac.uk)

JOHAN PAUWELS, ARNAUD J. TOURNIER, AND LORENZO PICINALI, *AES Member*
(j.pauwels@qmul.ac.uk) (arnaud.tournier17@imperial.ac.uk) (l.picinali@imperial.ac.uk)

Audio Experience Design (www.axdesign.co.uk), Imperial College London, London, United Kingdom

Immersive audio technologies, ranging from rendering spatialized sounds accurately to efficient room simulations, are vital to the success of augmented and virtual realities. To produce realistic sounds through headphones, the human body and head must both be taken into account. However, the measurement of the influence of the external human morphology on the sounds incoming to the ears, which is often referred to as head-related transfer function (HRTF), is expensive and time-consuming. Several datasets have been created over the years to help researchers work on immersive audio; nevertheless, the number of individuals involved and amount of data collected is often insufficient for modern machine-learning approaches. Here, the SONICOM HRTF dataset is introduced to facilitate reproducible research in immersive audio. This dataset contains the HRTF of 120 subjects, as well as headphone transfer functions; 3D scans of ears, heads, and torsos; and depth pictures at different angles around subjects' heads.

0 INTRODUCTION

Head-related transfer functions (HRTFs) are filters that characterize how both ears receive sound from a source in the surrounding space (see also [1]). They are specific to the morphology of a particular individual's external ear, head, shoulder, and torso, and once acoustically measured or numerically synthesized, they can be used for simulating 3D sound fields through a simple pair of headphones. This technique is generally referred to as binaural spatialization [2].

In the past years, several HRTF datasets have been created by various research institutions around the world using both human [3–11] and artificial heads [12–14, 6, 7, 15, 11, 16]. The number of human subjects in each of these publicly available datasets ranges between 10 and 220, with most of them containing less than 100 subjects. For a fraction of these human subjects, anthropometric data and a 3D scan of the subjects' ears and/or heads have also been collected alongside their HRTF [3, 5, 7, 9–11, 8].

However, combining these datasets in practice, e.g., for machine-learning applications [17], is not straightforward [18]. First, the speakers, microphones, and room characteristics vary across institutes. Second, the directions and distance of the source chosen by the researchers are inconsistent across datasets. Finally, and most importantly, the methods used to record these HRTFs have evolved over time, and different labs have chosen different techniques and options, from Golay-code signals [3] to exponential sweeps [10], from static loudspeaker arrays [3] to moving ones [8], and from two-way loudspeakers [10] to single-driver ones [8]. Because of these discrepancies, HRTFs are significantly different across these datasets, even for the same head [15].

Created within the remit of the SONICOM project [19],[1] this technical paper introduces the publicly-released SONICOM HRTF dataset, which aims at facilitating reproducible research (i.e., allowing researchers to ensure that they can repeat the same analysis multiple times with the same results [20]) in the spatial acoustics and immersive audio domain by including in a single database HRTFs measured from an increasingly large number of subjects (currently 120), as well as headphones transfer functions; 3D scans of ears, head, and torso; and RGB + depth pictures at different angles around the subject's head.

SEC. 1 introduces the hardware and software components, both off-the-shelf and custom-designed; this is followed by an overview of the measurement (SEC. 2) and post-processing (SEC. 3) procedures, looking in each section at various types of collected data [i.e., headphone transfer functions (HpTFs), HRTFs, photos, and 3D scans]. SEC. 4 presents the results from numerical and model-based perceptual evaluations on the measured data. Then, SEC. 5 summarizes the dataset features, providing information about the released data. Finally, SEC. 6 closes the paper, also looking at future developments.

[1] www.sonicom.eu.

Fig. 1. Measurement setup.

Table 1. Reverberation time (RT60) of the room per frequency band.

Frequency (Hz)	250	500	1,000	2,000	4,000	8,000
RT (ms)	110	110	60	60	60	70

1 HARDWARE AND SOFTWARE SETUP

The measurements take place in a room at Imperial College London. The room, dome-shaped with a square floor, measures 5 × 5 × 7.5 m and is acoustically treated with 150-mm foam wedges covering all walls, ceiling surfaces, and carpet on the floor. Its reverberation time (RT60) is indicated in Table 1. Acoustic doors and wall/floor insulation allow a background noise of 23-dB equivalent continuous A-weighted sound pressure level.

The room hosts an aluminum arch, on which loudspeakers were mounted such that their drivers are aligned along the arch (see Fig. 1). The rectangular speakers with off-center drivers are mounted horizontally because of size constraints, and their orientation is alternated for optimal weight distribution. The distance between the loudspeaker drivers and the center of the arch is 1.5 m. In its current iteration, the arch holds 23 loudspeakers between –45° and 225° of elevation, spaced every 10° between 30°/–30° and 150°/210° and spaced every 15° in the other positions.

This symmetric configuration allows for the completion of a complete HRTF measurement in half the time because the subject needs to be rotated only 180° in azimuth from the initial position. This allows for fast measurements (approximately 5 min) of HRTFs with an elevation resolution between 10° and 15°.

The loudspeakers (Wilmslow Audio Ltd, Leicester, United Kingdom) are custom one-way medium-density–fiberboard cabinets mounting a full-range Peerless 830987 3-in driver and operating in a frequency range between 100 Hz and 20 kHz. These are connected to two multichannel Triad TS-PAMP8-100 class-D amplifiers and a MOTU 24Ao audio interface. A free-field measurement is displayed in Fig. 2, which was performed with the same microphones used for the HRTF measurements and without the turntable or chair present. It can be seen from Fig. 2 that the effective frequency range of the system extends past 16 kHz, which is enough to include perceptually relevant spatial cues of the HRTF, and is in line with other HRTF databases [21].

A Tama First Chair drum seat with a backrest was placed along the vertical symmetric axis of the arch so that the subject's head could be positioned at the center of the arch by adjusting the chair height. The chair was mounted on a custom-made turntable operated by an Arduino UNO and a Nema 23 3-Nm stepper motor that can be controlled via open sound control messages through an ethernet cable (details about the turntable construction can be found in the dataset website—see the URL in SEC. 5). The turntable rotation error was measured to be smaller than 1° after a full-circle rotation in 5° increments; therefore, it was below any of the minimum audible angle thresholds [22]. Additionally, a custom-made chin rest (the 3D model is available for download from the dataset website) was mounted on the chair to help the subjects to remain still during the measurements.

For the acoustic measurements, a pair of Knowles FG-23329-P07 electret microphone capsules was used, phantom-powered through a couple of RØDE VXLR+ adapters. These capsules have been successfully employed for HRTF measurements in the past [21]. An RME Babyface Pro was used to record the input signals, using the MOTU interface as the master clock. A sampling rate of 96 kHz was used for all input and output audio signals. For each subject, a pair of SONICFOAM Memory Foam Earbud Tips were placed at the ear canal entrance to block the ear canal and provide housing for the microphones, as shown in Fig. 3.

In order to track the subject's position during the measurement, a tracking system based on infrared cameras (Op-

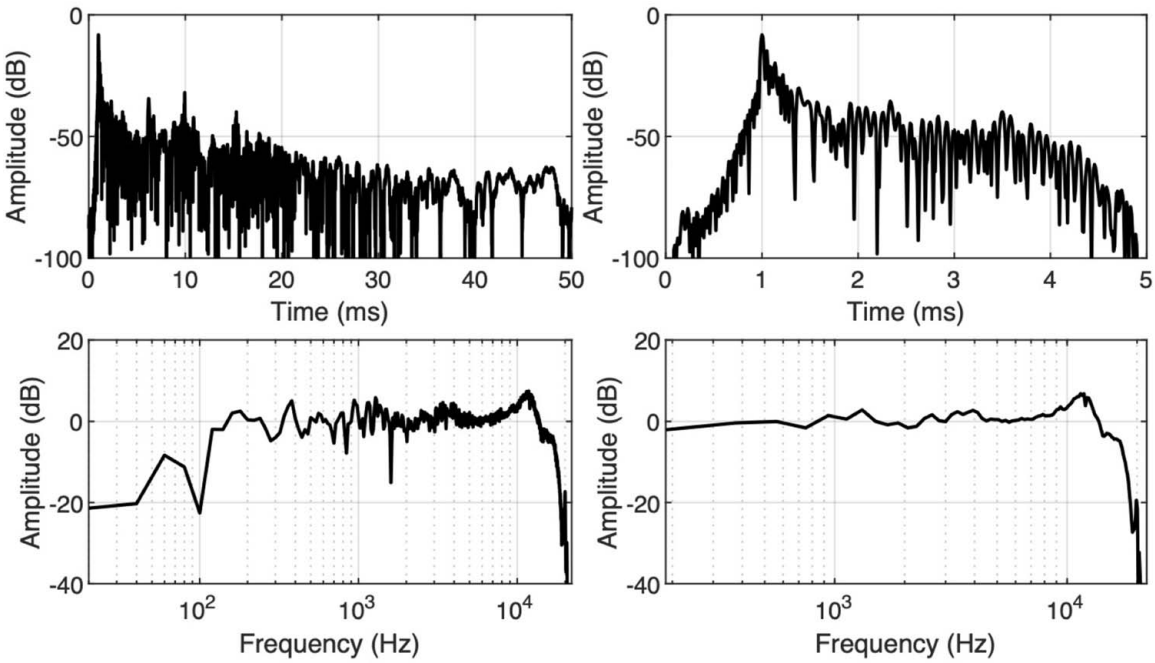


Fig. 2. Free-field impulse response of the loudspeaker at 0° elevation. Top: energy time curve. Bottom: magnitude response. Left: raw impulse response, truncated at 50 ms. Right: impulse response windowed at a length of 5 ms.

Fig. 3. Microphone inserted in the subject's ear.

titrack, Flex 3 cameras, Motive 1.5 API) was used. To this end, a thin headband with reflective markers was placed on the subject's head. Furthermore, three self-leveling alignment lasers (Bosch GCL 2-15) were employed to define a fixed coordinate system within the room and ensure that the subject's initial position was aligned with the center of the loudspeaker arch.

The acoustic measurements were controlled from and recorded on a Windows machine using another custom application[2] based on AMTatARI v7.0.23. This application allows for high-level control of the measurement workflow. It creates instances of other applications for low-level specific tasks: Pure Data for the real-time audio input and output, MATLAB for the signal processing, and a custom executable for the head tracking (based on Motive 1.5 C++ API) and moving the turntable.

For the 3D scanning, an EinScan Pro 2X 2020 3D scanner and EXScan Pro v3.7.0.3 software were employed. It is a structured-light 3D scanner that projects an infrared pattern and derives depth information from the deformation of this known pattern as captured by stereo cameras. Subjects were asked to wear a wig cap to prevent hair from being detected as a rigid surface during the scanning; for consistency, the wig cap was also worn during the HRTF measurements.

In addition to the 3D model, still images with depth information of the subjects' heads were captured with an iPhone XS. The turntable was used to obtain photos at every 5° angle from a fixed vantage point (72 photos per subject). A custom iOS app (the link to the publicly released code can be found in the dataset website) controlling the turntable and camera was developed to this end.

2 MEASUREMENT PROCEDURE

The whole measurement procedure is broken into five stages: measurement of the HpTF (which also served as a microphone sanity check), alignment of the subject's head, HRTF measurement, depth photographs, and 3D scanning. The in-ear microphones are positioned before the first stage

[2]https://github.com/Audio-Experience-Design/expsuite-code.

and removed before the fourth stage. The subject wears a wig cap at all times, as explained in the previous section.

2.1 Headphone Measurement

First, a pair of Sennheiser HD 650 headphones are measured on the subject using the same microphones as in the HRTF measurement. Then, to measure the HpTF, two exponential sweeps [23] from 20 to 22,000 Hz and with a length of 500 ms are used. This measurement is repeated five times, asking the subject to remove and put the headphones back on each time to account for fit-to-fit variations [24].

An automatic validation is employed, which warns the experimenter if any headphone measurements display a signal-to-noise-ratio (SNR) lower than 60 dB or if the difference between the left and right channels' energy is above 3 dB. This could be caused by, for instance, one of the microphones moving out of position or the subject producing self-noises (coughing, loud breathing, etc.). When this happens, the experimenter is shown a plot of the measurements and can decide whether to repeat that particular headphone measurement.

2.2 Alignment

The subject is asked to sit on the chair, and their position is adjusted with the help of the alignment lasers to ensure that their head is aligned and at the center of the arch. Right after the subject's ears are aligned with the lasers, the tracking system is initialized, and the lasers are turned off. Then, an acoustically absorbent curtain is used to hide the computer and experimenter to reduce specular reflections prior to the HRTF measurement.

2.3 HRTF Measurement

The HRTF measurement starts at the initial position with the subject facing the loudspeaker arch (0° azimuth). To allow for fast measurements, the multiple exponential sweep method [25] is used. Sweeps go from 20 Hz to 22 kHz, have a length of 500 ms, and are played sequentially from all 23 loudspeakers with an interval of 180 ms between each sweep. This interval was chosen to minimize the measurement time without compromising the SNR [25].

After all 23 sweeps have been played, the software automatically checks whether the subject's position or orientation has gone outside one of the tolerance ranges (±2.5° for azimuth/elevation; ±5° for roll; ±10 cm for X/Y/Z displacements) at any time during the measurement, in which case, the subject is given auditory feedback to help them correct their posture, and the measurement is repeated. It has to be noted that, thanks to the presence of a chinrest, significant translational shifts were never experienced during the alignment/measurement procedures. In order to perform an accurate re-alignment, the tolerance range for azimuth/elevation/roll is reduced to ±1° during the correction process.

Once the measurement for a given position is completed, the turntable is rotated by 5° counter-clockwise, and the process is repeated. This continues until the subject has rotated 175° from their initial position.

Fig. 4. Photogrammetry setup before starting to spin the turntable.

Automatic validation is employed, which warns the experimenter if any of the HRTF measurements displays an SNR lower than 60 dB or if the estimated sound energy of windowed head-related impulse response (HRIR) (discussed in SEC. 3.2) in either ear is below 90% of the impulse response's total energy. When this happens, the experimenter is shown a plot of the measurements, and, using their professional judgement, they decide whether to repeat the measurement for that particular position. HRIRs with only minor deviations from thresholds are usually kept, but HRIRs with major anomalies are repeated.

2.4 Depth Photographs

After the acoustic measurements have finished, the subject remains seated, and the microphones are taken off. An iPhone XS is positioned at head height using a tripod. The phone is orthogonally aligned to the subject's right side so that the head fits in the viewfinder. The front-facing camera of the iPhone was redirected sideways with a custom 3D-printed mirror bracket to use Apple's TrueDepth technology to capture depth information, which is not available on the rear cameras. On activating the shutter button in the custom app, the turntable starts spinning 360° clockwise, and pictures get taken automatically every 5°. A view of the setup before starting the measurement is shown in Fig. 4.

2.5 3D Scanning

Next, the subject is moved to a free-standing chair. There, a 3D scan is made by manually sweeping the hand-held scanner around the head and upper torso. Special attention is paid to covering the pinnae from all angles. Nonetheless, small unscanned patches remain because direct lines of

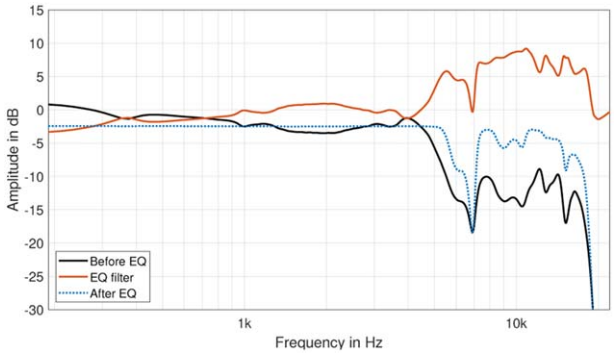


Fig. 5. Magnitude response of a headphone transfer function before and after applying the HpEQ filter, as well as the response of the actual HpEQ filter. Note how the regularization prevented the inversion of the notch at around 7 kHz. All measurements correspond to subject P0001, left ear.

sight are necessary with structured light scanners. These missing patches typically appear around the back of the ear and other narrow concave structures. Additionally, less time is spent scanning the face because the stroboscopic effect caused by the camera's high refresh rate is unpleasant to look at, even with closed eyes. Furthermore, a high level of detail in the 3D scan of the face was deemed unnecessary for the authors' purposes. Finally, any facial or long hair protruding from the wig cap also causes unscanned patches in the model because of their low reflectivity.

3 POST-PROCESSING

3.1 Headphone Measurement

Each of the five measured headphone impulse responses is cropped to a length of 4,096 samples, leaving the onset approximately at $t = 1$ ms. Then, fade-in and fade-out are applied via Hann windows of 16 and 128 samples, respectively.

In order to calculate the headphone equalization (HpEQ) filters, the average of the five measurements is calculated. To that end, the windowed impulse responses are transformed to the frequency domain via discrete Fourier transform, and the average of their magnitudes is calculated. Then, a minimum-phase HpEQ filter is constructed by inverting that average magnitude response and using the Hilbert transform to generate the phase response, according to [26].

Regularization was applied to prevent excessive amplification when inverting low magnitude values, which could cause the HpEQ to introduce ringing artefacts, following [27]. Fig. 5 shows an example of a headphone magnitude response and its corresponding HpEQ filter. It can be seen how the deep notch at around 7 kHz was not compensated by the HpEQ filter due to the regularization.

All five headphone measurements, the average response, and the HpEQ filter are saved in the database at three sampling frequencies: 44.1, 48, and 96 kHz. Those could directly be used by each measured subject to compensate for the filtering generated by this specific model of headphones on their specific head, which might be beneficial in order to improve the quality of the binaural rendering (see [27]). Including the raw measurements allows for the database users to design custom HpEQ filters that are less conservative than the regularized inversion employed here, e.g., by manually tuning the inversion of deep notches.

3.2 HRTF Measurement

First, the HRIRs for each azimuth and elevation are extracted from the measurements and truncated at a length of 50 ms. This is saved in the database as the "raw" version of the HRTF in SOFA format.

Then, the HRIRs are truncated at a length of 5 ms, leaving at least 1 ms of room for the onset. Fade-in and fade-out of 16 and 128 samples, respectively, are applied via Hann windows. This is also saved in the database as the "windowed" version of the HRTF.

Next, the windowed HRIRs are equalized by the free-field measurement (see Fig. 2) of the corresponding loudspeakers. In order to do this, a regularized inverse filter is calculated from the free-field measurements, using the same technique as in the HpEQ. This HRTF is saved in the database as "FreeFieldComp."

An alternative version of the free-field–compensated HRTF is also calculated, in which a minimum-phase equivalent of the inverse filter is used for the equalization. This would be useful to prevent phase artefacts that may arise from the inversion process, such as pre-ringing. Note how in Fig. 6, the energy time curve of this HRTF (fourth column) shows less energy before the onset than the non–minimum-phase version (third column). This HRTF is saved in the database as "FreeFieldCompMinPhase."

Then, for all four HRTF types mentioned above, an "ear-aligned" version is also calculated. In this version, the interaural time differences (ITDs) are estimated for all HRIRs, using the "threshold" estimation method from [28] with a threshold of 10 dB and a low-pass frequency of 3 kHz. Then, the ITDs are removed from all HRIRs and saved as metadata in the SOFA files. This format is useful for HRTF interpolation [29] and for compatibility with the 3D Tune-In Toolkit [2].[3] All SOFA files are exported in three different sampling frequencies: 44.1, 48, and 96 kHz.

3.3 Depth Photographs

The RGB+depth photographs are stored separately as high-resolution color files in HEIC format (7 gigapixels at 8 bit) and as lower-resolution depth files in TIFF format (VGA – video graphics array resolution at 32 bit). One potential use of these pictures is combining them into a 3D model through photogrammetry. This would then provide a 3D model that is representative of the level of detail that is achievable with current consumer hardware, which would make an interesting comparison with the professional-grade 3D capture that was employed (see next section).

[3] https://github.com/3DTune-In/3dti_AudioToolkit.

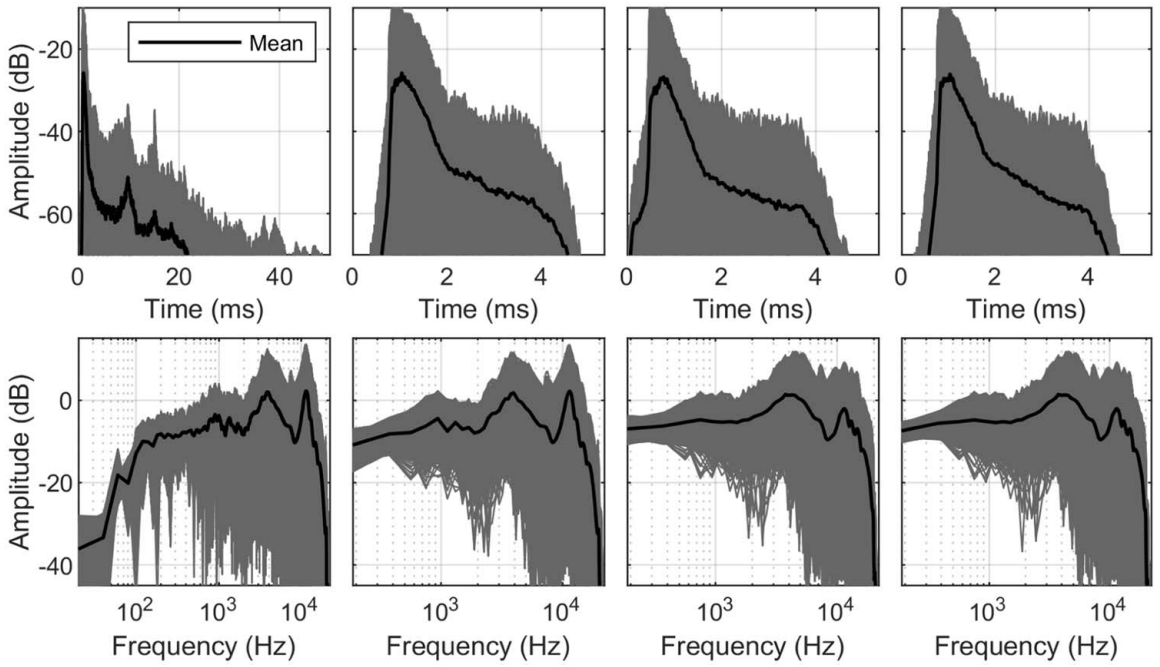


Fig. 6. Example of the different versions of an HRTF measurement (subject P0001, left ear). From left to right: raw, windowed, free-field compensated, and free-field compensated with minimum phase filter. The top row shows the energy time curve, and the bottom row shows the magnitude response. All 793 directions are plotted in gray, whereas the mean value is shown in black.

3.4 3D Scanning

The raw point cloud coming from the scanner is saved in ASC file format. The resolution (distance between two points) of the generated point cloud is 0.5 mm. This is then transformed into an un-watertight mesh by the EXScan Pro software using minimal processing, with a low amount of filtering and smoothing, and only holes of a perimeter smaller than 10 cm are filled in. This ensures as much of the detail as possible is saved. Then another watertight mesh is created. Both are saved as STL files. Fig. 7 displays the three different versions of the 3D scan of the KEMAR manikin. From the left-hand side to the right-hand side, there is the point cloud scan version, followed by the un-watertight and watertight mesh versions. These 3D scans could be used, for example, to synthesize HRTFs using methods such as Boundary Element Methods (e.g., [30]) or to develop new methods based on machine learning approaches.

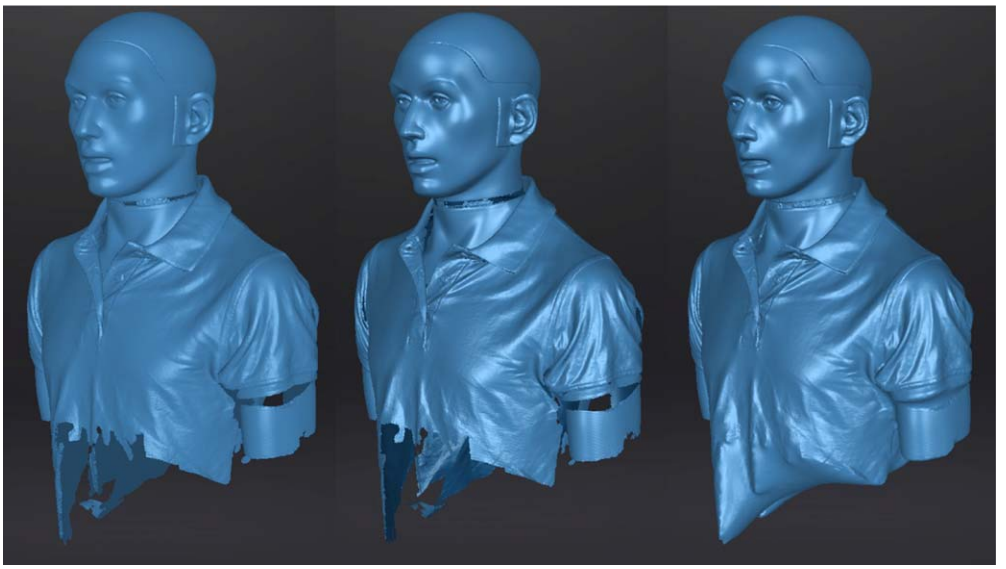


Fig. 7. Example of the different versions of the 3D scan of the KEMAR Manikin. From left to right: point cloud scan, un-watertight, and watertight mesh.

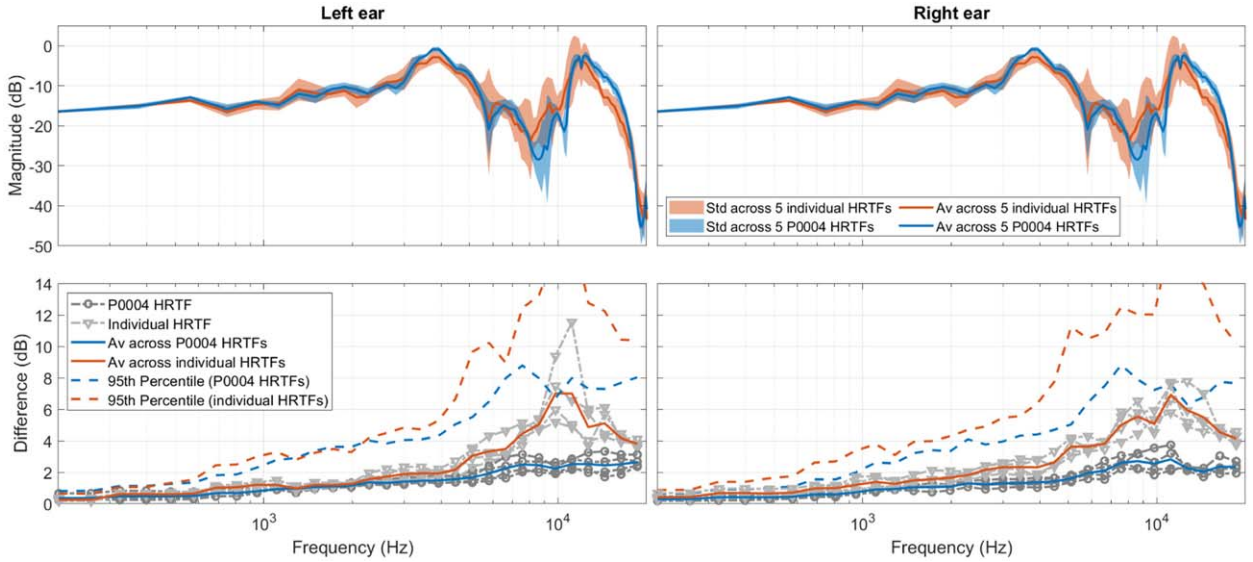


Fig. 8. Top panels: Average magnitude (solid lines) and standard deviation (shaded areas) of the front HRTF at the left (left panel) and right (right panel) ears for five different subjects or for five measurements of P0004. Bottom panels: Average magnitude difference between repeated and reference P0004 measurements across all source positions, together with the paired comparisons. The dashed lines shows the 95th percentiles. The same plots are shown also for the five different subjects (one subject is randomly chosen for reference).

4 MEASUREMENT VALIDATION

The repeatability of the HRTF measurement system was evaluated by comparing five measurement iterations of subject P0004. For each measurement, the whole setup procedure was rerun; the microphones were removed and then re-inserted in the ears, and the head alignment process was repeated. For comparison purposes, five measured HRTFs from five different subjects, randomly selected within the first participants to the measurement sessions (namely, P0001, P0002, P0003, P0006, and P0007) were also considered. The HRTFs were compared by computing numerical errors (SEC. 4.1) and running some predictions with perceptual models (SEC. 4.2).

4.1 Numerical Evaluation

Our first numerical analysis focuses on the HRTFs measured in the frontal position, because it allows for verification of the consistency of the head alignment and microphone placement across measurements. The "windowed" versions of the HRTFs are used for this comparison. Fig. 8 displays on the top panels the average front HRTF magnitude ("Av," solid lines) and standard deviation ("Std," shaded areas) across the five different subjects and five measurements from subject P0004. The standard deviation for the five different subjects is 4.5 and 4.1 dB (averaged across frequencies) for the left and right ears, respectively, whereas it is about 3.0 and 2.8 dB for the five measurements from P0004. Below 3,000 Hz, the standard deviation is 0.5 dB larger (averaged across ear and frequencies) for the five different subjects if compared with the repeated measures; this value grows to 1.8 dB when calculated between 3 and 20 kHz. This suggests that, in the authors' measurements, the influence of the shape of the ears, head, and torso is greater than the influence of the head alignment and microphone location in the ear canal.

The HRTFs spectra are also compared across all source positions in Fig. 8, bottom panels. This allows to compare the overall repeatability of the measurements, as opposed to the head and microphones alignment, which were the focus of the previous analysis. Each P0004 measurement is compared to a reference P0004 HRTF (i.e., the one that will be released), whereas for the other set of HRTFs, one is randomly chosen as the reference within the group. For each pair of HRTFs, the average difference is computed as the average across source positions of the difference between the HRTF spectra (this is shown as gray lines in the bottom panels in Fig. 8). Then, the average between the four comparisons is considered and plotted as solid lines. The dashed lines display the 95th percentile distribution across all differences.

Generally, the lines corresponding with P0004 are below the ones corresponding with the other HRTFs, meaning that the differences between the repeated measure of P0004 HRTFs are lower than those between the HRTFs measured from the five individuals. For example, the average difference is up to 2.5 dB for frequencies above 3 kHz for the P0004 HRTFs, and 6 dB when looking at the five individuals' HRTFs.

In addition to confirming the validity of these measurements by showing that the differences between repeated measures of the same subject are smaller than the differences between measurements done on different subjects, these values are comparable with results from previous research, e.g., [21]. It has to be noted though that, in the authors' case, the validation was done with five HRTFs and averaging all source positions, whereas in [21] it was done on the frontal location, and for 2 HRTFs only.

4.2 Model-Based Perceptual Evaluation

HRTFs are used to create a spatial sound experience for a human listener by imitating the perceptual spatial hearing cues. Therefore, a perceptual consistency evaluation of the measured HRTF dataset may complement the numerical analysis of spectrum differences, presented in the previous section, in validating the coherence of the database.

Ideally, this would be done via subjective listening tests. However, computational auditory models provide a more rapid alternative and allow for simulation of exhaustive experiments that often would be too demanding for an actual human listener to perform reliably. For this evaluation, a binaural speech intelligibility model [31] and an auditory sound localization model [32], available through the Auditory Modeling Toolbox [33], were employed to assess the similarity of various HRTF measurements from the database.

4.2.1 Speech Intelligibility Predictions

The Jelfs binaural speech intelligibility model predicts the spatial benefit in decibels for a given speech-masker location compared to the reference where the target and masker are co-located. It takes as input the HRIRs of the target and masker locations to compute the spectra, ITDs and masker coherence (assessing how similar the masker is at the individual's ears) per frequency band using a gammatone filter-bank. Based on the spectra, the model computes the SNR at each ear and keeps the higher of the two within each band. A formula is applied to assess the spatial benefit related to ITDs. This formula considers the masker and target ITD and the coherence of the masker at the ears. The values are then integrated across frequencies giving more weight to frequency bands relevant for speech. Finally, the broadband SNR is added to the broadband spatial benefit related to the ITD, providing an output in decibels that can be interpreted as a prediction of the spatial benefit in terms of speech intelligibility. For instance, if the output is 5 dB for given speech-masker locations, this means that the target level could be decreased by 5 dB while preserving the same intelligibility of a co-located target/masker condition.

These intelligibility predictions are computed for assessing the overall spatial benefit of HRTFs. These are performed only in the horizontal plane, where the model has been previously validated. To compute the overall benefit of one HRTF, predictions are made for each possible target-masker location and then averaged. The benefits obtained for the five HRTFs from P0004 range from 7.0 to 7.3 dB, whereas the benefits range from 7.1 to 7.5 dB for the five individual HRTFs. These values represent a validation of these measurements, because they are in line with previous evaluations of the model [31]. Nevertheless the similarity between the repeated measurements and the different individuals suggest that the numerical differences observed in SEC. 4.1 do not influence speech intelligibility predictions.

4.2.2 Sound Localization Predictions

The spherical sound localization model estimates how well a human listener would perform in a sound localization task when presented with binaural stimuli processed with a target HRTF that differs from their own true HRTF (template). The inputs of the model are two HRTFs (target and template) in SOFA format and the experimental conditions, such as the evaluated sound source directions and the number of repetitions. The model outputs are the predicted sound direction estimations made by the listener.

Generally, if the target and template HRTFs are similar, the estimations will be close to the actual sound source directions. On the other hand, if they are different, the estimations will contain larger localization errors, e.g., because of front-back and up-down confusions. For an accurate estimate of absolute localization errors, a set of free model parameters, which control non–HRTF-related aspects of sound localization performance (e.g., individual pointing accuracy), have to be calibrated using individual sound localization test data. However, personal sound localization data was unavailable, and only relative differences in localization errors were of interest to this study. Therefore, median parameter values, calculated using data from previous sound localization studies [34, 35], were used in this study.

For the evaluation, the five measurements of subject P0004 were used as template HRTFs, representing five hypothetical listeners. For each of them, sound localization tasks were simulated for the same ten target HRTFs that were analyzed in the previous section: the five measured ones from subject P0004 plus the ones from the five randomly selected subjects. This is analogous to asking each listener to evaluate their own HRTF, four individually measured HRTFs, and five non-individual HRTFs. In each task, the listener would have to localize one of 1,706 directions, interpolated from the supplied HRTF, for each of the ten target HRTFs, with 300 repetitions (to account for stochasticity), resulting in 5,118,000 localization estimations per listener.

Fig. 9 shows the results of the evaluation. The top row displays the predicted sound localization errors when using the five P0004 HRTFs as targets, whereas the bottom row shows the errors when using the other subjects' HRTFs as targets. The errors are divided into RMS local polar error (first column), RMS lateral error (middle column) and quadrant errors (third column), as defined in [36]. The results are plotted in a matrix form to show the predicted errors of each target-template pair. Therefore, in the top row, diagonal positions represent modeled error using the same HRTF as both the target and template.

Overall, estimated localization errors when using different versions of the same subject's HRTF are smaller than when using HRTFs from different subjects (i.e., non-individual). This can be inferred by comparing the overall shade difference between the top-row plots and bottom-row plots, particularly regarding local polar error (the bottom row is brighter). Furthermore, the variation of the errors within the same subject (top row) is smaller than across subjects (bottom row), as can be observed from the more considerable color contrasts in the latter. Additionally, the overall localization errors across HRTFs of the same subject are comparable to the own-HRTF errors (top row, an-

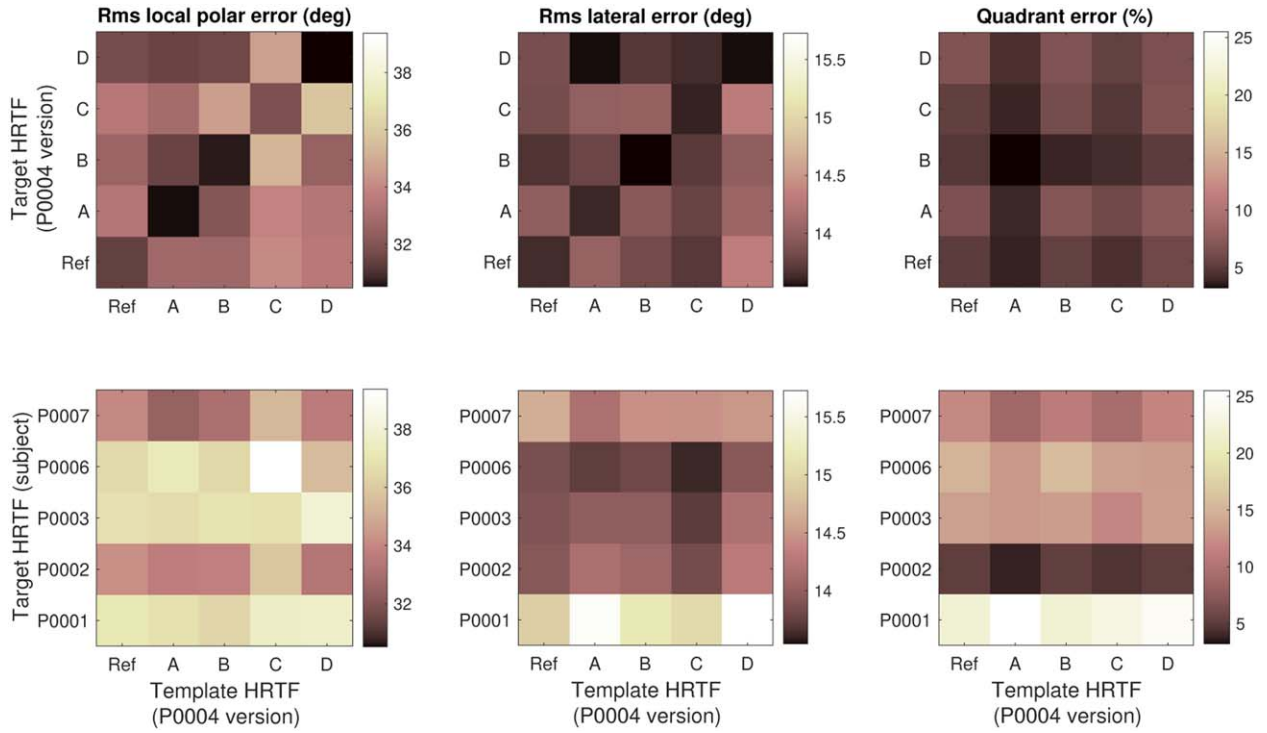


Fig. 9. Modeled sound localization errors for five versions of subject P0004 HRTFs when using different measurements of the individual HRTF (top row) and non-individual HRTFs from other subjects (bottom row).

tidiagonal), suggesting that subject P0004 would perform comparably well in a localization task regardless of which version of their measured HRTF is provided. Finally, the row-wise color similarity in the bottom-row plots indicates that the choice of template HRTF did not significantly affect the localization error. This suggests that the differences across subject P0004's HRTFs are relatively irrelevant compared to another subject's HRTF. Color differences between rows are due to overall differences between HRTFs measured from different subjects and the HRTFs measured from P0004. For example, subject P0002's HRTF is likely to be more similar to P0004's HRTFs if compared with P0001's HRTFs.

These results confirm the conclusions of the numerical analysis, namely that the differences in the measurements due to the placement of the microphones and the subject alignment are minor compared to the differences in the morphology of the ears and head of different human subjects.

5 FEATURES SUMMARY AND RELEASE

At the moment of publication (measurements are still ongoing and, therefore, these numbers will increase), the SONICOM HRTF dataset includes HRTFs measured from 120 subjects; 3D scans of their ears, head, and torso; and a set of depth pictures of the head taken every 5° on the horizontal plane. A total of 828 source positions have been measured around each subject's head at 1.5-m distance with the azimuth sampled every 5° and the elevation ranging from –45° to 90° (sampled every 10° between –30° and 30°, and every 15° below and above that). There are 72 azimuths for each elevation, and only one measurement of the top elevation at 90° has been included in the dataset (the one measured at 0° of azimuth). 793 different locations have therefore been released for every individual. In the SOFA files of the dataset, the azimuth is stored as an integer value between 0 and 355, encoding the azimuth angle, and the elevation as an integer between –45° and 90°.

In order to avoid potential issues with anti-aliasing filters and allow a wide dynamic range, the HRIRs are sampled at 96 kHz and 24 bits, but lower sample rates (44.1 and 48 kHz) are also included. For each sample rate, the following versions of the HRTFs (see SEC. 3.2) can be downloaded in SOFA format [37]:

- Raw,
- Windowed,
- Free-field compensated, and
- Free-field compensated with a minimum phase filter.

All the above are also provided with the ITDs removed from the signals and stored as metadata in the SOFA files. Some of the versions above are also released in the .3dti format in order for them to be directly usable with the 3D Tune-In Toolkit [2].

HpTFs (Sennheiser HD650) for every subject is also included in .mat format at each of the sample rates listed above. Additionally, the 3D models of ears, head, and torso are available in .stl format, with and without watertight post-processing, and the image data is available as .tif (depth).

The SONICOM HRTF dataset, together with other relevant information and data about the measurement system, can be accessed at the following link:

https://www.axdesign.co.uk/tools-and-devices/sonicom-hrtf-dataset. Please note that, due to privacy matters, RGB images are not publicly released. If these are required, please contact the team via email.

6 CONCLUSION

An HRTF dataset, which currently contains the accurately measured HRTFs of 120 subjects, has been created and publicly released. The dataset also contains, associated with each of these HRTFs, an HpTF, a 3D scan of the subject's head, and RGB+depth pictures of the subject at multiple angles, which can be used for photogrammetry. This dataset aims to aid researchers in creating reproducible research in the field of immersive audio.

In addition to continuing to measure subjects and releasing their HRTFs, several further developments could also be considered in the future. Because loudspeaker mounting points are available every 5°, down to 60° below the horizon, alternative speaker placements are possible. One option would be to shift half of the loudspeakers by 5° on one side of the arch to enable a 5°-resolution elevation measurement, albeit at the cost of double the measurement time.

7 ACKNOWLEDGMENT

The authors would like to thank Matt Speechley (for the custom design of the turntable), Neel Le Penru (for the chin rest), Kevin Sum and Oscar Jones (for the help in various stages of the setup), Michele Geronazzo, Piotr Majdak, and the ARI team (for the precious advice). This study was supported by the SONICOM project (www.sonicom.eu), funded by the European Union's Horizon 2020 research and innovation program under grant agreement 101017743.

8 REFERENCES

[1] L. Picinali and B. F. Katz, "System-to-User and User-to-System Adaptations in Binaural Audio," in M. Geronazzo and S. Serafin (Eds.), *Sonic Interactions in Virtual Environments*, Human–Computer Interaction Series, pp. 115–143 (Springer, Cham, Switzerland, 2023).

[2] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, et al., "3D Tune-In Toolkit: An Open-Source Library for Real-Time Binaural Spatialisation," *PLOS ONE*, vol. 14, no. 3, paper e0211899 (2019 Mar.). http://doi.org/10.1371/journal.pone.0211899.

[3] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF Database," in *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 99–102 (New Paltz, NY) (2001 Oct.). http://doi.org/10.1109/ASPAA.2001.969552.

[4] O. Warusfel, "LISTEN HRTF Database," http://recherche.ircam.fr/equipes/salles/listen/index.html (2003).

[5] N. Gupta, A. Barreto, M. Joshi, and J. C. Agudelo, "HRTF Database at FIU DSP Lab," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 169–172 (Dallas, TX) (2010 Mar.). http://doi.org/10.1109/ICASSP.2010.5496084.

[6] P. Majdak, B. Masiero, and J. Fels, "Sound Localization in Individualized and Non-Individualized Crosstalk Cancellation Systems," *J. Acoust. Soc. Am.*, vol. 133, no. 4, pp. 2055–2068 (2013 Apr.). http://doi.org/10.1121/1.4792355.

[7] 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 (2013 Nov.). http://doi.org/10.1250/ast.35.159.

[8] H. S. Braren and J. Fels, "A High-Resolution Individual 3D Adult Head and Torso Model for HRTF Simulation and Validation: 3D Data," Tech. Rep. (2020 Jun.). http://doi.org/10.18154/RWTH-2020-06760.

[9] F. Brinkmann, D. Manoj, P. Robert, et al., "The HUTUBS Head-Related Transfer Function (HRTF) Database," FG Audiokommunikation (2019 May). http://dx.doi.org/10.14279/depositonce-8487.

[10] R. Sridhar, J. G. Tylka, and E. Choueiri, "A Database of Head-Related Transfer Functions and Morphological Measurements," presented at the *143th Convention of the Audio Engineering Society* (2017 Oct.), e-Brief 357.

[11] 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, paper 2029 (2018 Oct.). http://doi.org/10.3390/app8112029.

[12] W. G. Gardner and K. D. Martin, "HRTF Measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908 (1995 Jun.). http://doi.org/10.1121/1.412407.

[13] M. Zhang, W. Zhang, R. A. Kennedy, and T. D. Abhayapala, "HRTF Measurement on KEMAR Manikin," in *Proceedings of ACOUSTICS*, vol. 9, paper 8 (Adelaide, Australia) (2009 Nov.).

[14] H. Wierstorf, M. Geier, and S. Spors, "A Free Database of Head Related Impulse Response Measurements in the Horizontal Plane With Multiple Distances," presented at the *130th Convention of the Audio Engineering Society* (2011 May), e-Brief 6.

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

[16] S. Spagnol, R. Miccini, and R. Unnthorsson, "The Viking HRTF Dataset v2," *Zenodo* (2020 Oct.). https://zenodo.org/record/4160401#.Y7SUi3ZKi5c.

[17] P. Siripornpitak, I. Engel, I. Squires, S. J. Cooper, and L. Picinali, "Spatial Up-Sampling of HRTF Sets Using Generative Adversarial Networks: A Pilot Study," *Front. Signal Process.*, vol. 2, paper 904398 (2022 Aug.). http://doi.org/10.3389/frsip.2022.904398.

[18] J. Pauwels and L. Picinali, "On The Relevance Of The Differences Between HRTF Measurement Setups For Machine Learning," in *Proceedings of the IEEE*

International Conference on Acoustics, Speech and Signal Processing (Rhodes Island, Greece) (2023 Jun.).

[19] L. Picinali, B. F. Katz, M. Geronazzo, et al., "The SONICOM Project: Artificial Intelligence-Driven Immersive Audio, From Personalization to Modeling [Applications Corner]," *IEEE Signal Process. Mag.*, vol. 39, no. 6, pp. 85–88 (2022 Nov.). http://doi.org/10.1109/MSP.2022.3182929.

[20] J. M. Alston and J. A. Rick, "A Beginner's Guide to Conducting Reproducible Research," *Bull. Ecol. Soc. Am.*, vol. 102, no. 2, paper e01801 (2021 Jan.). http://doi.org/10.1002/bes2.1801.

[21] F. Brinkmann, M. Dinakaran, R. Pelzer, et al., "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 Sep.). http://doi.org/10.17743/jaes.2019.0024.

[22] A. W. Mills, "On the Minimum Audible Angle," *J. Acoust. Soc. Am.*, vol. 30, no. 4, pp. 237–246 (1958 Apr.). http://doi.org/10.1121/1.1909553.

[23] A. Farina, "Advancements in Impulse Response Measurements by Sine Sweeps," presented at the *122nd Convention of the Audio Engineering Society* (2007 May), paper 7121.

[24] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Transfer Characteristics of Headphones Measured on Human Ears," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 203–217 (1995 Apr.).

[25] 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 Jul.).

[26] III J. O. Smith, *Spectral Audio Signal Processing*, http://ccrma.stanford.edu/∼jos/sasp/ (2011).

[27] I. Engel, D. L. Alon, K. Scheumann, J. Crukley, and R. Mehra, "On the Differences in Preferred Headphone Response for Spatial and Stereo Content," *J. Audio Eng. Soc.*, vol. 70, no. 4, pp. 271–283 (2022 Apr.). http://doi.org/10.17743/jaes.2022.0005.

[28] A. Andreopoulou and B. F. G. Katz, "Identification of Perceptually Relevant Methods of Inter-Aural Time Difference Estimation," *J. Acoust. Soc. Am.*, vol. 142, no. 2, pp. 588–598 (2017 Aug.). http://doi.org/10.1121/1.4996457.

[29] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions," *J. Audio Eng. Soc.*, vol. 69, no. 1/2, pp. 104–117 (2021 Jan.). http://doi.org/10.17743/jaes.2020.0070.

[30] B. F. G. Katz, "Boundary Element Method Calculation of Individual Head-Related Transfer Function. I. Rigid Model Calculation," *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2440–2448 (2001 Oct.). http://doi.org/10.1121/1.1412440.

[31] S. Jelfs, J. F. Culling, and M. Lavandier, "Revision and Validation of a Binaural Model for Speech Intelligibility in Noise," *Hear. Res.*, vol. 275, no. 1–2, pp. 96–104 (2011 May). http://doi.org/10.1016/j.heares.2010.12.005.

[32] 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. (forthcoming).

[33] P. Majdak, C. Hollomey, and R. Baumgartner, "AMT 1.x: A Toolbox for Reproducible Research in Auditory Modeling," *Acta Acust.*, vol. 6, paper 19 (2022 May). http://doi.org/10.1051/aacus/2022011.

[34] P. Majdak, M. J. Goupell, and B. Laback, "3-D Localization of Virtual Sound Sources: Effects of Visual Environment, Pointing Method, and Training," *Atten. Percept. Psychophys.*, vol. 72, no. 2, pp. 454–469 (2010 Feb.). http://doi.org/10.3758/APP.72.2.454.

[35] P. Majdak, T. Walder, and B. Laback, "Effect of Long-Term Training on Sound Localization Performance With Spectrally Warped and Band-Limited Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 2148–2159 (2013 Jul.). http://doi.org/10.1121/1.4816543.

[36] J. C. Middlebrooks, "Virtual Localization Improved by Scaling Nonindividualized External-Ear Transfer Functions in Frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1493–1510 (1999 May). http://doi.org/10.1121/1.427147.

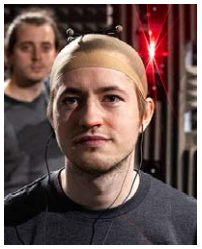
[37] P. Majdak, Y. Iwaya, T. Carpentier, et al., "Spatially Oriented Format for Acoustics: A Data Exchange Format Representing Head-Related Transfer Functions," presented at the *134th Convention of the Audio Engineering Society* (2013 May), paper 8880.

THE AUTHORS


Isaac Engel


Rapolas Daugintis


Thibault Vicente


Aidan O. T. Hogg


Johan Pauwels


Arnaud J. Tournier


Lorenzo Picinali

Isaac Engel received a B.Sc. degree in Electronic Systems Engineering and Master's degree on Telematics and Telecommunication Networks from the University of Málaga in 2015 and 2016, respectively. He was a doctoral (and briefly, postdoctoral) researcher at Imperial College London between 2016 and 2021, investigating spatial audio perception and signal processing methods for binaural audio rendering, and he obtained his Ph.D. in 2021. In 2018 and 2019, he worked as a Research Intern at Facebook Reality Labs. As of 2022, he is a Senior Research Engineer at Huawei's German Research Center.

•

Rapolas Daugintis is a Ph.D. student in the Audio Experience Design group at Imperial College London, working on binaural spatial audio personalization in extended reality applications. He was awarded an M.Sc. (Technology) in Acoustics and Audio Technology from Aalto University (Espoo, Finland) in 2021 and B.Sc. (Hons.) in Physics and Music from the University of Edinburgh in 2017. He has also worked as an acoustics consultant in AECOM (Leeds, United Kingdom) and Akukon (Helsinki, Finland).

•

Thibault Vicente is a Research Associate in the Audio Experience Design group at Imperial College London. Thibault is investigating the effect of head-related transfer function on auditory spatial mechanisms. Before joining Imperial College London, he graduated with a Ph.D. degree from Ecole Nationale des Travaux Publics de l'Etat (Lyon, France) and Macquarie University (Sydney, Australia) for his work on modeling the effect of hearing impairment on binaural speech intelligibility in noise. He also received an engineering degree from Ecole Nationale des Travaux Publics de l'Etat (Diplome d'Ingenieur Grandes Ecoles) and an M.Sc. in Acoustics from the University of Lyon.

•

Aidan O. T. Hogg received an M.Eng. degree in electronic and information engineering and Ph.D. degree from Imperial College London in 2017 and 2022, respectively. He is currently a Research Associate in spatial audio and virtual reality with the Audio Experience Design group at Imperial College London. He has also worked in various engineering roles with Broadcom, Dialog Semiconductor, and Nuance Communications. His current research focuses on using deep learning to capture head-related transfer functions and, more generally, spatial acoustics and immersive audio. Other research interests include speaker diarization and statistical signal processing for audio applications. More information about current research projects can be found here: https://aidanhogg.uk/.

•

Johan Pauwels is a Lecturer at the Centre for Digital Music of Queen Mary University of London. He received Master of Science degrees in Electrical Engineering (KU Leuven, 2006) and Artificial Intelligence (KU Leuven, 2007). In 2016, he obtained a Ph.D. from Ghent University on the topic of automatic harmony recognition from audio. He has contributed to multiple national and international research projects in the United Kingdom, Belgium, and France. His main research interests are signal processing, artificial intelligence, and big data science applied to music and immersive audio.

•

Arnaud J. Tournier is a postdoctoral researcher at Imperial College London, where he also obtained his Ph.D. in Machine Learning in 2021. Arnaud currently works with the Neural Reckoning group and the Audio Experience Design group on meta-learning, deep learning, synthetic head-related transfer functions, and spatial audio more generally. His work focuses on enabling immersive audio

technologies, particularly for augmented and virtual reality consumer applications. Before his Ph.D., Arnaud studied Stochastics and Machine Learning at Pierre and Marie Curie University (Master, Paris VI), Pure Mathematics at Paris-Sud (Master, Paris XI), and Engineering at Ecole Centrale Paris (Diplome d'Ingenieur Grandes Ecoles).

Lorenzo Picinali is a Reader in Audio Experience Design at Imperial College London. In the past few years, he worked in Italy, France, and the United Kingdom on projects dealing with 3D binaural sound rendering, interactive applications for visually and hearing impaired individuals, audiology and hearing aids technology, audio and haptic interaction, and, more in general, acoustical virtual and augmented reality. More information about the projects in which Lorenzo is involved can be found here: https://www.axdesign.co.uk/.