SimpleHRTF-3D: From Head Mesh to Immersive Spatial Audio

Augusto M. P. de Mendonça
Abraão de Santana
Jan M. Teixeira
Kedson A. Silva
Mariza Ferro
Igor M. Coelho
augustompm@id.uff.br
ac_santana@id.uff.br
janmt@id.uff.br
alves_kedson@id.uff.br
mariza@ic.uff.br
imcoelho@ic.uff.br
Universidade Federal Fluminense
Niterói, RJ, Brazil

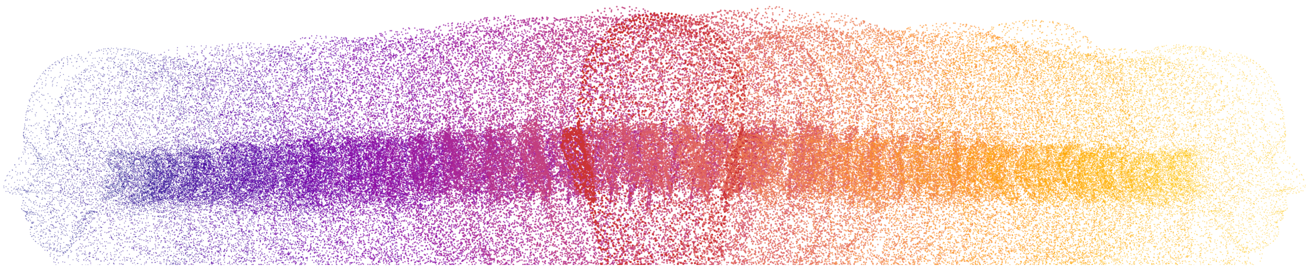

Figure 1: Visualization of 3D point clouds from all HUTUBS subjects, illustrating morphological variability.

ABSTRACT

Head-Related Transfer Functions (HRTFs) are essential for immersive spatial audio in multimedia applications like virtual reality and gaming, yet personalization remains challenging due to precise anthropometric measurement requirements. This paper introduces an open source pipeline for HRTF customization through three objectives. First, we validate a published Random Forest model on HUTUBS, achieving similar R^2=89.8% and SD=4.45 dB with manual measures. Second, SimpleHRTF-3D automates extraction from 3D head meshes using two-step PSO, achieving 10.96% mean extraction error (a 2.31 percentage point absolute reduction from 13.27% single-step error), yielding R^2=89.3% and SD=5.04 dB. As proof of concept, we extended the method to use photogrammetry, enabling photo-to-HRTF personalization. Validated on 58 HUTUBS subjects, our pipeline integrates manual, mesh, and image methods, providing reproducible tools for multimedia HRTF adaptation. The results demonstrate high-fidelity spatial audio capabilities for diverse immersive applications.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

KEYWORDS

HRTF personalization, spatial audio, machine learning, particle swarm optimization, metaheuristics, 3D mesh extraction, anthropometric measurements

1 INTRODUCTION

Spatial audio represents a fundamental component for immersive experiences in virtual reality (VR), augmented reality, and metaverse applications [16]. Head-Related Transfer Functions (HRTFs) enable the simulation of three-dimensional sound perception by encoding how sound waves interact with the listener's head, torso, and pinnae before reaching the eardrums [4]. These acoustic filters are essential for creating convincing spatial audio in multimedia systems, particularly in research on multisensory VR environments and 360-degree video tools, where immersive applications continue to expand [22, 28].

The effectiveness of spatial audio systems is critically dependent on HRTF personalization. A generic HRTF, while functional, can produce localization errors and front-back confusions that compromise the immersive experience [30]. An individualized HRTF, tailored to the listener's specific anthropometric features, significantly improves localization accuracy and reduces perceptual artifacts [7].

However, the acquisition of a personalized HRTF remains challenging due to three primary barriers: the cost of the acoustic measurement facilities, the complexity of the measurement procedure that requires hours in anechoic chambers, and the variability between subjects that demands a precise capture of individual anatomical features [4].

Practical barriers extend beyond the complexity of the measurement. Individual HRTF measurements are very time-consuming procedures that require specialized acoustic facilities in controlled environments [2]. The measurement setup involves rotating loudspeakers around subjects who wear in-ear microphones to capture responses from all spatial directions [16]. These technical requirements have confined HRTF personalization to research laboratories, preventing broader adoption in multimedia consumer applications and assistive technologies for hearing-impaired patients.

A traditional method for acquiring HRTF data is exemplified by the HUTUBS dataset from the Hearing Research Laboratory at TU Berlin, selected for analysis due to its large size and the availability of data in various formats [4]. In this study, 96 subjects underwent complete spherical measurements in 440 spatial positions. This approach, while providing reliable reference data, remains impractical for widespread deployment due to the high costs and time requirements of acoustic measurement facilities. Figure 1 visualizes this challenge by presenting an aggregated point cloud from all HUTUBS subjects, where the color gradient intensity reveals the substantial geometric variability in head morphology and ear structures across the population. This visualization demonstrates the complexity of developing a universal extraction pipeline that must accommodate such anatomical diversity while maintaining measurement precision.

Practical barriers have motivated computational alternatives. Recent advances in Machine Learning (ML) have demonstrated promising results for predicting HRTFs from anthropometric measurements. Multiple approaches exist in the literature, including deep neural networks that combine anthropometric data with ear images [13], PCA-based individualization using anthropometric dimensions [2], and Random Forest models for HRTF recommendation [21]. Teng and Zhong (2023) [30] achieved $R^2 = 90\%$ using Random Forest models with 23 manual measurements from the HUTUBS database with a spectral distortion (SD) of 4.74dB. However, as noted by Fantini et al. (2025) [7], these approaches share a common limitation: they require precise manual measurements that are difficult to obtain outside laboratory settings. This limitation motivates the development of automated extraction methods that can derive these measurements from more accessible data sources, such as 3D scans or photographs.

Motivated by these challenges, this study addresses the requirement for precise manual anthropometric measurements that are difficult to obtain outside of laboratory settings. The main objective is to develop SimpleHRTF-3D, an automated pipeline that enables HRTF personalization without manual measurements, progressing from laboratory-measured data to consumer-accessible input.

To achieve this objective, we establish three progressive components with corresponding contributions. First, we validate Teng and Zhong's Random Forest model [30] using manual anthropometric measurements from the HUTUBS database, establishing a performance baseline of R^2=89.8% with an open source implementation. Second, we develop automated feature extraction from 3D head meshes using Particle Swarm Optimization (PSO) [12], a metaheuristic algorithm that optimizes measurement parameters to minimize extraction errors. Our novel two-step PSO approach separates ear and head measurements, achieving a 17.4% relative error reduction (from 13.27% to 10.96% absolute error) compared to single-step optimization while maintaining prediction performance. Third, as a proof of concept, we extend the pipeline to photogrammetry, demonstrating that consumer photographs can achieve R^2=67.9% HRTF prediction accuracy using COLMAP reconstruction [26] [27], exceeding generic HRTF performance. Our complete pipeline and code are made available to support reproducible research in spatial audio personalization.

2 LITERATURE REVIEW

This section reviews the theoretical foundations and current approaches in HRTF personalization, establishing the context for our contributions while identifying gaps that motivate our work. We focus on three key areas that directly inform our methodology: anthropometry-based HRTF prediction, automated measurement extraction from 3D data, and photogrammetric reconstruction for consumer applications.

HRTFs encode the acoustic transformations that occur as sound travels from source to the listener's eardrums, capturing binaural cues essential for spatial perception [16]. While the theoretical foundation was established decades ago, practical personalization remains challenging due to measurement complexity and cost barriers. Current approaches generally fall into three categories with distinct trade-offs between accuracy and accessibility.

Selection-based methods match users and existing HRTF databases, with Pelzer et al. (2020) [21] achieving 75% NDCG scores using Random Forest models in anthropometric feature ratios. Synthesis approaches modify existing HRTFs through scaling techniques, as demonstrated by Middlebrooks (1999) and later refined by Mokhtari et al. (2008), who achieved near-individual localization accuracy, though elevation perception remained problematic [17, 18]. The most relevant category for our work involves anthropometry-based prediction, where Bomhardt et al. (2017) [2] established that PCA weights can be expressed through anthropometric dimensions using linear regression, while Fels and Vorländer (2009) [8] quantified relationships between anatomical features and HRTF characteristics across age groups.

Machine learning has significantly advanced HRTF prediction capabilities, with Teng and Zhong (2023) [30] achieving $R^2 = 90\%$ with a mean spectral distortion of 4.74 dB using Random Forest models with 23 manual measurements from HUTUBS. We adopt this work as our performance benchmark since it represents the state-of-the-art on the same HUTUBS database, enabling direct quantitative comparison. Alternative approaches include deep neural networks proposed by Lee and Kim (2018) [13], who combined anthropometric measurements with ear photographs, although computational complexity remains a concern. Beyond performance metrics, the choice of algorithm carries environmental implications, with Ferro et al. (2023) demonstrating that decision tree ensembles like Random Forest offer superior energy efficiency compared to deep learning approaches, whose training can produce CO_2 emissions equivalent to five cars over their lifetime [9]. Random Forest algorithms have thus proven particularly effective for HRTF tasks due to their ability to handle non-linear relationships while providing interpretable feature importance rankings [3], making them not only accurate but also a sustainable choice for spatial audio applications.

An accessibility gap persists between research capabilities and practical deployment. Current HRTF personalization methods require either specialized acoustic measurement facilities with controlled environments and professional equipment, or precise manual anthropometric measurements that require trained personnel and significant time investment. This limitation has confined personalization to research settings, preventing widespread adoption in consumer multimedia applications and assistive technologies. Recent surveys by Fantini et al. (2025) [7] identified automated feature extraction as a key research gap, noting that although ML models achieve high accuracy with manual measurements, automated extraction from accessible 3D data sources remains challenging.

This automation of anthropometric measurements presents unique challenges, particularly for acoustic applications that require submillimeter precision in ear geometry. The general approaches of Yan et al. (2020) [33] achieved errors ranging from 2.5 to 16.0 mm on body measurements, while Giachetti et al. (2014) [11] found that automated landmark detection consistently exhibits higher errors than manual methods. Ear-specific work has focused primarily on biometric applications rather than acoustic measurements. Lin et al. (2024) [14] achieved submillimeter precision in CT-reconstructed ears, although medical imaging requirements limit practical deployment. Prakash and Gupta (2012) [23] developed efficient ear localization, while Mursalin and Islam (2021) [19] extended detection to 3D domains but focused on identification rather than measurement. This disconnect between detection capabilities and acoustic measurement needs, highlighted by Ganapathi et al. (2023) [10], motivates our PSO-based optimization approach. Abualigah (2025) [1] highlighted PSO's robustness in high-dimensional search spaces and adaptation to nonlinear constraints, while requiring minimal implementation overhead.

Photogrammetric reconstruction offers potential pathways toward consumer accessibility, though application to anthropometric measurement remains challenging. COLMAP, developed by Schönberger and Frahm (2016) [26], enables robust reconstruction from unordered image collections, while Matuzevičius and Serackis (2022) [15] demonstrated 12× quality improvements for smartphone-based head reconstruction, although ear regions remain particularly difficult due to occlusions and fine geometric details. Statistical outlier removal methods from Ross (2021) [24] and Point Cloud Library implementations by Rusu and Cousins (2011) [25] provide essential noise reduction capabilities for processing such data.

Our work addresses these limitations through a progressive approach that bridges the accessibility gap. We first validate the Random Forest baseline, then introduce automated extraction using novel two-step PSO optimization, and finally demonstrate proof-of-concept photogrammetric reconstruction. This progression from manual measurements to consumer photography directly tackles the fundamental barrier that limits widespread adoption of HRTF personalization.

3 THE HUTUBS DATABASE AND ANTHROPOMETRIC FEATURES

The HUTUBS database is publicly available and provides a foundation for HRTF research and personalization [6]. This cross-evaluated database contains acoustically measured and numerically simulated HRTFs from 96 subjects across 440 spatial positions. Each measurement covers the audible frequency range with 256 samples at 44.1 kHz sampling rate [4].

Beyond acoustics, HUTUBS includes 3D head meshes in PLY format, enabling automated anthropometric extraction research. Of the 96 subjects, 58 possess complete 3D meshes suitable for our pipeline, with subject 80 having a mesh but missing SOFA measurements. The database provides 25 anthropometric features per subject (39 when considering both ears), measured semi-automatically using standardized protocols. These features include head dimensions, ear geometry, and angular measurements for HRTF prediction. These meshes serve as the foundation for our automated feature extraction pipeline [6].

3.1 Anthropometric Features

The anthropometric characteristics are divided into two primary categories: head/torso measurements and ear-specific parameters. Head measurements include fundamental dimensions (x1-x3) representing width, height, and depth, respectively, along with pinna offsets (x4-x5) indicating ear position relative to the center of the head. The ear measurements comprise detailed geometric parameters (d1-d10) that capture the structure of the pinna, as shown in Table 1.

Feature importance analysis by Fels and Vorländer (2009) demonstrated that head dimensions primarily influence low-frequency ITD and ILD cues below 3 kHz, while pinna geometry controls high-frequency spectral features above 4 kHz [8].

Not all features were proven to be equally extractable from 3D meshes. Direct measurements such as the general dimensions of the pinna (d5, d6) and the head parameters (x1-x3) can be reliably obtained through geometric analysis. However, internal ear structures (d1-d4, d8-d10) require either manual annotation or statistical derivation. Our analysis revealed consistent proportional relationships, enabling feature estimation when direct measurement proves infeasible. For example, the height of the cavum concha is strongly correlated with the height of the general pinna following the relationship $d1 = 0.296 \times d5$ (correlation coefficient $r = 0.842$), validated across the 58 subjects with complete data.

Table 1 summarizes the core anthropometric features used throughout our pipeline. Following the baseline methodology established by Teng and Zhong (2023) for transparency and reproducibility, we employ both anthropometric parameters and frequency features as inputs to the Random Forest models.

4 METHODOLOGY

Our method comprises three components, each addressing a specific objective: Component 1 validates HRTF prediction from manual measurements to establish the baseline (EO1), Component 2

Table 1: Anthropometric features used for HRTF prediction from HUTUBS database used in this work.

Feature	Description	Unit
Head measurements		
x1	Head width	cm
x2	Head height	cm
x3	Head depth	cm
x4	Pinna offset down	cm
x5	Pinna offset back	cm
x16	Head circumference	cm
Ear measurements (per ear)		
d1	Cavum concha height	cm
d2	Cymba concha height	cm
d3	Cavum concha width	cm
d4	Fossa height	cm
d5	Pinna height	cm
d6	Pinna width	cm
d7	Intertragal incisure	cm
d8	Cavum concha depth (down)	cm
d9	Cavum concha depth (back)	cm
d10	Crus of helix depth	cm
θ_1	Pinna rotation angle	degrees
θ_2	Pinna flare angle	degrees

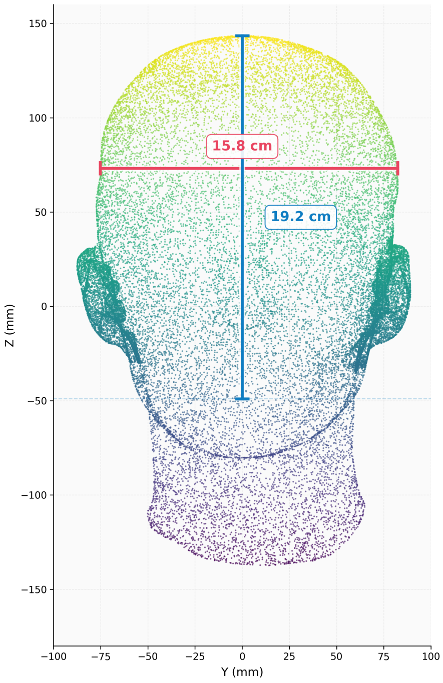

Figure 2: Front view of a subject from the HUTUBS database showing ($x1$ and $x2$) head measurements.

presents SimpleHRTF-3D for automated feature extraction from 3D meshes using PSO optimization (EO2), and Component 3 demonstrates a proof-of-concept photogrammetry-based HRTF personalization approach for enhanced consumer accessibility (EO3). Each component builds upon the previous one, creating a sequential pipeline that transforms diverse input modalities into individualized HRTFs for immersive spatial audio experiences.

4.1 Component 1: HRTF Prediction from Manual Measurements

This component addresses EO1 by validating the Random Forest models following closely Teng and Zhong (2023) [30], who demonstrated high performance in predicting HRTFs from anthropometric measurements. The methodology serves two purposes: validating the original results for reproducibility, considering the lack of available sources, and establishing a performance baseline.

We employ two model configurations: HRTF-40 uses all available anthropometric features from Table 1, including the difficult-to-measure internal parameters (d9, d10, θ_1, θ_2) plus frequency bins. In contrast, HRTF-32 excludes these four internal measurements while maintaining all other features and frequency bins. Both configurations process 64 frequency bins spanning 1-12 kHz, with frequency included as an additional input feature alongside anthropometric measurements. This frequency-enhanced approach enables a single model per ear-position combination to predict a high spectral response.

The Random Forest implementation uses 500 estimators with hyperparameters optimized based on the original work. Maximum features for splitting to follow $max\_features = \lfloor \sqrt{n_{features}} \times 3.5 \rfloor$, yielding 17 for HRTF-40 and 15 for HRTF-32. Additional parameters include $min\_samples\_split = 2$ and $min\_samples\_leaf = 5$, balancing the complexity of the model and generalizability.

Training uses the HUTUBS dataset with 90 valid subjects. The 80/10 train-test split as the original methodology. Ten models are trained for each configuration: 5 spatial positions × 2 ears, with positions selected to represent critical localization scenarios: frontal $(0, 0)$, lateral $(40, 0)$ and $(320, 0)$, and elevation $(0, 30)$ and $(0, -30)$.

Following the methodology of Teng and Zhong (2023), HRTF magnitude spectra are converted to the decibel scale before training, with models predicting responses on the logarithmic scale directly. Anthropometric measurements are used without normalization, preserving their physical units (centimeters and degrees). Out-of-bag (OOB) validation mitigates overfitting and provides reliable R^2 estimation during training, while spectral distortion metrics are evaluated on the held-out test set of 10 subjects.

Performance assessment employs two complementary metrics. The coefficient of determination (R^2) quantifies the accuracy of the prediction using a correlation-based approach:

$$R^2 = \left[\frac{\sum_{i=1}^{N} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})^2}} \right]^2$$

where y_i represents ground truth HRTF magnitude in dB, $\hat{y}_i$ the prediction, and $\bar{y}$, $\bar{\hat{y}}$ their respective means. Spectral distortion (SD) measures perceptual accuracy as the root mean square error in decibels:

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Additionally, anthropometric extraction accuracy uses the mean percentage error:

$$\epsilon = \frac{1}{M}\sum_{j=1}^{M}\frac{|x_j^{pred} - x_j^{gt}|}{x_j^{gt}} \times 100\%$$

where M denotes the number of measurements.

4.2 Component 2: SimpleHRTF-3D - Automatic Feature Extraction

This component achieves EO2 by developing automated extraction of anthropometric measurements from 3D head meshes presents challenges due to geometric variability, noise, and the need for precise landmark identification. Yan et al. (2020) proposed a full processing pipeline for anthropometric measurements from 3D body scans, achieving mean absolute errors from 2.5 to 16.0 mm depending on the specific measurement [33]. Giachetti et al. (2014) organized a benchmark competition for automatic landmark location, finding that automated methods had higher errors than interobserver measurements, with only geometrically salient landmarks achieving accuracy close to manual performance [11].

We employ PSO to automatically tune the extraction parameters, minimizing the difference between the extracted and ground-truth measurements from the HUTUBS parameters. PSO, introduced by Kennedy and Eberhart (1995), uses a population of particles exploring the parameter space guided by the best individual and collective solutions [12]. Each particle represents a complete set of extraction parameters, with its position encoding parameter values and velocity determining search direction.

To handle the multi-objective nature of minimizing errors across multiple anthropometric measurements simultaneously, we employ a weighted sum scalarization approach. The objective function computes the mean percentage error across all target measurements: $f(\mathbf{p}) = \frac{1}{N \cdot M}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{|x_{ij}^{pred}(\mathbf{p}) - x_{ij}^{gt}|}{x_{ij}^{gt}} \times 100\%$, where $\mathbf{p}$ represents the parameter vector, N is the number of subjects, M is the number of measurements per step, and x_{ij}^{pred} and x_{ij}^{gt} are the predicted values and ground-truth, respectively. This equal-weight scalarization treats all measurements as equally important, avoiding bias toward any particular anatomical feature.

For PSO hyperparameters, rather than employing automatic hyperparameter optimization, we performed a systematic grid search due to the project's time constraints. The final configuration uses a swarm size of 15 particles, 80 iterations, inertia weight $\omega = 0.9$, and best global weights $\phi_p = \phi_g = 0.5$. These values balance exploration and exploitation to avoid local minima while maintaining convergence efficiency in the 58-subject dataset.

Our novel contribution lies in decomposing the optimization into two sequential steps, recognizing that ear and head measurements require fundamentally different extraction strategies. The two-step PSO algorithm proceeds as follows (detailed in Algorithm 1):

Step A - Ear Parameter Optimization: Focuses exclusively on minimizing the extraction error for ear measurements $d5$ (height) and $d6$ (width). The optimization space includes four parameters: $bottom\_cut$ (percentile for vertical cropping), $lateral\_ear$ (percentile for selecting lateral points), ear_height (vertical offset adjustment), and $sigma$ (Gaussian smoothing kernel). This focused approach allows the algorithm to adapt specifically to ear topology challenges.

Step B - Head Parameter Optimization: Independently optimizes head measurements $x1$ (width), $x2$ (height), and $x3$ (depth). The parameters include bottom_cut (neck removal threshold), $x1\_method$ (selection between anatomical temple-based or percentile-based extraction), $x2\_cut\_adjust$ (fine-tuning vertical bounds), and $x3\_level$ (depth measurement plane). The anatomical method for $x1$ leverages domain knowledge by measuring at $65 - 85\%$ of head height in a frontal view, where temples are typically located.

PSO configuration employs 15 particles over 80 iterations with inertia $\omega = 0.9$ and cognitive/social weights $\phi_p = \phi_g = 0.5$. Step A optimizes ear extraction through vertical cropping ($bottom\_cut \in [0, 100]$), lateral selection ($lateral\_ear \in [0, 100]$), offset adjustment ($ear\_height \in [-1, 1]$), and smoothing ($\sigma \in [0, 5]$). Step B optimizes head measurements via neck removal ($bottom\_cut$), method selection ($x1\_method \in [0, 1]$), bound adjustment ($x2\_cut\_adjust \in [-5, 5]$), and depth positioning ($x3\_level \in [0, 2]$).

This decomposition reduces the combinatorial search complexity from a single 8-dimensional space to two independent 4-dimensional problems, yielding substantial computational savings through dimensional reduction. The approach leverages the anatomical independence between skull-based measurements ($x1, x2, x3$) and observable ear features ($d5, d6$), while preserving the high intra-group correlations (Pearson $\rho > 0.7$). Empirically, this strategy reduced the execution time to 34 minutes (from 58) while improving the mean error from 13.27% (single-step PSO) to 10.96% (two-step PSO), demonstrating both computational efficiency and enhanced accuracy.

For anthropometric features not directly measurable from external mesh geometry, we derive values using population-based proportional relationships. To establish these relationships, we performed an exhaustive grid search over plausible anatomical ratios using ground-truth measurements from HUTUBS. For each candidate proportion (e.g., $d1/d5$ ranging from 0.2 to 0.4 in steps of 0.001), we calculated the mean absolute error across all 58 subjects between the derived value and the actual measurement. The proportions yielding minimum population-wide error were selected:

- Ear proportions from base measurements d5 and d6: $d1 = 0.296 \times d5$ (cavum concha height), $d2 = 0.162 \times d5$ (cymba concha height), $d3 = 0.481 \times d5$ (cavum concha width), $d4 = 0.635 \times d5$ (fossa height), $d7 = 0.267 \times d6$ (intertragal incisure width), $d8 = 0.472 \times d6$ (cavum concha depth)
- Head-related offsets from x1: $x4 = 0.018 \times x1$ (vertical ear offset), $x5 = 0.025 \times x1$ (horizontal ear offset)

These proportions enable the estimation of internal ear structures and fine anatomical details while maintaining consistency with population statistics. The complete extraction pipeline combines PSO-optimized direct measurements with these derived features to produce the full set of anthropometric inputs required for HRTF prediction.

Algorithm 1 Two-Step PSO for Anthropometric Extraction

```
**Input:** 3D meshes $\mathcal{M} = \{M_1, ..., M_{58}\}$, ground truth $\mathcal{G}$
**Output:** Optimized parameters $\mathbf{p}_A^*, \mathbf{p}_B^*$

// Step A: Ear Optimization (d5, d6)
Initialize particle swarm $\mathcal{S}_A$ with 15 particles
for each particle $i$ in $\mathcal{S}_A$ do
    $\mathbf{p}_i \leftarrow random([bottom\_cut, lateral\_ear, ear\_height, \sigma])$
    $\mathbf{v}_i \leftarrow random(\mathbf{v}_{min}, \mathbf{v}_{max})$ // PSO native: velocity
end for
for iteration $t = 1$ to 80 do
    for each particle $i$ in $\mathcal{S}_A$ do
        // Problem-specific: anthropometric extraction
        $\mathcal{E}_i \leftarrow \text{ExtractEarMeasurements}(\mathcal{M}, \mathbf{p}_i)$
        $f_i \leftarrow \text{MeanPercentageError}(\mathcal{E}_i, \mathcal{G}_{ears})$
        // PSO native: update personal and global best
        if $f_i < f(\mathbf{pbest}_i)$ then
            $\mathbf{pbest}_i \leftarrow \mathbf{p}_i$
        end if
        if $f_i < f(\mathbf{gbest}_A)$ then
            $\mathbf{gbest}_A \leftarrow \mathbf{p}_i$
        end if
    end for
    for each particle $i$ in $\mathcal{S}_A$ do
        // PSO native: velocity and position update
        $\mathbf{v}_i \leftarrow \omega\mathbf{v}_i + \phi_p\mathbf{r}_1 \odot (\mathbf{pbest}_i - \mathbf{p}_i) + \phi_g\mathbf{r}_2 \odot (\mathbf{gbest}_A - \mathbf{p}_i)$
        $\mathbf{p}_i \leftarrow \mathbf{p}_i + \mathbf{v}_i$
    end for
end for
$\mathbf{p}_A^* \leftarrow \mathbf{gbest}_A$ // Optimized ear parameters

// Step B: Head Optimization (x1, x2, x3)
Initialize particle swarm $\mathcal{S}_B$ with 15 particles
[Repeat PSO process with head-specific parameters and constraints]
$\mathbf{p}_B^* \leftarrow \mathbf{gbest}_B$ // Optimized head parameters

return $\mathbf{p}_A^*, \mathbf{p}_B^*$
```

4.3 Component 3: Photo-HRTF Pipeline

This component fulfills EO3, the third component, that presents a proof-of-concept photogrammetric pipeline demonstrating HRTF personalization from consumer-grade photography. Using HUTUBS subject “pp1” (the first ID) as a case study, we captured multiple images while simulating typical user movement patterns to later use these photographs to reconstruct the 3D head geometry. The primary objective is to demonstrate the feasibility of extending our mesh-based approach to accessible image data.

We implement 3D reconstruction using COLMAP, an open-source Structure-from-Motion (SfM) [26] and Multi-View Stereo (MVS) [27] pipeline. The incremental SfM approach iteratively builds the reconstruction through feature extraction, matching, geometric verification, and bundle adjustment. The MVS stage subsequently densifies the sparse point cloud through patch-based photometric consistency optimization.

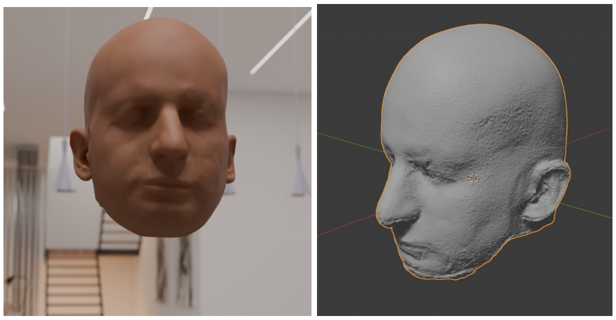

Figure 3: Photo-HRTF pipeline validation. Left: Original HUTUBS "*pp*1" ID model in simulated photography environment. Right: COLMAP reconstruction showing characteristic noise and artifacts in the photogrammetric mesh.

Table 2: Photogrammetric reconstruction specifications

Parameter	HUTUBS	COLMAP
Vertices	141,538	284,147
Faces	283,024	568,039
Density ratio	1.0×	2.0×
Capture method	Structured light	70 photos
Processing time	Real-time	~30 min
Geometric noise	Low	Moderate

Following Matuzevičius and Serackis (2022), who achieved 12 reconstruction quality improvement through background removal and frame selection strategies [15], we capture 70 images circumferentially around the subject. The acquisition protocol maintains approximately $50cm$ distance with overlapping coverage to ensure robust feature matching across views.

COLMAP processes the image sequence through: (1) SIFT feature extraction and exhaustive matching, (2) geometric verification through RANSAC-based fundamental matrix estimation, (3) incremental reconstruction initialized from a high-quality image pair, and (4) dense MVS reconstruction. The resulting point cloud undergoes surface reconstruction to generate the final mesh.

Coordinate system alignment employs three anatomical landmarks (nasion and bilateral pre-auricular points) to compute the rigid transformation between the COLMAP and HUTUBS spaces. The photogrammetric reconstruction exhibits 2.0× vertex density (284,147 vs 141,538 vertices), but suffers from characteristic noise in regions with limited photographic coverage or specular reflections. Additional reconstruction specifications are detailed in Table 2.

Lin et al. (2024) demonstrated that automated surface-based measurements achieve precision improvements of 5 plus, compared to manual methods [14]. Applying our two-step PSO extraction to photogrammetric data yields 34.2% mean measurement error, mainly concentrated in the ear regions (55.7% error) versus head measurements (5.6% error). This degradation reflects the fundamental challenges of reconstructing anatomical details from photography.

5 EXPERIMENTS SETUP

All experiments were performed on Ubuntu 22.04 LTS with an Intel Core i7-13700 processor (16 cores, 24 threads), 32GB DDR5 RAM, and NVIDIA GeForce RTX 3060 6GB GPU. The development was done in Visual Studio Code 1.101.

The implementation utilizes Python 3.10 with the following core libraries: NumPy 1.24.0 for numerical computation, Pandas 2.0.0 for data manipulation, scikit-learn 1.3.0 for Random Forest models [20], Trimesh 4.0.0 for 3D mesh processing [5], and netCDF4 [31] 1.6.0 for SOFA file handling. COLMAP 3.8 [26] provides photogrammetric reconstruction capabilities with Nvidia CUDA acceleration enabled.

6 ETHICAL CONSIDERATIONS

The HUTUBS database contains sensitive anthropometric data from 90 participants that could enable re-identification, raising privacy concerns. Although participants consented to data use, these measurements are increasingly used as biometric identifiers, requiring careful ethical handling to prevent misuse.

The limited sample size may generate poorly adjusted HRTFs for underrepresented groups. While synthetic data can mitigate sample scarcity [29], current methods prioritize database expansion over diversity considerations.

Positively, our pipeline democratizes HRTF personalization by enabling deployment on consumer hardware, reducing both computational costs and environmental impact compared to deep learning alternatives. This accessibility aligns with sustainable computing practices while maintaining prediction quality.

We disclose that Writefull was used for orthographic revision. All scientific content and methodology remain the authors' responsibility [32].

7 RESULTS

7.1 HRTF Prediction Results

Both HRTF-40 and HRTF-32 configurations achieve comparable performance ($R^2 = 89\%$, $SD = 4.5$ dB), demonstrating robustness to anthropometric feature reduction. The minimal performance difference when removing d9, d10, θ_1, and θ_2 validates our strategy of excluding these difficult-to-extract measurements from automated pipelines.

Frequency-dependent analysis reveals consistent prediction accuracy across the audible spectrum, with slightly reduced performance above 10 kHz where individual HRTF variations increase. A position-specific evaluation shows uniform R^2 values ($89.2 - 90.1\%$) across the five target locations, confirming spatial generalizability. These results closely match Teng & Zhong [30], who reported $R^2 = 90\%$ and 4.74 dB SD for their Random Forest implementation using similar anthropometric features. Our implementation achieves $R^2 = 89.8\%$ and SD=4.45 dB with the HUTUBS database [4], validating their approach while establishing the baseline for our automated extraction methods.

7.2 PSO Extraction Results

The two-step PSO decomposition achieves 10.96% mean extraction error across 58 subjects, compared to 13.27% for single-step PSO—an absolute improvement of 2.31 percentage points, representing a 17.4% relative error reduction. This improvement stems from parameter specialization: Step A optimizes ear measurements with $bottom\_cut = 15.53\%$, $lateral\_ear = 5.61\%$, $ear\_height = 0.00$, and Gaussian $\sigma = 3.00$, achieving 16.02% mean error for d5/d6 features. Step B independently optimizes head dimensions using $bottom\_cut = 26.30\%$, anatomical x1 method, $x2\_cut\_adjust = -1.0$, and $x3\_level = 1.0$, achieving remarkable 4.22% error for x1/x2/x3 measurements.

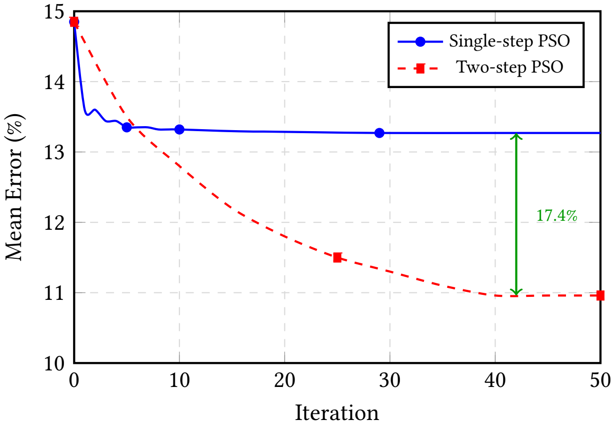

Figure 4: PSO convergence comparison showing rapid optimization in the first 50 iterations. Two-step decomposition achieves 17.4% improvement over single-step PSO through specialized parameter optimization, with both approaches maintaining stable convergence thereafter.

The two-step decomposition enables specialized configurations. Notably, head measurements benefit from higher bottom_cut values (26.30% vs 15.53%) to exclude neck regions, while ear detection requires lower lateral percentiles (5.61%) with stronger Gaussian smoothing ($\sigma = 3.00$) for robust pinna boundary detection.

Figure 4 illustrates the convergence behavior, with a single-step PSO plateauing at iteration 40 while the two-step method exploits parameter independence for superior performance. The best extraction case (Subject 31) achieves a 3.06% mean error, as visualized in Figures 5 and 6, demonstrating the potential of automated anthropometric measurement when the parameters are properly optimized.

Computational efficiency remains practical with 34 minutes total execution time for the complete 58-subject dataset using 15 particles and 80 iterations per step. This demonstrates the viability of PSO optimization for real-world deployment where extraction parameters must adapt to varying mesh characteristics.

7.3 Photo-HRTF Results

The complete photogrammetric pipeline achieves $R^2 = 67.9\%$ and $SD = 5.64$ dB when predicting HRTF from smartphone-captured images. Initial mesh reconstruction using COLMAP [26] produces 284,000 vertices with 34.2% anthropometric extraction error, primarily due to reconstruction artifacts and noise in the ear regions. Head measurements show excellent accuracy (average error 5.6%),

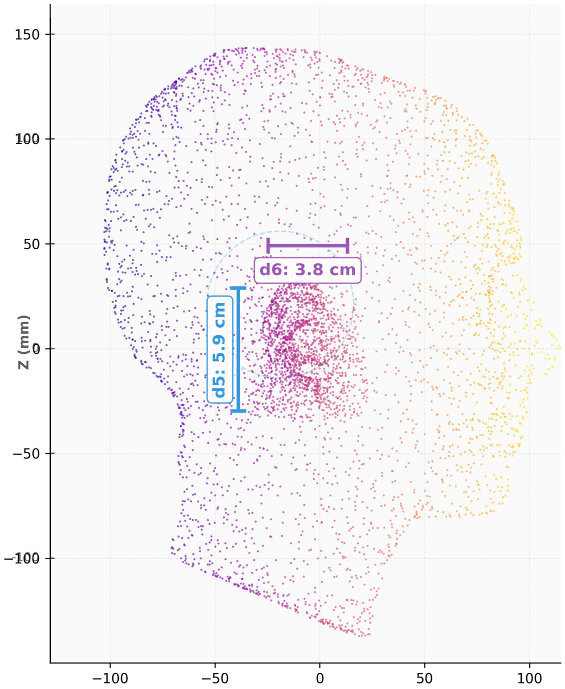

Figure 5: PSO-optimized anthropometric extraction on subject 31 showing lateral view showing the $d5$ and $d6$ measurements.

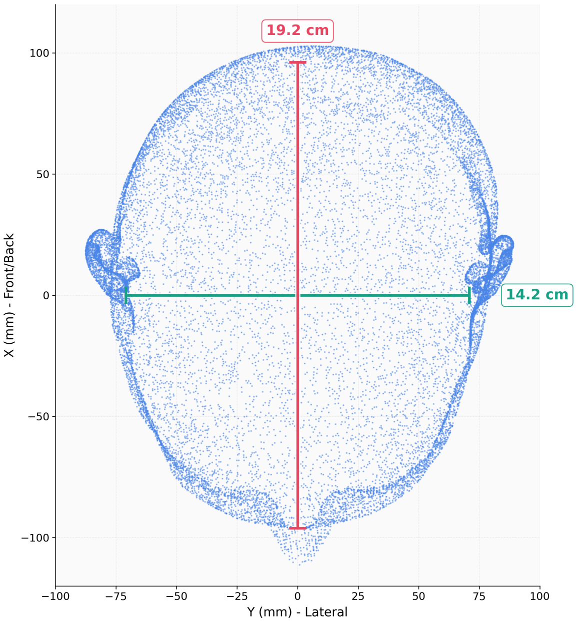

Figure 6: PSO-optimized anthropometric extraction on subject 31 showing top view.

while ear measurements present significant challenges (average error 55.7%), particularly for ear width (d6) extraction.

The pipeline demonstrates a 24.0% degradation from baseline performance (R^2 = 89.3%) when using photogrammetric meshes instead of high-quality scans. This degradation is in line with the magnitude of the extraction error, as each 1% of measurement error translates to approximately 0.7% reduction in the accuracy of the HRTF prediction. Future improvements in 3D reconstruction quality, particularly for ear regions, could significantly enhance prediction accuracy. Lin et al. [14] recently proposed specialized algorithms for precise auricle measurement achieving submillimeter accuracy, suggesting potential pathways toward the theoretical limit established by mesh-based extraction.

8 DISCUSSION AND FUTURE WORK

This work successfully addresses three objectives. Our first objective validated the Random Forest methodology on HUTUBS data, achieving the targeted R^2 = 89.8% and SD = 4.45 dB performance with manual measurements. This validation confirms the reproducibility of machine learning approaches for HRTF prediction and provides a baseline for subsequent automated methods.

The second objective, automated extraction from 3D meshes through SimpleHRTF-3D, demonstrates that PSO-optimized measurement extraction can maintain high prediction accuracy (R^2 = 89.3%, SD = 5.04 dB) despite a 10.96% extraction error. The novel two-step PSO decomposition has been proven to be particularly effective, reducing the mean error by 17.4% compared to single-step optimization. This improvement validates our hypothesis that head and ear measurements benefit from independent parameter optimization due to their distinct geometric characteristics.

As a third objective and as proof-of-concept, the photogrammetric pipeline extends accessibility by enabling HRTF prediction from consumer photographs. Although R^2 = 67.9% with SD = 5.64 dB, represents a degraded performance compared to high-quality meshes, it substantially exceeds generic HRTF models and demonstrates feasibility for consumer applications where acoustic measurement facilities remain inaccessible.

We highlight that the main technical contribution of our work is the two-step PSO approach, completing 58 subjects in 34 minutes. However, internal ear measurements rely on population averages, potentially introducing errors for atypical anatomy. Photogrammetric reconstruction shows 55.7% ear measurement error, highlighting challenges in consumer photography applications. Validation remains limited to HUTUBS database.

As the next steps in our research, we aim to: (1) explore gradient boosting and lightweight neural networks, (2) cross-database validation with CIPIC and ARI datasets, (3) perceptual evaluation in VR environments, and (4) integration with consumer applications.

Code and data are available at https://github.com/augustompm/SimpleHRTF-3D/. Contributions are welcome.

ACKNOWLEDGMENTS

This work was supported by CAPES - Finance Code 001, CNPq (grant 312165/2025-0), and FAPERJ (grant JCNE 2021 E-26/201.451/2022).

REFERENCES

[1] Laith Abualigah. 2025. Particle Swarm Optimization: Advances, Applications, and Experimental Insights. *Computers, Materials and Continua* 82, 2 (2025), 1539–1592. https://doi.org/10.32604/cmc.2025.060765

[2] Ramona Bomhardt, Hark Braren, and Janina Fels. 2017. Individualization of head-related transfer functions using principal component analysis and anthropometric dimensions. *Proceedings of Meetings on Acoustics* 29, 1 (09 2017), 050007. https://doi.org/10.1121/2.0000562

[3] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[4] Fabian Brinkmann, Manoj Dinakaran, Robert Pelzer, Peter Grosche, Daniel Voss, and Stefan Weinzierl. 2019. A Cross-Evaluated Database of Measured and Simulated HRTFs Including 3D Head Meshes, Anthropometric Features, and Headphone Impulse Responses. *Journal of the Audio Engineering Society* 67 (09 2019), 705–718. https://doi.org/10.17743/jaes.2019.0024

[5] Dawson-Haggerty et al. [n. d.]. *trimesh*. https://trimesh.org/

[6] Brinkmann Fabian, Dinakaran Manoj, Pelzer Robert, Wohlgemuth Jan Joschka, Seipel Fabian, Voss Daniel, Grosche Peter, and Weinzierl Stefan. 2019. The HUTUBS head-related transfer function (HRTF) database repository. https://doi.org/10.14279/depositonce-8487

[7] Davide Fantini, Michele Geronazzo, Federico Avanzini, and Stavros Ntalampiras. 2025. A Survey on Machine Learning Techniques for Head-Related Transfer Function Individualization. *IEEE Open Journal of Signal Processing* 6 (2025), 30–56. https://doi.org/10.1109/OJSP.2025.3528330

[8] Janina Fels and Michael Vorlaender. 2009. Anthropometric Parameters Influencing Head-Related Transfer Functions. *ACTA ACUSTICA united with ACUSTICA* 95 (03 2009), 331–342. https://doi.org/10.3813/AAA.918156

[9] Mariza Ferro, Gabrieli D. Silva, Felipe B. de Paula, Vitor Vieira, and Bruno Schulze. 2023. Towards a sustainable artificial intelligence: A case study of energy efficiency in decision tree algorithms. *Concurrency and Computation: Practice and Experience* 35, 17 (2023), e6815. https://doi.org/10.1002/cpe.6815 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.6815

[10] Iyyakutti Iyappan Ganapathi, Syed Sadaf Ali, Surya Prakash, Ngoc-Son Vu, and Naoufel Werghi. 2023. A Survey of 3D Ear Recognition Techniques. *ACM Computing Surveys, Volume 55, Issue 10* 55, 10, Article 204 (2023), 36 pages. https://doi.org/10.1145/3560884

[11] A. Giachetti, E. Mazzi, F. Piscitelli, M. Aono, A. Ben Hamza, T. Bonis, P. Claes, A. Godil, C. Li, M. Ovsjanikov, V. Pătrăucean, C. Shu, J. Snyders, P. Suetens, A. Tatsuma, D. Vandermeulen, S. Wuhrer, and P. Xi. 2014. Automatic location of landmarks used in manual anthropometry. In *Eurographics Workshop on 3D Object Retrieval* (Strasbourg, France) *(3DOR 14)*. Eurographics Association, Goslar, DEU, 93–100.

[12] J. Kennedy and R. Eberhart. 1995. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, Vol. 4. 1942–1948 vol.4. https://doi.org/10.1109/ICNN.1995.488968

[13] Geon Woo Lee and Hong Kook Kim. 2018. Personalized HRTF Modeling Based on Deep Neural Network Using Anthropometric Measurements and Images of the Ear. *Applied Sciences* 8, 11 (2018). https://doi.org/10.3390/app8112180

[14] Yangyang Lin, Johannes G. G. Dobbe, Nadia Lachkar, Elsa M. Ronde, Theo H. Smit, Corstiaan C. Breugem, and Geert J. Streekstra. 2024. A three-dimensional algorithm for precise measurement of human auricle parameters. *Scientific Reports* 14, 1 (2024), 10760. https://doi.org/10.1038/s41598-024-61351-5

[15] Dalius Matuzevičius and Artūras Serackis. 2022. Three-Dimensional Human Head Reconstruction Using Smartphone-Based Close-Range Video Photogrammetry. *Applied Sciences* 12, 1 (2022). https://doi.org/10.3390/app12010229

[16] Microsoft Research. 2025. Spatial Audio–Project Overview. https://www.microsoft.com/en-us/research/project/spatial-audio/. Industrial context: spatial-audio applications, personalised HRTF challenges, and technology transfer to Windows 10, Xbox One, Soundscape and HoloLens.

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

[18] Parham Mokhtari, Ryouichi Nishimura, and Hironori Takemoto. 2008. Toward HRTF personalization: an auditory-perceptual evaluation of simulated and measured HRTFs. In *Proceedings of the 14th International Conference on Auditory Display*.

[19] Md. Mursalin and Syed Mohammed Shamsul Islam. 2021. Deep Learning for 3D Ear Detection: A Complete Pipeline From Data Generation to Segmentation. *IEEE Access* 9 (2021), 164976–164985. https://doi.org/10.1109/ACCESS.2021.3129507

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[21] Robert Pelzer, Manoj Dinakaran, Fabian Brinkmann, Steffen Lepa, Peter Grosche, and Stefan Weinzierl. 2020. Head-related transfer function recommendation based on perceptual similarities and anthropometric features. *The Journal of the Acoustical Society of America* 148, 6 (12 2020), 3809–3817. https://doi.org/10.1121/10.0002884

[22] Myllena Prado, Lucas Althoff, Sana Alamgeer, Alessandro Rodrigues e Silva, Ravi Prakash, Marcelo M. Carvalho, and Mylène C. Q. Farias. 2022. 360RAT: A Tool for Annotating Regions of Interest in 360-degree Videos. In *WebMedia '22* (Curitiba, Brazil). Association for Computing Machinery, 272–280. https://doi.org/10.1145/3539637.3557930

[23] Surya Prakash and Phalguni Gupta. 2012. An efficient ear localization technique. *Image and Vision Computing* 30, 1 (2012), 38–50. https://doi.org/10.1016/j.imavis.2011.11.005

[24] Sheldon M. Ross. 2021. *Introduction to Probability and Statistics for Engineers and Scientists* (6 ed.). Academic Press, Amsterdam. 32–33 pages.

[25] Radu Bogdan Rusu and Steve Cousins. 2011. 3D is here: Point Cloud Library (PCL). In *2011 IEEE International Conference on Robotics and Automation*. 1–4. https://doi.org/10.1109/ICRA.2011.5980567

[26] Johannes L. Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4104–4113. https://doi.org/10.1109/CVPR.2016.445

[27] Johannes L. Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Computer Vision – ECCV 2016*. 501–518. https://doi.org/10.1109/CVPR.2016.445

[28] Aleph Silveira, Roope Raisamo, Fotios Spyridonis, Alexandra Covaci, George Ghinea, and Celso A. S. Santos. 2023. Guidelines for conducting biofeedback-enhanced QoE studies in mulsemedia-enhanced virtual reality. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web* (Ribeirão Preto, Brazil) *(WebMedia '23)*. Association for Computing Machinery, New York, NY, USA, 32–40. https://doi.org/10.1145/3617023.3617029

[29] F. Stärz, S. Van De Par, S. Roskopf, L. O. H. Kroczek, A. Mühlberger, and M. Blau. 2025. Comparison of binaural auralisations to a real loudspeaker in an audiovisual virtual classroom scenario: Effect of room acoustic simulation, HRTF dataset, and head-mounted display on room acoustic perception. *Acta Acustica* 9 (2025), 31. https://doi.org/10.1051/aacus/2025012

[30] Yuqi Teng and Xiaoli Zhong. 2023. An Individualized HRTF Model Based on Random Forest and Anthropometric Parameters. In *2023 15th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*. 143–146. https://doi.org/10.1109/IHMSC58761.2023.00041

[31] Unidata. 2025. *Network Common Data Form (netCDF)*. Boulder, CO. https://doi.org/10.5065/D6H70CW6 [software].

[32] Writefull. 2025. Writefull for Overleaf - AI Language Feedback for LaTeX. https://www.writefull.com/writefull-for-overleaf.

[33] Song Yan, Johan Wirta, and Joni-Kristian Kämäräinen. 2020. Anthropometric clothing measurements from 3D body scans. *Machine Vision and Applications* 31, 1 (jan 2020), 7. https://doi.org/10.1007/s00138-019-01054-4