DOI: 10.61782/fa.2025.0345

FORUM ACUSTICUM
EURONOISE 2025

ENHANCING PHOTOGRAMMETRY RECONSTRUCTION FOR HRTF SYNTHESIS VIA A GRAPH NEURAL NETWORK

Ludovic Pirard[1*] **Katarina Poole**[1] **Lorenzo Picinali**[1]
[1] Dyson School of Design Engineering, Imperial College London, London, United Kingdom

ABSTRACT

Traditional Head-Related Transfer Functions (HRTFs) acquisition methods rely on specialised equipment and acoustic expertise, posing accessibility challenges. Alternatively, high-resolution 3D modelling offers a pathway to numerically synthesise HRTFs using Boundary Elements Methods and others. However, the high cost and limited availability of advanced 3D scanners restrict their applicability. Photogrammetry has been proposed as a solution for generating 3D head meshes, though its resolution limitations restrict its application for HRTF synthesis. To address these limitations, this study investigates the feasibility of using Graph Neural Networks (GNN) using neural subdivision techniques for upsampling low-resolution Photogrammetry-Reconstructed (PR) meshes into high-resolution meshes, which can then be employed to synthesise individual HRTFs. Photogrammetry data from the SONICOM dataset are processed using Apple Photogrammetry API to reconstruct low-resolution head meshes. The dataset of paired low- and high-resolution meshes is then used to train a GNN to upscale low-resolution inputs to high-resolution outputs, using a Hausdorff Distance-based loss function. The GNN's performance on unseen photogrammetry data is validated geometrically and through synthesised HRTFs generated via Mesh2HRTF. Synthesised HRTFs are evaluated against those computed from high-resolution 3D scans, to acoustically measured HRTFs, and to the KEMAR HRTF using perceptually-relevant numerical analyses as well as behavioural experiments, including localisation and Spatial Release from Masking (SRM) tasks.

Keywords: *HRTF, HRTF synthesis, Photogrammetry Reconstruction, Graph Neural Network, GNN, Mesh2HRTF*

**Corresponding author*: *l.pirard@imperial.ac.uk.*

1. INTRODUCTION

1.1 Related Works

Head-Related Transfer Functions are unique acoustic filters that characterise how sound waves from locations around the listener interact with their anatomy, including the shape, size, and position of the ears, as well as the head and torso [1].

Generic HRTFs can be derived from average human morphology using mannequins such as the KEMAR [2], performing acoustic measurements or numerical synthesis from 3D meshes [3]. While these are useful for universal consumer applications, they often fail to provide a sufficiently accurate spatial audio experience, resulting in front-back confusion and impaired elevation perception [4].

Individual HRTFs are fundamental in immersive audio as they can result in enhanced rendering quality, higher sound localisation accuracy, and potentially also better spatial release from masking performances [5–7].

Traditional individual HRTF acquisition methods require specialised equipment and expertise, which pose significant accessibility challenges [8]. Alternatively, high-resolution 3D modelling provides a pathway to numerically synthesised HRTFs widely available using tools such as Mesh2HRTF [3]. However, the limited availability and high cost of advanced 3D scanners restrict their applicability [9, 10].

Photogrammetry can be seen as an alternative; it is more accessible as well as affordable, and can be per-

formed employing consumer equipment, e.g., a smartphone or a digital camera. Photogrammetry has been explored as a solution to obtain HRTFs via mesh reconstruction and HRTFs synthesis in previous works [9, 11, 12]. The lack of resolution and ear details in the mesh reconstruction has limited the use of photogrammetry for personal HRTFs synthesis [13, 14].

1.2 Research aims

To address these limitations, this study investigates the feasibility of using Graph Neural Networks (GNN) [15] implementing neural subdivision techniques for upsampling low-resolution photogrammetry-reconstructed (PR) meshes into high-resolution ones, which can then be employed to synthesise individual HRTFs. The objective is to improve the resolution and ear morphology details from photogrammetry-reconstructed meshes.

The overall goal is to create an accessible method for any user to obtain a personal HRTF that closely resembles what they could obtain from acoustically measured HRTF, offering more accurate spatial cues compared to a generic HRTF.

2. METHODOLOGY

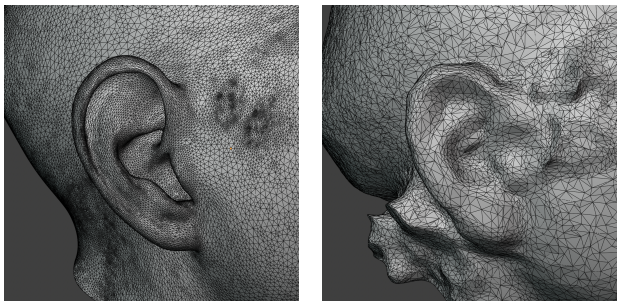
HRTF synthesis is first performed on raw photogrammetry-reconstructed meshes using Mesh2HRTF to establish a baseline and assess the limitations of photogrammetry for individual HRTF computation, without the use of neural networks. Both numerical and perceptual evaluations are conducted. Subsequently, a Graph Neural Network (GNN) is trained and evaluated to generate upsampled meshes. HRTFs are synthesised from these refined meshes and compared against the baseline results obtained from the raw photogrammetry data. Finally, numerical and perceptual evaluations are also carried out to assess the improvements achieved through refinement.

2.1 Photogrammetry Reconstruction

For each subject in the SONICOM dataset [1], 72 images were captured at 5-degree intervals to achieve a full 360-degree representation. The photogrammetry data include high-resolution RGB photographs, depth maps, and gravity data, all acquired using an iPhone XS with a custom 3D-printed mirror bracket to use Apple's TrueDepth technology.

To evaluate the effectiveness of different photogrammetry reconstruction methods, multiple software solutions were tested, including Reality Capture, Agisoft Metashape, Autodesk Recap Photo, and Apple's photogrammetry API. An informal assessment of mesh quality, resolution, and ear morphology details was conducted, revealing that Apple's photogrammetry API produced the most accurate and visually consistent reconstructions. Consequently, this method was implemented using Swift within Xcode for batch processing.

The resulting 244 subject meshes are generated in the .stl format from the SONICOM dataset photogrammetry data. Each subject mesh obtained from photogrammetry has a corresponding high-resolution 3D scan acquired with an EXScan Pro. The main differences between the two meshes are the number of faces and vertices, which impacts the resolution of the meshes and the ear details. The photogrammetry mesh still represents the general shape of the head and ears but lacks ear features that represent individual ear morphology.



(a) 3D scan mesh : right ear

(b) PR mesh using Apple Photogrammetry API : right ear

Figure 1: Same subject right ear meshes with different acquisition methods

2.2 HRTF synthesis

Meshes of varying quality including raw photogrammetry reconstructions, GNN upsampled meshes, and high-resolution 3D scans are processed using Mesh2HRTF [3] to generate simulated HRTFs with chosen locations that correspond to the SONICOM HRTF measurement setup. Prior to simulation, several post-processing steps are applied, including mesh alignment, beheading, clean-up, curvature-adaptive mesh grading, and the assignment of distinct mesh faces for the skin, right ear, and left ear.

Curvature-adaptive mesh grading refines the mesh resolution by increasing the density of elements in high-curvature regions, such as the pinnae, while reducing complexity in flatter areas [16]. This optimisation maintains essential geometric features for accurate acoustic diffraction modelling while minimising computational cost. Due to the high memory requirements of the boundary element method (BEM) used in Mesh2HRTF, simulations are performed on a High-Performance Computing (HPC) system. The HRTF synthesis process follows the guidelines provided in the Mesh2HRTF documentation and tutorials to ensure methodological consistency.

2.3 Numerical and perceptual evaluation

Synthesised HRTFs obtained from photogrammetry-reconstructed and GNN-upsampled meshes are evaluated numerically and perceptually against those computed from high-resolution scans, acoustically measured HRTFs, and the KEMAR HRTF.

The numerical evaluation is conducted using the Log-Spectral Distortion (LSD) metric [17], which quantifies spectral differences between HRTFs in the frequency domain, as well as interaural cue variations, including Interaural Time Differences (ITDs) and Interaural Level Differences (ILDs). Numerical analysis is performed using the Spatial Audio Metrics 0.1.2, K. C. Poole, AXD, Imperial College London `https://github.com/Katarina-Poole/Spatial-Audio-Metrics`.

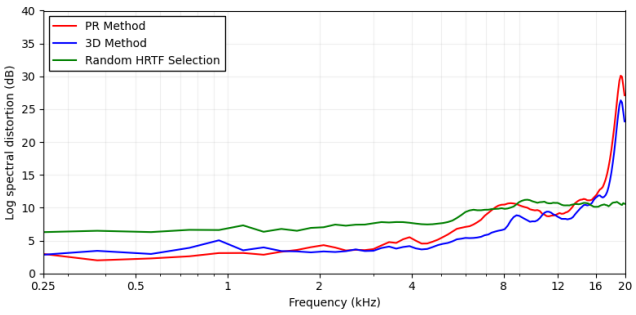


Figure 2: Average LSD comparison across 10 subjects : PR vs 3D vs Random HRTF selection against measured HRTF

The average Log-Spectral Distortion (LSD) is computed across 10 subjects to compare HRTFs synthesised from photogrammetry-Reconstruction (PR) and high-resolution 3D scan meshes [18] against measured HRTFs. For each subject, LSD is calculated as the mean distortion between the synthesised and measured HRTFs for both ears, and then averaged across subjects. Results in Fig. 2 indicate that PR-derived HRTFs exhibit similar trends to those from high-resolution scans but with increased degradation. Both PR and 3D scan-based methods generally produce lower LSD values than a randomly selected non-individual HRTF across most of the frequency spectrum up to 12 kHz. However, significant discrepancies arise above 16 kHz, possibly due to errors in the BEM calculations as these appear for both reconstructions, with the PR one showing slightly larger deviations. As reported in [19], a randomly assigned non-individualised HRTF typically results in an average LSD of 5–10 dB across the spectrum.

Perceptual evaluation of the synthesised HRTFs is conducted through both localisation tasks and Spatial Release from Masking (SRM) experiments [5, 20]. The first assesses the ability of participants to accurately identify the direction of virtual sound sources presented at various azimuth and elevation angles. SRM experiments evaluate the impact of different HRTF sets on speech intelligibility in complex auditory scenes, asking participants to recognise target speech signals in the presence of spatially distributed maskers.

2.4 Graph Neural Network

The neural network developed in this study is based on a Graph Neural Network (GNN) architecture, where 3D meshes are represented as graphs in which vertices and edges encode geometry and connectivity data. Liu et al. introduced a novel framework for data-driven coarse-to-fine geometry modelling [21], taking a coarse triangle mesh as input and recursively subdivides it to a finer geometry. This framework is adapted to our dataset to enhance the resolution of photogrammetry-reconstructed meshes.

A key challenge lies in the lack of a bijective map between the low- and high-resolution meshes, as they originate from different acquisition techniques. Unlike [21], where a direct mapping exists, our approach leverages the method proposed by Schmidt et al. to compute continuous and bijective maps (surface homeomorphisms) between genus-0 triangle meshes [22]. This requires adapting the dataset to meet the method's constraints.

Once the bijective maps computed, the dataset of paired low- and high-resolution meshes trains a Graph Neural Network (GNN) to upscale low-resolution in-

puts to high-resolution outputs, optimised using a specific Hausdorff Distance-based loss function [21]. The model's performance is validated by generating high-resolution meshes from unseen photogrammetry reconstructed meshes and evaluated geometrically with ground truth 3D scan meshes using the Hausdorff Distance.

3. EXPECTED RESULTS AND CHALLENGES

In photogrammetry reconstruction, the quality of the reconstructed mesh depends not only on the quality of the data but also on the algorithm and software used. Employing photogrammetry data taken at a fixed distance from the subject and limited on the horizontal plane, results in the reconstructed meshes to lack ear features and contain an incorrectly reconstructed overhead region.

The numerical evaluation of HRTFs synthesised from raw PR meshes is expected to exhibit similar trends to those computed from high-resolution meshes, although with greater degradation due to geometric inaccuracies. Despite these limitations, perceptual evaluation through localisation tests is expected to demonstrate improved spatial accuracy compared to the KEMAR HRTF and approach the performance of acoustically measured HRTFs.

On the model side, the GNN is expected to learn how to modify a mesh to upscale its resolution and improve ear morphology. However, a potential challenge lies in the risk of the GNN generalising ear morphology, which could lead to inaccurate fitting of personal anatomy. Future work is essential to better understand the perceptual relevance of various components of the pinna, as it remains unclear which parts of the ear are most crucial for accurate sound localisation and HRTF synthesis. By identifying and prioritising these key anatomical features, the GNN's performance could be optimised for personal HRTF synthesis. The intended outcomes for the HRTFs synthesised from GNN-refined meshes are to achieve equivalent numerical and perceptual performance to those computed from high-resolution meshes. The data will be presented at the conference.

4. ACKNOWLEDGMENTS

Horizon-MSCA-2022-DN-01: CherISH is a European Doctorate Network project funded by the European Union's Horizon 2020 framework program for research and innovation under the Marie Sklodowska-Curie Grant Agreement No: 101120054.

5. REFERENCES

[1] I. Engel, R. Daugintis, T. Vicente, A. Hogg, J. Pauwels, A. Tournier, and L. Picinali, "The SONICOM HRTF Dataset," *Journal of the Audio Engineering Society*, vol. 71, pp. 241–253, May 2023.

[2] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *The Journal of the Acoustical Society of America*, vol. 97, pp. 3907–3908, June 1995.

[3] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "MESH2HRTF: AN OPEN-SOURCE SOFTWARE PACKAGE FOR THE NUMERICAL CALCULATION OF HEAD-RELATED TRANFER FUNCTIONS," July 2015.

[4] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A Cross-Evaluated Database of Measured and Simulated HRTFs Including 3D Head Meshes, Anthropometric Features, and Headphone Impulse Responses," *Journal of the Audio Engineering Society*, vol. 67, pp. 705–718, Sept. 2019.

[5] D. González-Toledo, M. Cuevas-Rodríguez, T. Vicente, L. Picinali, L. Molina-Tanco, and A. Reyes-Lecuona, "Spatial release from masking in the median plane with non-native speakers using individual and mannequin head related transfer functions," *The Journal of the Acoustical Society of America*, vol. 155, pp. 284–293, Jan. 2024.

[6] L. Picinali and B. F. G. Katz, "System-to-User and User-to-System Adaptations in Binaural Audio," in *Sonic Interactions in Virtual Environments* (M. Geronazzo and S. Serafin, eds.), pp. 115–143, Cham: Springer International Publishing, 2023.

[7] J. Meyer and L. Picinali, "On the generalization of accommodation to head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 157, pp. 420–432, Jan. 2025.

[8] J. Pauwels and L. Picinali, "On the Relevance of the Differences Between HRTF Measurement Setups for Machine Learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, June 2023. ISSN: 2379-190X.

[9] K. Pollack, W. Kreuzer, and P. Majdak, "Perspective Chapter: Modern Acquisition of Personalised Head-

Related Transfer Functions – An Overview," in *Advances in Fundamental and Applied Research on Spatial Audio* (B. F. G. Katz and P. Majdak, eds.), Rijeka: IntechOpen, 2022. Section: 2.

[10] H. Ziegelwanger, A. Reichinger, and P. Majdak, "Calculation of listener-specific head-related transfer functions: Effect of mesh quality," *Proceedings of Meetings on Acoustics*, vol. 19, p. 050017, May 2013.

[11] M. Dellepiane, N. Pietroni, N. Tsingos, M. Asselot, and R. Scopigno, "Reconstructing head models from photographs for individualized 3D-audio processing," *Comput. Graph. Forum*, vol. 27, pp. 1719–1727, Oct. 2008.

[12] A. Meshram, R. Mehra, H. Yang, E. Dunn, J.-M. Franm, and D. Manocha, "P-HRTF: Efficient personalized HRTF computation for high-fidelity spatial sound," in *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 53–61, Sept. 2014.

[13] K. Pollack, P. Majdak, and H. Furtado, "Application of non-rigid registration to photogrammetrically reconstructed pinna point clouds for the calculation of personalised head-related transfer functions. sl," Hamburg, 2023.

[14] K. Pollack, P. Majdak, and H. Furtado, "Combination of photogrammetry and non-rigid pinna registration for the calculation of personalised head- related transfer functions," in *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, (Turin, Italy), pp. 4147–4150, European Acoustics Association, Jan. 2024.

[15] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or, "MeshCNN: a network with an edge," *ACM Transactions on Graphics*, vol. 38, pp. 1–12, July 2019.

[16] T. Palm, S. Koch, F. Brinkmann, and M. Alexa, "Curvature-adaptive mesh grading for numerical approximation of head-related transfer functions," *Proceedings of the Fortschritte der Akustik (DAGA)*, pp. 1111–1114, 2021.

[17] X. Hu, J. Li, L. Picinali, and A. Hogg, "HRTF Spatial Upsampling in the Spherical Harmonics Domain Employing a Generative Adversarial Network," Sept. 2024.

[18] K. C. Poole, J. Meyer, V. Martin, R. Daugintis, N. Marggraf-Turley, J. Webb, L. Pirard, N. La Magna, and L. Picinali, "The Extended SONICOM HRTF Dataset," 2025. Forum Acusticum, Malaga, Spain.

[19] A. O. T. Hogg, M. Jenkins, H. Liu, I. Squires, S. J. Cooper, and L. Picinali, "HRTF Upsampling With a Generative Adversarial Network Using a Gnomonic Equiangular Projection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2085–2099, 2024. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[20] R. Daugintis, B. Alary, M. Geronazzo, and L. Picinali, "Effects of binaural rendering personalisation and reverberation on speech-on-speech masking," *Journal of the Audio Engineering Society*, Aug. 2024.

[21] H.-T. Liu, V. Kim, S. Chaudhuri, N. Aigerman, and A. Jacobson, "Neural subdivision," *ACM Transactions on Graphics*, vol. 39, July 2020.

[22] P. Schmidt, D. Pieper, and L. Kobbelt, "Surface Maps via Adaptive Triangulations," in *Computer Graphics Forum*, vol. 42, 2023. Issue: 2.