Audio Engineering Society

Convention Paper 10685

Presented at the AES 156th Convention
2024 June 15–17, Madrid, Spain

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

An approach for mesh-based generation of spatially dense, lowpass-filtered, individualized HRTFs using dynamic data acquisition

Andrea Quattrini[1], Alberto Vancheri[1], Tiziano Leidi[1], Thierry Heeb[1], Loris Grossi[1], Gilles Oldano[1], and Noah Spagnoli[1]

[1]*SUPSI, Institute for Information Systems and Networking, Polo Universitario Lugano, Campus Est, Via la Santa 1, 6962 Lugano-Viganello, Switzerland*

Correspondence should be addressed to Tiziano Leidi (tiziano.leidi@supsi.ch)

ABSTRACT

Head Related Transfer Functions (HRTFs) capture the binaural information required for correct identification of a sound source position in 3D space. They are individual to each user and heavily depend on the direction of arrival of the sound from the considered source. Acquisition of personalized HRTFs for all possible directions is a lengthy process requiring carefully calibrated measurements which is difficult to apply in practical situations. In this paper we propose a data-driven, machine-learning based approach for the generation of personalized, direction-dense, lowpass-filtered HRTFs computed from a mesh of the user's head and from a non-uniform set of dynamic data measurements. We present the different steps of our approach and show results from laboratory experiments. Comparison with a state-of-the-art, BEM-based HRTF generation method confirms the effectiveness of the proposed solution.

1 Introduction

Head Related Transfer Functions (HRTFs) describe how the ears of a listener receive sound from a sound source. They represent the information used by our auditory system to properly assess the position of a sound source in a 3D environment as shown by Bregman [1] or Majdak et al [2]. In this paper, we present a data-driven, machine-learning based HRTF acquisition and generation scheme for use in the context of our geometrical approach to crosstalk cancellation [3]. This scheme fully supports the multi-band version of our crosstalk cancellation principle [4] and is a key element for its adaptive version where the user can take arbitrary positions [5]. The proposed approach is general purpose and could prove useful in other application areas, in particular when combined with spatialization and binaural technologies for immersive 3D audio rendering in multimedia, virtual reality and gaming.

Every individual has unique HRTFs. Sound waves travelling from the source to the listener are affected by the listener's upper body, head and ears, and thus the resulting transfer functions not only depend on the direction of arrival of the sound, but also greatly change from one person to the other.

Personalized HRTFs offer the promise of the most realistic 3D audio reproduction, and it has been shown that

use of generic HRTFs generally results in non-optimal sound source localization perception, as studied by Majdak et al. [6]. Unnatural coloration and other artefacts have also been reported when using non-personalized HRTFs, for instance in the work of Brinkmann [7]. Different approaches for acquiring or generating personalized HRTFs have been studied, ranging from simple recording using in-ear microphones for static listener positions to complex multidirectional schemes. Promising approaches for the generation of personalized HRTFs include computation of HRTFs from a 3D mesh of the head as presented by Guezenoc and Seguier [8] or from pictures or anthropometric features as reviewed by Pollack et al. [9].

Acquiring and storing individualized HRTFs for every possible direction of arrival is time consuming and resources hungry. Therefore, a solution based on interpolation from a non-uniform set of dynamic measurements and maintaining proper perception of direction of arrival by the listener is highly desirable. Many methods for HRTFs interpolation have been proposed such as the time-domain bilinear technique by Savioja et al. [10], its extension adding auxiliary inter-positional transfer functions by Biscainho et al. [11] or the works of Bruschi et al. based on inverse distance weighting [12]. Frequency-domain based interpolation of HRTFs has also been explored. For instance, Hartung et al. [13] propose a method exhibiting increased performance compared to time-domain approaches, albeit at higher computational cost. In [14], Gamper discloses another frequency-domain interpolation scheme based on gains applied to neighbouring HRTFs and on phase information derived from a spherical head model. In addition, machine learning techniques are now commonly used for HRTFs generation. For instance, in [15], Zurale and Dubnov use neural networks for spatial up-sampling of sparse HRTFs. Machine learning has also been successfully used for generation of personalized HRTFs from scanned head geometry by Wang et al. [16] or from ear shapes by Zhou et al. [17]. In [18] Fantini et al. propose an approach where machine learning is applied to median plane HRTF generation based on a user's pinna anthropometric features extracted from a 3D mesh. During the review of this paper, a similar study was proposed by Zhao et al. [19]. We reference this work as supportive evidence for our approach, as it aligns with our methodology.

The acquisition and generation approach that we present in this paper is based on a combination of recording, tracking, scanning, fitting and interpolation techniques that leverages several machine learning models. In section 2, we describe the procedure followed to generate a triangulated mesh of a real head using a 3D sensor. This is a crucial, initial step for the subsequent computations. In section 3, we describe the experimental acquisition and processing method used to produce datasets that feed the machine-learning models for generation the final HRTFs. The procedure followed to acquire these data is based on audio recordings from in-ear microphones worn by a user freely moving in space. This requires careful tracking of the head position during motion. The datasets acquired by the proposed procedure are in fact noisy HRTFs obtained with source positions randomly sampled among the possible directions. We refer to this dataset as the "scattered HRTFs". Section 4 is devoted to the description of the machine-learning based interpolation procedure of the data contained in the scattered HRTF dataset that leads to personalized HRTFs in any direction. We call these HRTFs the "experimental HRTFs". Finally, in section 5, experimental HRTFs are compared with equivalent HRTFs obtained by numerically solving the Kirchhoff-Helmoltz equation using a Boundary Element Method (BEM) based approach. These numerical HRTFs are designated as the "theoretical HRTFs". The mesh constructed in section 2 is used both in the construction of the experimental HRTFs via machine-learning methods and in the computation of the theoretical BEM-based HRTFs. To summarize our contributions in this paper, we:

- Present a strategy based on upper-body 3D reconstruction from which we extract a topological mesh used both for the interpolation of laboratory HRTFs and for the calculation of theoretical HRTFs through BEM.
- Introduce a technique for dynamic data acquisition in the laboratory that streamlines the dataset creation process.
- Release our dataset for future purposes and for community testing.

2 Head Scanning and Fitting

In this section, we present our approach to 3D scanning of a listener's head and the subsequent construction of a triangulated mesh from the resulting point cloud.

Lately, the evolution of 3D reconstruction techniques has been marked by significant advancements. Long-established solutions include photogrammetry [20], whereas recent developments profit from Neural Radiance Fields (NeRF) [21] and Gaussian Splatting [22], which are currently playing a pivotal role. These techniques, all grounded in the use of RGB images, have significantly advanced the ability to produce complex three-dimensional reconstructions. However, current state-of-the-art RGB image-based 3D methods share a crucial limitation: scale ambiguity. While good at capturing shapes and textures in detail, their output can be accurate only up to a scale factor. This limitation becomes especially problematic in applications where precise, absolute metric measurements are crucial. Considering this fact, our research has pivoted towards the use of 3D sensors-based reconstruction techniques. These methods differ fundamentally from their image-based counterparts by their ability to capture not only the geometry and texture of objects but also their exact dimensions, thus overcoming the scale ambiguity issue. Common outputs of these 3D reconstruction techniques are point clouds or meshes. In the case of 3D sensors, the point cloud is a direct output of the measurement instrument and the mesh consists in 3D points connected by polygons to form a surface. However, 3D sensors also come with a limitation: their Field of View (FoV) is restricted, allowing them to capture and reconstruct only part of a scene at a time. To overcome this limitation, it is imperative to maneuver the sensor in relation to the object or vice versa. Employing Simultaneous Localization and Mapping (SLAM) techniques to discern relative movements and to align all measurements to a unified reference system, allows for comprehensive reconstruction of the full scene [23]. The density and total number of points in the final point cloud or mesh, which are critical factors for the quality of 3D models, depend on several variables. For depth image-based methods, the resolution of the depth sensor plays a key role. The resulting large-scale data from these processes often pose challenges for subsequent applications, such as storage, processing, and analysis. To address this challenge, a well-known approach is to reduce the dimensionality of the data to a more manageable level. Dimensionality reduction in this context does not merely imply a decrease in the number of points or density of the mesh, but involves transforming the data into a form retaining essential geometric and textural features while being more efficient to handle. Techniques such as principal component analysis (PCA), deep learning autoencoders, or parametric models like FLAME [24] or SMPL [25] can be employed for this purpose.



Fig. 1: LiDAR-equipped tablet used to scan the upper-body of the individual.

In our approach, the upper-body scanning of individuals is conducted using a Light Detection And Ranging (LiDAR) sensor mounted on a tablet (see figure 1), and positioned to allow for real-time observation of acquisition results as they are grabbed. This device is also equipped with an integrated RGB camera. The scanner initial depth accuracy is approximately $5mm$. However, this value can be significantly enhanced through a series of processes, which will be detailed later in this paper. These steps include temporal filtering and advanced fitting techniques, each contributing to the overall improvement of the scanner accuracy. Its FoV is aligned with the depth sensor through a calibration and registration procedure. The resulting RGBD sensor has to be moved all around the person to achieve a complete reconstruction. In this phase, individuals are instructed to remain as still as possible from the start to the end of the acquisition procedure. Minor movements are addressed during the processing of point clouds. The acquisition frequency is set to 1Hz to enable the LiDAR to undergo sufficient roto-translation to cover partially different facial regions for subsequent acquisitions. The relative position between an acquisition and the next one is computed using the Perspective-n-Point algorithm [26], with matched keypoints utilizing Oriented FAST and Rotated BRIEF (ORB) [27] as descriptors on RGB images. The estimated position is then refined using ICP [28] on depth images (i.e. point clouds) to

compensate for micro-movements. These calculated roto-translations are finally used for the construction of a pose graph. A pose graph is composed of two principal components: nodes and edges. Each node corresponds to a geometry P_i and is associated with a pose matrix T_i, which maps P_i to a global coordinate system. The global space is defined by the reference space P_0, making T_0 the identity matrix. The initialization of other pose matrices is achieved by cumulative transformations between adjacent nodes. Edges in the pose graph represent the connections between two overlapping nodes, or pieces of geometry. Every edge embeds a transformation matrix $T_{i,j}$ that serves to align the source geometry P_i with the target geometry P_j. To enhance the accuracy, each image is tested and computed against not just the two preceding images but also against the two subsequent ones. This approach facilitates the development of a pose graph represented by an adjacency matrix reflecting heterogeneously connected nodes. Unlike a graph with solely direct neighbor connections, this structure includes additional interconnections among nodes, a feature that has been demonstrated to significantly enhance the quality of the resultant 3D reconstruction. The pose graph is then optimized using the Levenberg-Marquardt technique [29], and the corresponding mesh is obtained through a Truncated Signed Distance Function (TSDF), created with the point clouds and the transformations derived from the optimization process. An example of the resulting reconstruction is shown in figure 2.

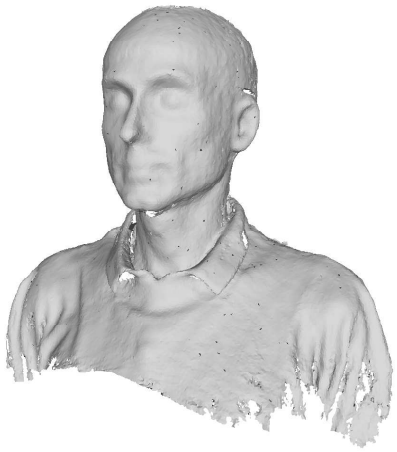
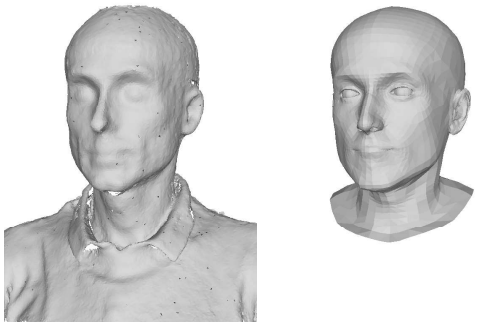


Fig. 2: Example of the scan of an upper body carried out with the reconstruction method described in section 2.

FLAME [24], which stands for Faces Learned with Articulated Model and Expressions, is a topological, machine-learning based model primarily used for detailed 3D reconstruction of human heads, capturing facial expressions, head poses, and even subtle anatomical differences. As a topological model, FLAME defines a consistent mesh topology that remains invariant across different identities and expressions. This consistency allows for the analysis and comparison of head features across a wide range of subjects in variable contexts. The strength of FLAME lies in its ability to model complex variations in head, pose and facial topology with a relatively small number of control parameters. This is achieved through a blend of linear shape spaces that account for identity, expressions, and pose, which are derived from a large dataset of 3D scans of human heads. The topology of the FLAME model ensures that vertex correspondence is maintained across different models, which is crucial for applications that require vertex-level comparison, such as facial animation and morphological studies. In the development of our framework, the decision to employ the FLAME model over other alternatives such as 3D Morphable Models (3DMM) [30] or Basel Face Model [31] was made after careful consideration of the specific requirements of our project. A significant advantage of FLAME lies in its comprehensive 360-degree modeling of the human head. It's worth noting that FLAME does not have a uniform distribution of vertices and triangles; rather, it consists of 5023 vertices, of which 1300 are dedicated to the ear areas. This level of detail is one of the key reasons why FLAME is well-suited for our methodology. The conversion of a 3D scan to a FLAME model (called fitting) is thorough and detailed, involving a sequence of steps designed to ensure precision. It begins with an initial pre-alignment phase, followed by both rigid and non-rigid alignment phases to adjust the parameters of the FLAME model. This fine-tuning is key to achieving an optimal overlay, capturing the intricate details of head dimensions (including those of nose, ears and mouth), head pose, facial expressions, and the overall facial structure, ensuring a close match between the model and the real-world scan. At this stage, the FLAME model, a comprehensive watertight 3D model trained on thousands of head reconstructions, not only fits the sparse point cloud data but also realistically supplements missing information that may be unavailable due



(a) 3D Reconstruction **(b)** FLAME Fitting

Fig. 3: On the left, the 360-degree reconstruction of the person's face, on the right, its corresponding FLAME fitting.

to occlusions during the LiDAR scanning process. For mathematical details, please refer to [24]. An example of head fitting is shown in figure 3. The topological nature of FLAME allows for a consistent reference system across all scanned heads. Here, it's important to note that the position of the FLAME reference system does not align with our convention. Details on our reference system can be found in the section 3. However, the topological nature of FLAME allows us to easily transform the mesh from one reference system to the other. This consistency is crucial for correlating the resulting head mesh with laboratory-acquired data, as will be described in the next section. Our convention leverages the coordinates of predefined vertices to construct a transformation matrix. This matrix is then applied to reposition the mesh within the specified reference system. Specifically, these vertices (the tip of the nose and the two ears) are consistently identified within the FLAME topology by their fixed indices.

3 HRTF Data Acquisition and Processing

In this section, we provide an overview of the experimental procedure and processing method used to obtain HRTF data. Our approach uses multi-sine based responses acquisitions to produce HRTFs for different user positions and orientations. Simultaneously, a Motion Capture (MoCap) system is used to track the user position and orientation, as explained in more details below. All the recordings conducted in our laboratory are carried out using the following pieces of equipment: a computer, an audio interface, a loudspeaker, in-ear microphones, and a measurement microphone. The recordings have been performed at a sampling frequency of $F_s = 48kHz$.

Multi-sine [32] is a technique based on FFT allowing for real-time acquisition of a system's frequency and impulse responses. For an FFT of length N (given by the desired frequency resolution), the excitation signal consists in all frequencies corresponding to the bins of the FFT, at equal level, and with phases spread for crest factor optimization. Its period is obviously equal to N. Such an excitation signal has a flat frequency response similar to the one of a pulse, but phase spreading allows for higher energy per frequency bin, hence higher signal-to-noise ratio and robustness to noise.

For responses acquisitions, both the excitation signal X and the resulting system output Y are averaged over a number $m \geq 1$ of periods to produce signals X_m and Y_m, each of length N. The frequency and impulse responses of the system are then simply given by:

$$F(\omega) = FFT(Y_m)/FFT(X_m) \quad (1)$$

$$IR(n) = IFFT(F(\omega)) \quad (2)$$

During acquisition, a human user wearing in-ear microphones L (left) and R (right) slowly moves in space while a source S continuously emits the multi-sine signal of period N. Our experiments have a duration of some minutes, but could last longer to acquire more data. The signal is simultaneously recorded by an external, fixed location, reference measurement microphone E. The recordings of the in-ear microphones and of the external microphone are designated by $x_L(t)$, $x_R(t)$ and $x_E(t)$ respectively.

Equations 1 and 2 are applied to short segments of duration $\Delta t = Nm$ extracted from the recordings $x_L(t)$, $x_R(t)$ and $x_E(t)$, and associated to different positions of the user during motion. Working with a moving user imposes some constraints on the period N of the multi-sine and the number m of periods used for averaging. Large values of N enable high frequency resolution but minimization of artefacts due to user movements calls for low values of m and N. We found an acceptable trade-off between high frequency resolution and short time duration Δt of audio segments with $N = 8192$ and $m = 1$, corresponding to a frequency resolution $\Delta f = 5.9Hz$ and $\Delta t = 0.17s$ at a sampling rate of $48kHz$.

Let θ and ϕ be the time dependant azimuth and elevation angles of the source S in a reference system co-moving with the head of the user. The movement of the user follows a defined scheme in such a way that values of azimuth and elevation angles are sampled as uniformly and densely as possible.

During the time span of an audio segment, the incidence angles θ and ϕ undergo variations $\Delta\theta$ and $\Delta\phi$ and, as a consequence, the HRTF obtained by processing the audio segment is the result of a sort of time average. This could lead to blurred HRTFs. The tracking system enables us to detect the variation of the head configuration during a given audio segment and then to evaluate the variations $\Delta\theta$ and $\Delta\phi$ of the incidence angles. Variations $\Delta\theta$ and $\Delta\phi$ above given thresholds enable us to discard data where the head configuration has undergone a large variation during the time segment.

In order to obtain HRTFs from this setting, the time dependant values of θ and ϕ need to be measured and associated to the recordings of the microphones. As previously introduced, values of θ and ϕ during motion are measured using a MoCap system. MoCap systems provide high precision tracking of the positions of point markers rigidly located on a moving body. Our system consists of six 0.3 MP cameras that cover a volume roughly $4m$ x $4m$ x $3m$, equipped with an active infrared lighting system. It boasts a 3D measurement precision of up to $0.5mm$ at 100 FPS. We designate the set of markers rigidly connected with the head of the user as the Head Rigid Body (HRB) markers and the associated reference system co-moving with the head as the HRB co-moving system. The tracked path of the HRB markers in space allows for the reconstruction of translations and rotations of the HRB co-moving system with respect to the laboratory's reference system. Unfortunately, the axes and origin of the HRB co-moving system are determined by the positions of the HRB markers and not by natural markers of a human head, such as ears and nose. This makes the results scarcely interpretable. In addition, the results produced with these tracking data are not reliably reproducible because, when the user needs to wear the HRB markers again, the position of the HRB co-moving system may change considerably. As a final issue, comparison with BEM simulations (see section 4 and 5) requires that the same reference co-moving system is used both in laboratory and in simulations. In order to overcome these difficulties, we start with a preliminary calibration session, where a static user wears three additional markers for three key points on the head (left and right ears, L and R respectively, and nose N) besides the HRB markers. The positions of these points enable the construction of a natural reference system (NRS) $Oxyz$ co-moving with the user's head as follows: the plane xy is defined by the triangle LRN, with the x axis passing through L and R and oriented from L to R. The y axis is orthogonal to x, passes through N and oriented towards the nose N. Finally, the z axis is defined by the requirement that the system $Oxyz$ be orthogonal and right-handed. This system is the same as the one used in the processing of the head mesh as described in the previous section. By doing so, the relation between HRB and NRS markers can be worked out and used to refer all the tracking data to the NRS. Such data is both reproducible and easily interpretable.

To obtain the HRTF, we first low-pass the signals $x_L(t)$, $x_R(t)$ and $x_E(t)$ to improve the subsequent analysis. In our experiment we used a cutoff frequency of $7000Hz$, as we are interested in HRTF for crosstalk cancellation, but it could be modified in other exploitation contexts. From these signals, we uniformly extract segments of length N, each associated to a value of θ and ϕ obtained from tracking data linearly fitted for times corresponding to the time span of the segment. The j-th segment extracted from $x_L(t)$ is defined by $x_L^{(j)}(t) = x_L(t+jh)$ with $0 \leq t \leq N-1$. In our experiment, we set a value of h equivalent to 0.05 seconds. Similar definitions apply to $x_R^{(j)}(t)$ and $x_E^{(j)}(t)$.

For each of these segments, we compute the transfer functions $H_{LE}^{(j)}(\omega) = \frac{X_L^{(j)}(\omega)}{X_E(\omega)}$ and $H_{RE}^{(j)}(\omega) = \frac{X_R^{(j)}(\omega)}{X_E(\omega)}$ connecting the external microphone with the in-ear microphones. The impulse responses derived from $H_{LE}^{(j)}(\omega)$ and $H_{RE}^{(j)}(\omega)$ by IFFT are then windowed with a Tukey window to obtain anechoic impulses responses $h_{LE,a}^{(j)}(t)$ and $h_{RE,a}^{(j)}(t)$ excluding early reflections and other contributions due to the acoustics of the environment (the subscript a refers to the anechoic contributions). The choice of the parameters of the window (length, position with respect to the peak of the pulse and cosine fraction) must be taken carefully. For instance, a too short cosine fraction could lead to artifacts due to a too abrupt windowing, whereas a too long window, perhaps as a consequence of the choice of larger cosine fraction, could lead to the inclusion of contributions of early

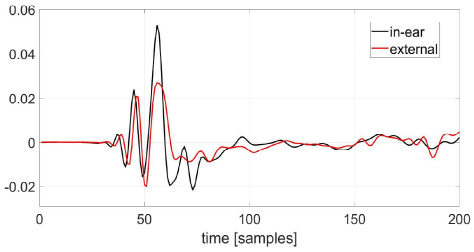


Fig. 4: Comparison of the filtered signals recorded by the left in-ear microphone and the external reference microphone for incidence angles of the sound near to $\theta = 1^\circ$ and $\phi = 5^\circ$. The peaks of the signals have been aligned for visualisation purposes.

reflections. In our current implementation, the parameters are the same for each configuration of the head, but in a more advanced implementation they should be adapted to the incidence angles. The FFTs of $h^{(j)}_{LE,a}(t)$ and $h^{(j)}_{RE,a}(t)$ have following form:

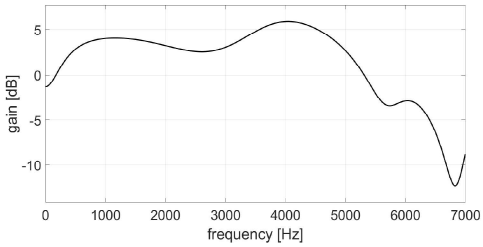
$$H^{(j)}_{LE,a}(\omega) = g^{(j)}_L(\omega)\exp(-i\omega\tau^{(j)}_L(\omega))) \qquad (3)$$

$$H^{(j)}_{RE,a}(\omega) = g^{(j)}_R(\omega)\exp(-i\omega\tau^{(j)}_R(\omega))) \qquad (4)$$

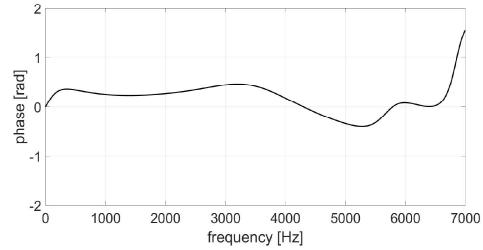
where $g^{(j)}_L(\omega)$, $g^{(j)}_R(\omega)$ are the gains factors in position j at the left and right ear respectively and $\tau^{(j)}_R(\omega)$, $\tau^{(j)}_L(\omega)$ are the corresponding phase delays. Finally, the values of $\tau^{(j)}_L(\omega)$, $\tau^{(j)}_R(\omega)$, $g^{(j)}_L(\omega)$ and $g^{(j)}_R(\omega)$ are corrected in order to move the reference point from the position of the external microphone E to the origin of the co-moving system as required by the definition of HRTFs. The use of an external microphone E together with a tracking system to detect the head configuration makes controlling the distance between the head and the loudspeaker unnecessary.

Figure 4 shows an example of the signals $x_L(t)$ and $x_E(t)$. Figure 5 shows the amplitude and phase of $H_{LE,a}(\omega)$ with respect to frequency, after filtering and windowing.

The recording procedure with a moving user enables collection of a large amount of scattered data. In our experiments we produce up to 8000 positions (see figure 6) in a relatively short time span, but the extraction



(a) Amplitude of $H_{LE,a}$



(b) Phase of $H_{LE,a}$

Fig. 5: Amplitude and phase of the anechoic transfer function $H_{LE,a}$ from the reference microphone E to the left in-ear microphone L for incidence angles of the sound near to $\theta = 1^\circ$ and $\phi = 5^\circ$. The signals received at the left ear L and at the external microphone E have been aligned for post-processing purposes.

procedure is configurable for different situations. However, compared to a controlled procedure using a static dummy head, the proposed solution can lead to noisier data. First, changes in head position during a period of the multi-sine can result in a sort of blurred photography effect, especially for positions where the HRTF is strongly sensitive to changes in θ and/or ϕ. Second, the recording itself takes place in non-optimally controlled conditions leading to noisy data and outliers, for instance from parasitic noise resulting from the user's movements. For these reasons we use filters to discard anomalous data. Such anomalous data are frequently manifested by low values of the cross correlation between the in-ear and external microphones recordings or by large phase variations of $H^{(j)}_{LE,a}(\omega)$ or $H^{(j)}_{RE,a}(\omega)$ in a selected frequency band. Indeed, in non-anomalous cases, phase variations tend to remain small (see figure 5b). For the approach presented in this paper, a value of 0.62 for cross correlation and a phase variation limit equal to π in the frequency band ranging from

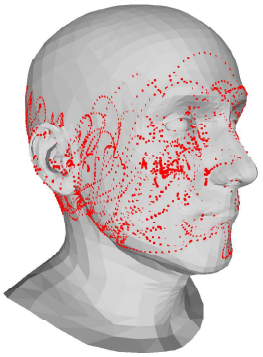
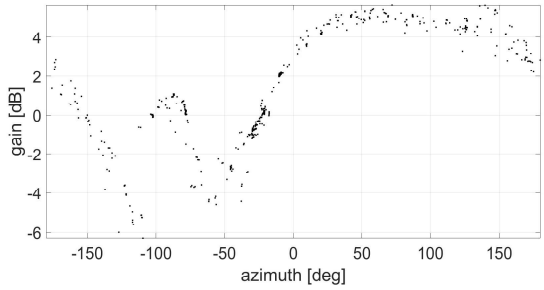
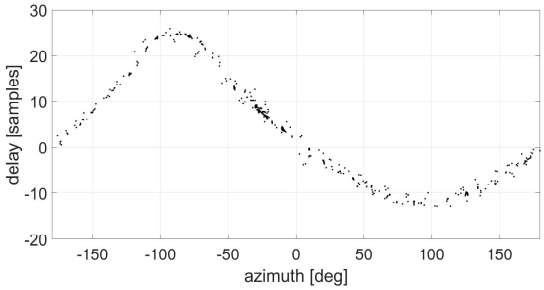


Fig. 6: Impact points of the sound wave on the head at different incidence angles sampled in the laboratory measures.



(a) Scatter plot of g_L



(b) Scatter plot of τ_L

Fig. 7: Scatter plots of the left gain and delay g_L and τ_L as a function of θ in the layer $-1^\circ \leq \phi \leq 1^\circ$

.

0 to $3100Hz$ have proven to be effective in discarding most of the anomalous data. These values have been obtained experimentally and are not the result of an optimisation process. Furthermore, other filtering criteria could be added, for instance by introducing thresholds on too fast translations and/or rotations of the head. The effectiveness of the selected values is shown by the absence of outliers in the quantities τ_L, τ_R, g_L and g_R. Figure 7 shows the scatter plots of g_L and τ_L as a function of θ in the layer $-1^\circ \leq \phi \leq 1^\circ$ (part of the noise in these data is explainable by variations of ϕ). A systematic inspection of pictures like figure 7 gives us confidence in having obtained a sufficiently qualitative dataset for subsequent treatment by the methods described in this paper.

4 HRTF Data Interpolation

The experimental procedure described in section 3 enables us to obtain scattered and noisy HRTF data of real heads. We now show how we take advantage of machine-learning techniques to interpolate this scattered data to obtain direction-dense HRTFs for any required head orientation. These interpolated HRTFs are called the "experimental HRTFs".

With the integration of machine-learning approaches, the domain of data interpolation has witnessed a paradigm shift. However, interpolating data by the use of machine-learning techniques still poses a challenging task for neural networks, particularly when the input data is noisy. In addition, the phenomenon of over-fitting is complex to control. In the context of Multi-Layer Perceptron (MLP)-based interpolation, the intricacies of over-fitting become especially pronounced due to the capability of MLP models to capture complex patterns within data. Rahaman et al. [33] have illuminated a crucial aspect of MLPs, demonstrating that these networks inherently exhibit a bias towards learning low frequency functions, a property that significantly impacts their behavior in data interpolation tasks. This finding is pivotal in understanding the limitations and strengths of MLPs in capturing data dynamics against frequency. To solve this, Mildenhall et al. [21] have proposed leveraging positional encoding to facilitate the learning of high-frequency elements. This innovative approach introduces considerations that are particularly relevant to our work. Specifically, interpolation often demands a careful balance in frequency components, where the presence of high-frequency elements, although indicative of complex patterns, can sometimes result in undesired artefacts or noise in the output. These interpolation techniques are usually employed in a supervised manner, necessitating complete and reliable data for training. In cases involving lab-

oratory data, where completeness is not guaranteed, application of these techniques and objective evaluation of results quality remain an open issue.

In our study, we introduce a neural network architecture designed to process mesh data for sound wave interaction prediction, effectively exploiting the strengths of Convolutional Neural Networks (CNNs) with a Fully Connected Head (FCH) for final output generation. By exploiting CNNs, we take advantage of the inherent superiority of a machine learning-based approach over traditional methods in analyzing and predicting sound wave interactions with meshes. Traditional interpolation methods may lack an understanding of the complex spatial dependencies present in the data, whereas the CNN-based architecture with a Fully Connected Head (FCH) enables a more sophisticated modeling of spatial relationships in mesh data, leading to more accurate and detailed predictions of sound wave interactions. The core of our architecture is built on the ability to handle input data represented as a set of vertices ordered according to the FLAME convention, with dimensions (5023, 3), alongside azimuth and elevation parameters. This specification is tailored to capture the intricate spatial relationships within the mesh, essential for accurate sound interaction modeling. It is integrated with a structured configuration of dense layers, reminiscent of those found in MLPs, enabling the extraction and interpretation of data global structures. The architecture starts with a specialized affine transformation layer, responsible for aligning the sound wave impact point along the X-axis. For each sample, azimuth and elevation are utilized to calculate a transformation matrix that reorients the head mesh so that the impact point always lies along the X-axis. This preprocessing step is crucial for standardizing the input data, ensuring that subsequent layers of the network can effectively learn from a consistent representation of sound wave interactions with the mesh surface. Following this initial transformation, the network employs a series of five convolutional blocks to extract and process features from the transformed mesh data. Each block is composed of two convolutional layers, with the number of filters increasing in depth as the data progresses through the network: the first two blocks apply 32 filters, the following two blocks increase that number to 64, and the final block further expand to 128 filters. This gradation in filtering complexity allows the network to incrementally capture features of increasing specificity and detail, facilitating a comprehensive understanding of the mesh's structural and spatial properties. Each convolutional block is augmented with a MaxPooling layer to reduce dimensionality and enhance the network's ability to focus on salient features, except for the final block, which incorporates a GlobalMaxPooling layer. The transition to GlobalMaxPooling in the last block is a strategic choice, aimed at summarizing the most critical features extracted across the entire mesh. Following the application of GlobalMaxPooling, the feature representation is enriched by concatenating azimuth and elevation data with the output vector. This extended feature vector is then forwarded to a Fully Connected Head (FCH) designed for the final prediction task. This FCH consists in seven hidden layers, each layer playing an important role in modeling the complex relationships present in the data. The final predicted values are g_L, g_R, τ_L and τ_R. The overall neural network architecture is summarized in figure 8.

The model training is conducted utilizing adaptive moment estimation by using the Adam optimization algorithm with Mean Squared Error (MSE) as loss function. The learning rate is decreased when validation loss plateaus are detected, allowing for a more exhaustive exploration of the solution space. The value of the weight assigned to each sample for the loss computation is inversely proportional to the number of neighbors of the sample. This approach is designed to prevent bias caused by non-uniform data distribution, particularly in areas where sample density is low. The weights are calculated using the Nearest Neighbors algorithm, with a fixed radius threshold. The laboratory dataset of a given individual is used to train the algorithm. This dataset comprises 8327 scattered samples spanning various elevations and azimuths, encompassing the range of [-179.69°, 181.58°] for azimuth and [-42.58°, 41.86°] for elevation. An example of the training results on the dataset of figure 6 is depicted in figures 9 and 10.

The HRTFs obtained with the proposed interpolation procedure for experimental HRTFs described above can be compared with theoretical HRTFs obtained using the Boundary Elements Method (BEM). The same head mesh is used to produce comparison data at the same frequencies for a dense grid of source positions.

In our approach, BEM calculations of HRTFs are conducted using the open-source software MESH2HRTF [34], which relies on 3D reconstruction of the head of the listener, as depicted in figure 3b. To ensure convergence of the numerical solver, the mesh undergoes

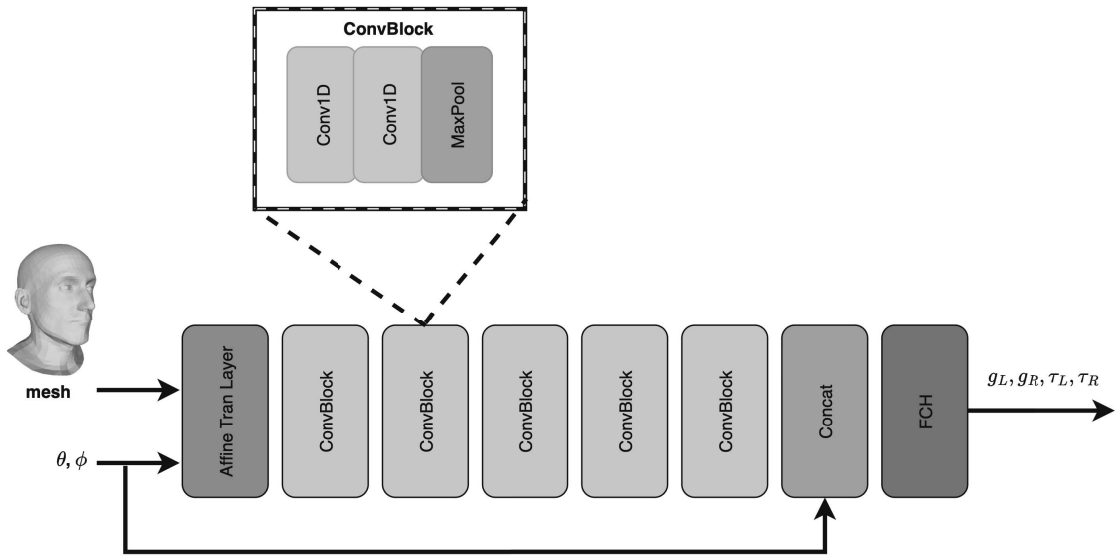


Fig. 8: Our CNN architecture. After an initial Affine Transformation Layer, five Convolutional Blocks are responsible for geometric features extraction. The final FCH (Fully Connected Head) predicts the HRTF values once the geometric features have been concatenated with azimuth and elevation.

re-meshing with a grading algorithm to meet quality requirements. This results in a 3D mesh where one ear is finely detailed, while the other is re-meshed with larger triangles. Hence, two 3D meshes are produced for each ear, and these are processed separately. Additionally, to replicate the position of in-ear microphones in laboratory experiments, the audio source for the calculation is positioned outside the ear (in our experiment, we used a distance of $10mm$). The MESH2HRTF simulation output provides a sound pressure value for each predefined position within an evaluation grid customized to limit calculations to the relevant coordinates. Following these computations, the resulting sound pressures are stored as complex numbers in the standardized audio format known as Spatially Oriented Format for Acoustics (SOFA).

5 Results, Conclusions and Next Steps

Results obtained by interpolating the dataset of the individual of figure 3b are shown in figure 9 for the frequency of $1500Hz$ for three different elevations ($-15°$, $0°$, $15°$), as well as in figure 10 for the frequencies of $2500Hz$, $3500Hz$, and $4500Hz$ for one elevation ($15°$). These particular frequencies and elevations were chosen for the sake of visualization clarity, yet it's important to note that our model possesses the capability to generate a dense grid of results spanning various frequencies and elevation angles. Comparison results obtained with BEM simulation are presented too. It can be noted how the multi-sine and tracking based acquisition procedure (described in section 3) allows for a sparse and noisy experimental data cloud that the machine learning-based interpolation scheme manages to interpolate with high fidelity, in particular for the τ values. Interpolation results of g show some residual noise, mainly caused by a wider dispersion of the data cloud in these dimensions.

Comparison of the experimental results with the theoretical data generated by BEM reveals good compatibility between the results. Values for τ are particularly well aligned, whereas the values of g show some deviations. However, one must consider that the experimental conditions and those used for theoretical simulations are subject to relevant differences, such as the absence of the upper body part in the mesh used for BEM computations and the approximations introduced by the scanning and fitting process. For example, the sequences of deep valleys and peaks in g, produced in BEM simulations as consequences of interference

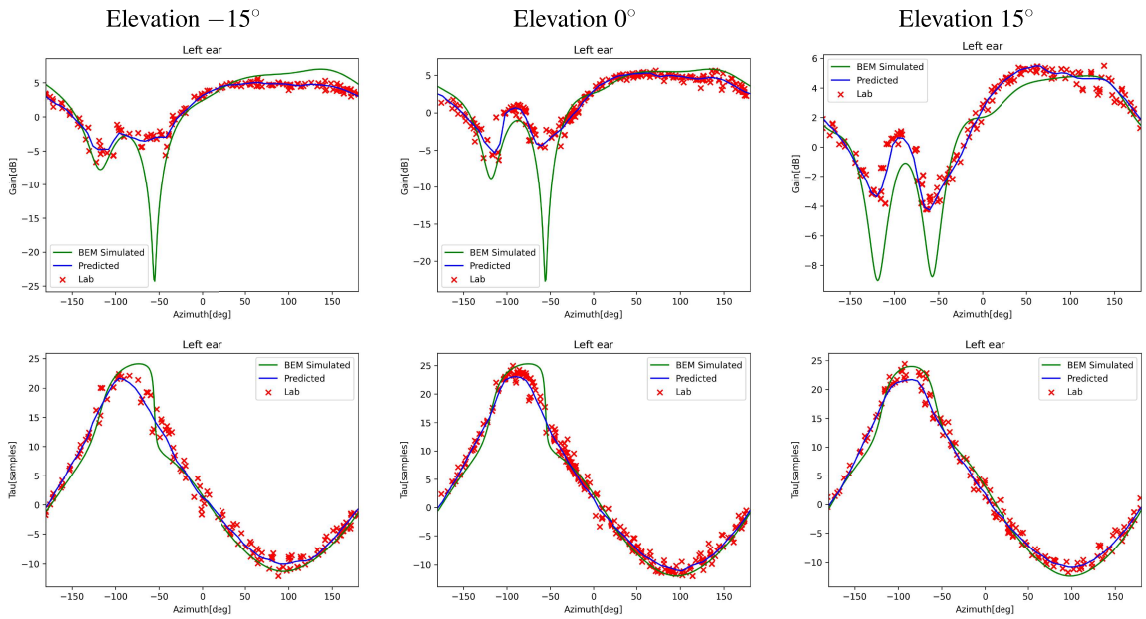


Fig. 9: Interpolation of g and τ for the left ear at $1500Hz$. The laboratory data are represented by red crosses, the interpolation with the CNN is the blue line, while the data simulated with the BEM are in green. For the laboratory data, a range of $\pm 1^\circ$ was considered with respect to the nominal elevation.

around contralateral ears ($\theta \simeq 90^\circ$ for g_R and $\theta \simeq -90^\circ$ for g_L), could be difficult to observe in experimental conditions where the reflections from torso are also present. Moreover, the systematic deviation for positive values of θ could also be explained by a missing destructive contribution from the shoulders.

For the curious reader, the complete experimental dataset and related interpolation figures for different frequencies and for a larger number of elevations are made available under the Creative Commons Attribution 4.0 licence (CC BY) at the following address: `https://doi.org/10.5281/zenodo.11109676`.

The observations provided so far confirm the effectiveness of the proposed combination of multi-sine/MoCap-based data acquisition, FLAME fitting and machine-learning based interpolation scheme for the generation of direction-dense, user-specific HRTFs from a set of dynamic measurements. First, application of the multi-sine technique provides a low-complexity, robust method for fast acquisition of source to ears responses, whilst the MoCap system ensures precise estimation of the user's position and orientation during the measurements. Secondly, the proposed interpolation scheme allows for consideration of features that might not have been captured by classical interpolation schemes such as Lagrange or other polynomial methods, resulting in direction-dense interpolated HRTFs. Finally, the consistency offered by FLAME across scans of heads of different users provides a ultimate advantage when working with laboratory-acquired data as shown in sections 2 and 3 of this paper.

The promising approach presented in this paper motivates further research in the field. Topics of our future research include the refinement of the methodology for the acquisition of denser experimental data clouds, in particular in the ears regions, for a large number of individuals. Regarding machine-learning models, our focus will be on the experimentation with alternative interpolation schemes, as well as on multi-individual, generative AI approaches, by using the geometry of the user's head to condition the HRTF generation strategy. Another important topic of our future research will be the integration of data obtained from BEM simulations with our data-processing pipeline, for example, to favour data set cleaning and augmentation. An additional goal will be the use of advanced machine learn-

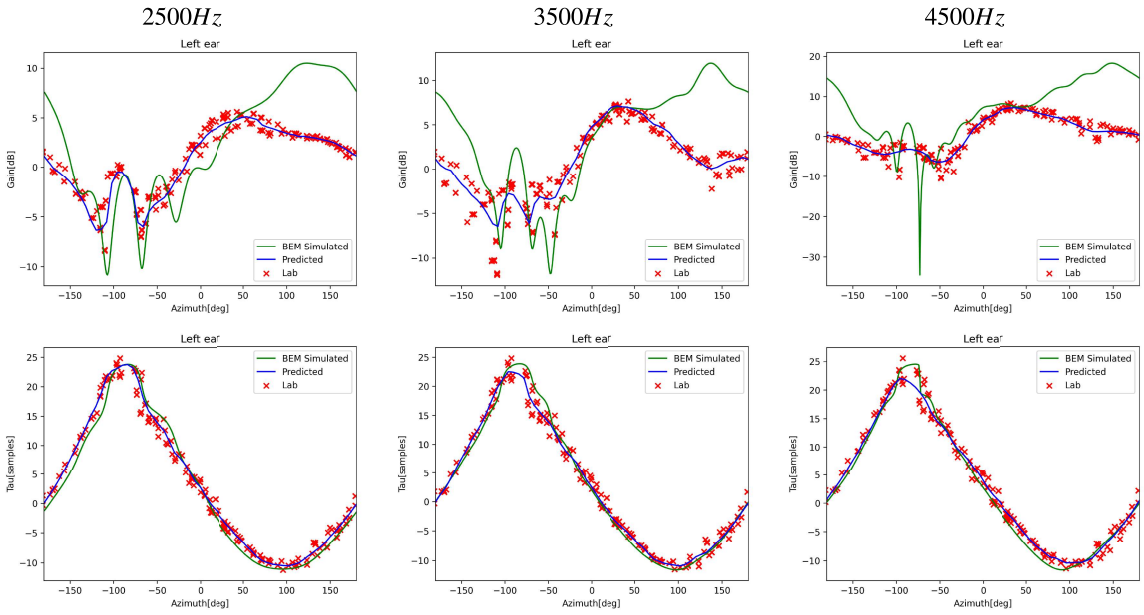


Fig. 10: Interpolation of g and τ for the left ear at $2500Hz$, $3500Hz$ and $4500Hz$, for an elevation of 15°. The laboratory data are represented by red crosses, the interpolation with the CNN is the blue line, while the data simulated with the BEM are in green. For the laboratory data, a range of $\pm 1^\circ$ was considered with respect to the nominal elevation.

ing techniques, such as domain adaptation, to close the gap between simulated and real data. Finally, in depth validation of the effectiveness of the obtained HRTFs in the context of our crosstalk cancellation research activities is part of our current research priorities and will potentially be the topic of future publications.

This work was conducted within the framework of the ASyViR research program funded by Innosuisse, the Swiss funding agencies for innovative technologies, under grant 105.603 IP-ICT.

References

[1] Bregman, A. S., "Auditory Scene Analysis. The Perceptual Organization of Sound," *MIT Press, Cambridge, MA*, 1994, doi:https://doi.org/10.7551/mitpress/1486.001.0001.

[2] Majdak, P., Baumgartner, R., and Jenny, C., "The Technology of Binaural Understanding," *Springer, Berlin, Germany*, 2020, doi:https://doi.org/10.1007/978-3-030-00386-9_5.

[3] Vancheri, A., Leidi, T., Heeb, T., Grossi, L., Spagnoli, N., and Weiss, D., "Geometrical Acoustics Approach to Crosstalk Cancellation," in *Proceedings of the Audio Engineering Society 152nd Convention*, 2022.

[4] Vancheri, A., Leidi, T., Heeb, T., Grossi, L., and Spagnoli, N., "Multiband time-domain crosstalk cancellation," in *Proceedings of the Audio Engineering Society 153rd Convention*, 2022.

[5] Vancheri, A., Leidi, T., Heeb, T., Grossi, L., and Spagnoli, N., "Dynamic Adaptation in Geometrical Acoustic CTC," in *Proceedings of the Audio Engineering Society 154th Convention*, 2023.

[6] Majdak, P., Masiero, B., and Fels, J., "Localization in Individualized and Non-Individualized Crosstalk Cancellation Systems," *J. Acoust. Soc. Am.*, (4), pp. 2055–2068, 2013, doi:https://doi.org/10.1121/1.4792355.

[7] Brinkmann, F., Lindau, A., and Weinzierl, S., "On the Authenticity of Individual Dynamic Bin-

aural Synthesis," *J. Acoust. Soc. Am.*, (4), pp. 1784–1795, 2017, doi:https://doi.org/10.1121/1.5005606.

[8] Guezenoc, C. and Seguier, R., "HRTF Individualization: A Survey," in *Proceedings of the Audio Engineering Society 145th Convention*, 2018.

[9] Pollack, K., Kreuzer, W., and Majdak, P., "Perspective Chapter: Modern Acquisition of Personalised Head-Related Transfer Functions – An Overview," in B. F. Katz and P. Majdak, editors, *Advances in Fundamental and Applied Research on Spatial Audio*, chapter 2, IntechOpen, Rijeka, 2022, doi:10.5772/intechopen.102908.

[10] Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R., "Creating Interactive Virtual Acoustic Environments," *Journal of the Audio Engineering Society*, (47), pp. 675–705, 1999.

[11] Biscainho, L. W. P., Freeland, F. P., and Diniz, P. S. R., "Using inter-positional Transfer Functionsin 3D-sound," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1961–1964, 2002, doi:doi:10.1109/ICASSP.2002.5745014.

[12] Bruschi, V., Dourou, N., Nobili, S., and Cecchi, S., "A New Method for HRTFs Interpolation based on Inverse Distance Weighting," in *Proceedings of the Audio Engineering Society 154th Convention*, 2023.

[13] Hartung, K., Braasch, J., and Sterbing, S. J., "Comparison of Different Methods for the Interpolation of Head-Related Transfer Functions," in *Proceedings of the 16th Audio Engineering Society Conference on Spatial Sound Reproduction*, 1999.

[14] Gamper, H., "Head-related transfer function interpolation in azimuth, elevation, and distance," *Journal of the Acoustical Society of America*, p. 547, 2013, doi:doi:10.1121/1.4828983.

[15] Zurale, D. and Dubnov, S., "Spatial Upsampling of Sparse Head Related Transfer Functions - A VQ-VAE & Transformer based approach," in *AES International Conference on Audio for Virtual and Augmented Reality*, 2022.

[16] Wang, Y., Zhang, Y., Duan, Z., and Bocko, M., "Predicting global head-related transfer functions from scanned head geometry using deep learning and compact representation," in *Audio and Speech Processing*, 2022, doi:doi:10.48550/arXiv.2207.14352.

[17] Zhou, Y., Jiang, H., and Ithapu, V. K., "On the predictability OF HRTFs from ear shapes using deep networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, doi:doi:10.1109/ICASSP39728.2021.9414042.

[18] Fantini, D., Avanzini, F., Ntalampiras, S., and Presti, G., "HRTF Individualization Based on Anthropometric Measurements Extracted from 3D Head Meshes," in *Immersive and 3D Audio: from Architecture to Automotive (I3DA), IEEE*, 2021, doi:doi:10.1109/I3DA48870.2021.9610904.

[19] Zhao, J., Yao, D., Gu, J., and Li, J., "Efficient prediction of individual head-related transfer functions based on 3D meshes," *Applied Acoustics*, 219, p. 109938, 2024, doi:10.1016/j.apacoust.2024.109938.

[20] Ullman, S., "The interpretation of structure from motion," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153), pp. 405–426, 1979.

[21] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R., "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, 65(1), pp. 99–106, 2021.

[22] Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G., "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Transactions on Graphics*, 42(4), 2023.

[23] Quattrini, A., Mascheroni, A., Vandone, A., Coluzzi, M., Barazzetti, A., Cecconi, F., and Leidi, T., "Real Estate Advisory Drone (READ): system for autonomous indoor space appraisals, based on Deep Learning and Visual Inertial Odometry," in *IOP Conference Series: Materials Science and Engineering*, volume 1226, p. 012112, IOP Publishing, 2022.

[24] Li, T., Bolkart, T., Black, M. J., Li, H., and Romero, J., "Learning a model of facial shape and expression from 4D scans," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017.

[25] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J., "SMPL: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866, 2023.

[26] Li, S., Xu, C., and Xie, M., "A robust O (n) solution to the perspective-n-point problem," *IEEE transactions on pattern analysis and machine intelligence*, 34(7), pp. 1444–1450, 2012.

[27] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G., "ORB: An efficient alternative to SIFT or SURF," in *2011 International conference on computer vision*, pp. 2564–2571, Ieee, 2011.

[28] Besl, P. J. and McKay, N. D., "Method for registration of 3-D shapes," in *Sensor fusion IV: control paradigms and data structures*, volume 1611, pp. 586–606, Spie, 1992.

[29] Levenberg, K., "A method for the solution of certain non-linear problems in least squares," *Quarterly of applied mathematics*, 2(2), pp. 164–168, 1944.

[30] Blanz, V. and Vetter, T., "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on pattern analysis and machine intelligence*, 25(9), pp. 1063–1074, 2003.

[31] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T., "A 3D face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pp. 296–301, Ieee, 2009.

[32] Beis, "Instant Audio Frequency Response Measurements," 2015, www.beis.de/Elektronik/FreqResp/InstFreqRespMeasure.html.

[33] Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A., "On the spectral bias of neural networks," in *International Conference on Machine Learning*, pp. 5301–5310, PMLR, 2019.

[34] Ziegelwanger, H., Kreuzer, W., and Majdak, P., "MESH2HRTF: An open-source software package for the numerical calculation of Head-Related Transfer Function," 2015, doi:10.13140/RG.2.1.1707.1128.