Evaluation of the accuracy of photogrammetry for head-related transfer functions acquisition using numerical methods

Fabio Di Giusto[1,3], Sjoerd van Ophem[1,3], Elke Deckers[2,3], Wim Desmet[1,3]

[1] *KU Leuven, Department of Mechanical Engineering, Celestijnenlaan 300 B, B-3001, Heverlee, Belgium*

[2] *KU Leuven, Campus Diepenbeek, Department of Mechanical Engineering, Wetenschapspark 27, 3590, Diepenbeek, Belgium*

[3] *DMMS lab, Flanders Make, Belgium* *Email: fabio.digiusto@kuleuven.be*

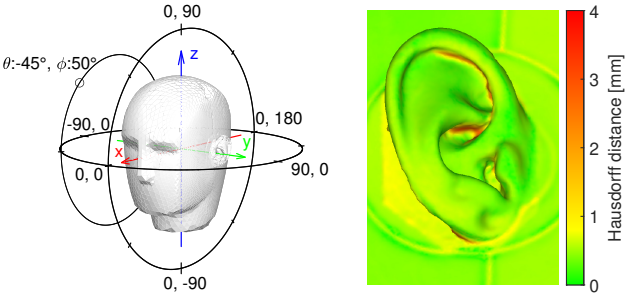
Introduction

The Head Related Transfer Function (HRTF), defined as the transfer function of the acoustical transmission path from a point source to the eardrums, contains the auditory cues used to localise sounds in space. These encompass binaural cues, i.e. Interaural Time and Level Difference (ITD, ILD) dominant in the horizontal plane, and monaural spectral cues mainly used in elevation perception. Headphone presentation of signals filtered with a subject's own HRTF results in a realistic spatial experience [1]. However, the high idiosyncrasy of the HRTF poses a challenge; the use of non-individual filters hinders the spatialisation, making individualisation necessary to obtain the most accurate rendering of a virtual scene [2]. The most conventional HRTF acquisition procedure is an acoustic measurement in anechoic conditions, which requires specialised equipment and is affected by noise and repeatability issues [3]. Alternative approaches employ numerical methods, such as the Boundary Element Method (BEM), to calculate HRTFs on individual geometries. Although not affected by the aforementioned problems, this process requires a precise geometry of the subject which is challenging to obtain given the complexity of the ear shape [1]. Different scanning techniques have been tested (e.g. laser scanning); however, they generally require expensive equipment. Photogrammetry, on the other hand, shows potential in terms of speed and affordability at the cost of a typically lower precision. This technique has been assessed for the reconstruction of ears from a set of pictures, showing a maximum deviation of around 2 mm from a reference geometry [4]. Comparison of HRTF amplitude numerically computed on reference and photogrammetric scans of two head geometries has shown good agreement up to 10 kHz [5]. Nevertheless, numerical results have been seldom compared to matching acoustic HRTFs and the analyses generally focused on objective metrics rather than on perceptual attributes of the transfer functions.

This study aims to assess the accuracy of numerical HRTFs acquired on a Neumann KU100 dummy head. The geometry, scanned with photogrammetry, is compared to a reference obtained through laser scanning. Numerically computed HRTFs are validated against a set of acoustic measurements, taking into account both objective and perceptual metrics. The employed perceptual metric is a localisation model evaluating elevation perception in the mid-sagittal plane [6]. This model has already been used to analyse repeated acoustic dummy head HRTF measurements showing differences in perceptual attributes between them [7]. The novelty of this study is found in using these metrics to assess the perceptual accuracy of numerically computed HRTFs.

Methods

The reference geometry of the dummy head is acquired using a line laser scanner. The head is scanned entirely while the ears are separately scanned to obtain a better accuracy of the pinna shape. The resulting scans are merged and meshed, producing a triangular surface mesh. The photogrammetric mesh is similarly obtained; however, given the uniform colour and texture of the dummy head, which hinders the photogrammetric reconstruction algorithm, chalk is applied in random patterns on its surface to mimic the effect of scanning spray. Videos of the head and ears are taken with a smartphone camera; around 100 frames are extracted from each video and used for the reconstruction, leveraging a photogrammetric algorithm with default settings [8, 9]. Processing is applied to obtain a watertight mesh, which is scaled to the correct dimensions based on measures taken on the dummy head. The meshes are then aligned to the axes and the measurement grid, corresponding to the direction and distance r at which the HRTFs are computed. The head alignment is shown in Figure 1a, along with the utilised interaural-polar coordinate system defined by the azimuth angle $\theta \in [-90°, 90°]$ and the elevation angle $\phi \in [-90°, 270°)$. The Hausdorff distance is employed as a metric to assess the deviation between the photogrammetric and reference geometries [5].



(a) Laser scanned mesh (coarse) and interaural-polar coordinate system. **(b)** Right ear photogrammetric mesh with Hausdorff distance from laser mesh.

Figure 1: Scanned geometry of KU100 dummy head.

Numerical HRTFs are calculated on the meshes using the Fast Multipole BEM implemented in "Mesh2HRTF" [10]. Prior to the calculations, a grading algorithm is applied for speeding up the computation while maintaining

accuracy of the results. One graded mesh per ear is created, with edge length ranging from 1 mm at one ear to 15 mm at the opposite side [11]. HRTFs are evaluated on a grid of 1550 points spanning all θ and a limited set of $\phi \in [-30°, 80°] \cup [100°, 210°]$ at $r = 1.2$ m, in frequency steps of 100 Hz from 100 Hz to 22 kHz. The results are then converted in Head Related Impulse Responses (HRIRs) through inverse Fourier transform, lowpass filtered at 18 kHz and compared to acoustic data from the "Club Fritz" database, containing KU100 HRIRs measured in different laboratories [3]. To avoid interpolation errors a subset of the database is extracted [7], selecting only HRIRs having $\phi \in [-30°, 70°] \cup [110°, 210°]$ with a 10° step, corresponding to the largest common set of ϕ at $\theta = 0°$. The resulting dataset, presented in Table 1, contains 8 acoustic HRTFs identified by their name within the original database (e.g. cf1) and the numerical HRTFs computed on the laser (las) and photogrammetric meshes (pho). It should be mentioned that, due to high noise, cf5 is not selected even though having the desired ϕ resolution, and cf12 is not taken from the database but post-processed from the available original measurement [13].

Table 1: Dataset of HRTFs

ID	Institute	Sources	Elevation range ϕ [°]
cf1	IRCAM	2016	-45:5:225
cf2	UMD	823	-35:5:215
cf3	NASA	432	-40:10:70 ∪ 110:10:210
cf4	IRCAM	1944	-40:10:70 ∪ 110:10:210
cf8	NICT	1008	-40:10:220
cf9	Tekada	2017	-50:5:230
cf10	RIEC	648	-80:10:260
cf12	ARI	1550	-30:10:80 ∪ 100:10:210
las	KUL	1550	-30:10:80 ∪ 100:10:210
pho	KUL	1550	-30:10:80 ∪ 100:10:210

ITDs and ILDs are evaluated on the dataset. ITDs are calculated by leading-edge detection, the code implemented in the Time-Of-Arrival (TOA) model within the Auditory Modelling Toolbox (AMT) is leveraged for a robust estimation [12]. ILDs are evaluated as:

$$\mathrm{ILD}(\theta, \phi, r, f) = 20 \log_{10} \left| \frac{\mathrm{H_R}(\theta, \phi, r, f)}{\mathrm{H_L}(\theta, \phi, r, f)} \right|, \qquad (1)$$

where $\mathrm{H_R}$ and $\mathrm{H_L}$ respectively correspond to right and left ear HRTF, and f is the frequency [1].

The sagittal plane localisation model, also included in the AMT, is used as a perceptual metric. This model takes two Directional Transfer Functions (DTFs) as inputs, obtained from the HRTFs, one defined as the template and corresponding to the internal representation of the virtual subject, and one as the target which is used to filter the sound presented to this subject. The two inputs are processed similarly to the auditory system and compared, predicting the localisation performance in terms of Quadrant Error (QE) and Polar Error (PE), relating to the rate of confusion and the local precision of the subject's response, respectively. Setting template and target to be the same DTF, the baseline condition is obtained, corresponding to a localisation with the subject's own auditory filters. A sensitivity parameter S is used to represent a listener's localisation ability, modelling the non-acoustic factors related to spatial sound perception. In this analysis S is set to 0.21 corresponding to the lowest value observed in the original study and modelling a "good localiser" [6]. HRTFs in the dataset are resampled to 48 kHz, converted to DTFs and fed to the model. Both QE and PE are evaluated at $\theta = 0°$ for a limited set of $\phi \in [-30°, 40°] \cup [140°, 210°]$ to avoid directions above the listener where localisation is generally poor [7]. The errors for each combination of DTFs as template and target are computed, resulting in two all-against-all matrices (QEM, PEM) which rows and columns share a common DTF as template and target, respectively,

$$\mathrm{QEM}_{i,j} = \mathrm{QE}(\mathrm{DTF}_i, \mathrm{DTF}_j), \qquad (2)$$

where i and j respectively correspond to template and target index. The baseline error of each template DTF is subtracted from the error distribution with different targets to normalise small inequalities in baseline performance, obtaining the ΔQEM and ΔPEM matrices,

$$\Delta\mathrm{QEM}_{i,j} = \mathrm{QEM}_{i,j} - \mathrm{QEM}_{i,i}. \qquad (3)$$

Distribution of the normalised error metrics for each template (resp. target) and different target (resp. template) DTFs, i.e. the values of each row (resp. column) of ΔQEM and ΔPEM excluding the diagonal element, are compared with a paired t-test with significance level set to 0.01 in order to assess perceptual similarity. The null-hypothesis is that a pair of template (resp. target) DTFs leads to a similar localisation error distribution when setting all the other DTFs as the target (resp. template); small differences in performance could be attributed to noise inherent to the acquisition process. Two DTFs are defined to be perceptually similar if the null-hypothesis is not rejected, leading to the construction of a diversity matrix highlighting the similarities. The photogrammetric DTF is not included in this analysis since initial tests show that it produces higher localisation error than all the other DTFs, generating non-normal error distributions, thus violating the assumptions made in the t-test. Finally, localisation error distributions obtained setting the laser DTF as template, and switching the targets to be the photogrammetric DTF, individual accurate and inaccurate DTFs from the dataset, and non-individual DTFs arbitrarily selected from the ARI database [14] (from subjects 2, 5, 8 and 10), are compared. This is done to assess how the localisation performance with photogrammetric DTFs compares to localising with individual and non-individual auditory filters.

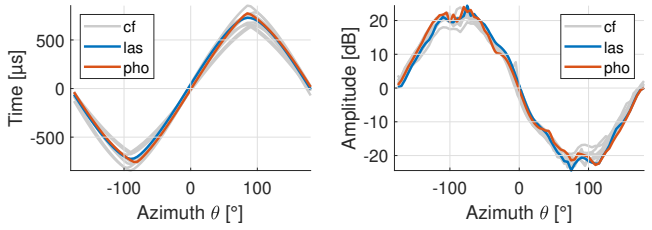
Results

Scanning Accuracy

The Hausdorff distance computed between the photogrammetry and reference meshes indicates a maximum deviation of 4 mm at the head and 2.5 mm at the pinnae. The results at the right ear are visualised in Figure 1b. It is noticeable that inaccuracies are mainly found on the convex ear structures due to their complex geometry; this is found to be in line with previous studies [4, 5].

ITD and ILD Analysis

Broadband horizontal plane ITDs and ILDs, computed with the TOA model and taking the mean of Equation 1 over f, are presented in Figure 2a and 2b, respectively. Numerical HRTF results are found to lie within the variability observed in the acoustic data. A maximum deviation of 240 µs is observed in the ITDs of the dataset, matching previous results on the full acoustic database [3], while the maximum deviation in ILDs is below 8 dB. The results indicate that the applied numerical acquisition procedure can provide interaural cues in line with the ones obtained in repeated acoustic measurements.



(a) Broadband ITD. **(b)** Broadband ILD.

Figure 2: Horizontal plane ITDs and ILDs for acoustic (cf), numerical laser (las) and photogrammetric (pho) HRTFs.

Localisation Performance Analysis

Figure 3 shows the localisation performance using one DTF as template and all others as targets for three representative cases, namely a relatively clean acoustic DTF (cf1), a noisy one (cf9) and the numerical laser DTF (las).

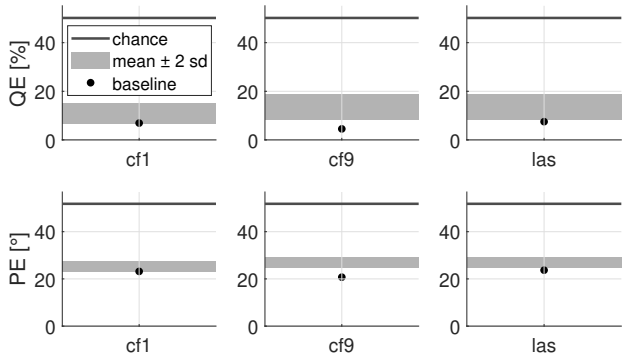


Figure 3: Localisation performance with fixed template and varied target DTFs, results of a clean (cf1), noisy (cf9) and numerical (las) DTF. Black dots show baseline performance, grey areas show distribution of the localisation error (mean ±2 standard deviations) with different target DTFs, the black line shows chance rate (i.e. random guessing).

The first and last show a low baseline error, generally lying within two standard deviations from the mean of the QE and PE distributions with different targets. The noisy DTF also shows a low baseline error, but a higher mean of the error distribution with other targets. This can be due to the noise which modifies the spectral cues, decreasing the similarity between this and the other DTFs. Similar trends can be observed in the error distributions with fixed target and varied template DTF.
The diversity matrix obtained running the paired t-test on the error distributions of ΔQEM and ΔPEM defined

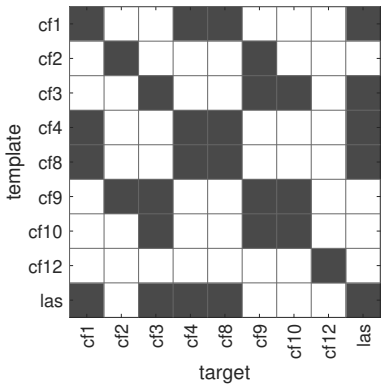


Figure 4: Diversity matrix of acoustic (cf1-12) and numerical laser (las) HRTFs.

in Equation 2 and 3 is shown in Figure 4. It is noticeable that several clusters tend to form. {cf1, cf4, cf8, las} are found to be similar while other groups are composed by {cf3, cf9, cf10}, {cf2, cf9} and {cf12}. DTFs in the first cluster are characterised by low noise and a localisation error with different targets close to the baseline condition. This seems to confirm the perceptual validity of the data in the group containing the laser DTF. The second cluster is affected by higher noise which tends to hinder the perceptual attributes of the DTFs. The clustering might be due to a common noise level in the data, resulting in a similar perceptual influence. Nonetheless, a similarity is also found between {cf3, las}, across the two discussed groups, which seems to indicate that the noise in the second group, even though perceptually relevant, is relatively low. The remaining two groups, showing low similarities, might be affected by a higher noise level capable of disrupting the perceptual characteristics of the data. Although these results are obtained with $S = 0.21$, similar trends with minor differences are observed increasing it up to $S = 1.0$, close to the highest value matching experimental data in the original study [6], modelling a worse localiser than the previous case.

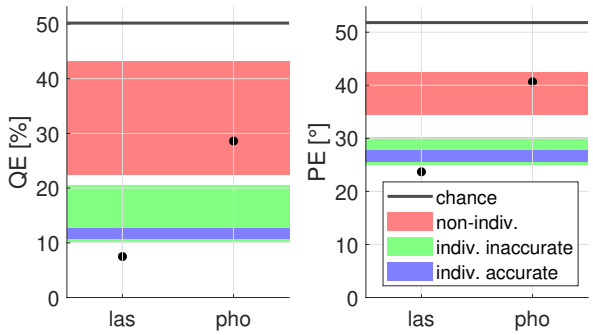


Figure 5: Localisation performance with template fixed to be the laser DTF and varied target set to be the laser (las, baseline) and photogrammetric (pho) DTF, and groups of individual accurate, inaccurate and non-individual DTFs. Areas show distributions of the localisation error (mean ±2 standard deviations), the black line shows chance rate.

The results of comparing the localisation performance with the photogrammetric and various individual and non-individual DTFs while using the laser DTF as ref-

erence are shown in Figure 5. The individual accurate DTFs are defined as the ones found to be similar to the numerical laser results, i.e. {cf1, cf3, cf4, cf8}, while the individual and inaccurate DTFs are the remaining ones, i.e. {cf2, cf9, cf10, cf12}. The localisation error distributions are plotted as areas defined by the mean ±2 standard deviations, laser (baseline) and photogrammetric DTF performance are also shown. The results indicate that the individual accurate DTFs produce slightly higher error than the baseline performance with numerical laser DTFs. Individual inaccurate DTFs are characterised by a higher mean and deviation; worse results are obtained with non-individual DTFs, with a performance relatively close to the chance rate, corresponding to random guessing. The photogrammetric DTF shows a poor localisation performance, having an error comparable to the one obtained when localising with non-individual auditory filters. This might be due to the poor accuracy of the photogrammetric ear geometry, especially at the convex ear structures, and can cause problems in the virtual rendering given that a correct elevation perception seems to relate to realism of the virtual scene [2]. Further analyses are needed to localise which parts of the photogrammetric ear geometry have the strongest influence on the HRTF and tackle the inaccuracies related to the photogrammetric acquisition of individual HRTFs.

Conclusion

In this study the accuracy of photogrammetric HRTFs acquired on a KU100 dummy head is assessed. The geometry of the head is scanned through photogrammetry from images taken with a smartphone camera and compared to a reference laser scan, showing a maximum deviation of 2.5 mm at the ear. Numerical HRTFs computed on these geometries are compared to acoustically measured ones from a database. ITD and ILD analysis shows agreement between the numerical and acoustic results. Perceptual validity is evaluated through a localisation model; a paired t-test is used to assess the similarity between the HRTFs, validating the perceptual accuracy of the laser scanned results. Analysis of the localisation error with individual, non-individual and photogrammetric HRTFs shows that the latter relates to a poor elevation perception, probably due to the low scanning accuracy at several parts of the ear, which is thought to hinder the virtual auditory experience. Improvements in the photogrammetric acquisition process are considered to be necessary to obtain perceptually valid HRTFs.

Acknowledgements

The European Commission is gratefully acknowledged for their support of the VRACE research project (GA 812719). The research of S. van Ophem (fellowship no. 1277021N) is funded by a grant from the Research Foundation – Flanders (FWO).

References

[1] Zhong, X. and Xie, B.,"Head-Related Transfer Functions and Virtual Auditory Displays", Soundscape Semiotics: Localization and Categorization, pp. 99-134, 2014

[2] Jenny, C. and Reuter, C., "Usability of individualized head-related transfer functions in virtual reality: Empirical study with perceptual attributes in sagittal plane sound localization", JMIR Serious Games, 8(3), 2020

[3] Andreopoulou, A., Begault, D. R. and Katz, B. F. G., "Inter-Laboratory Round Robin HRTF Measurement Comparison", Journal on Selected Topics in Signal Processing, 9(5), pp. 895–906, 2015

[4] Reichinger, A., Majdak, P., Sablatnig, R. and Maierhofer, S., "Evaluation of Methods for Optical 3-D Scanning of Human Pinnas", Proceedings of the 2013 International Conference on 3D Vision, pp. 390–397, 2013

[5] Mäkivirta, A., Malinen, M., Johansson, J., Saari, V., Karjalainen, A. and Vosough, P., "Accuracy of photogrammetric extraction of the head and torso shape for personal acoustic HRTF modeling", 148th Audio Engineering Society International Convention, 2020

[6] Baumgartner, R., Majdak, P. and Laback, B., "Modeling sound-source localization in sagittal planes for human listeners", The Journal of the Acoustical Society of America, 136, pp. 791–802, 2014

[7] Barumerli, R., Geronazzo, M. and Avanzini, F., "Round robin comparison of inter-laboratory HRTF Measurements - Assessment with an auditory model for elevation", IEEE 4th VR Workshop on Sonic Interactions for Virtual Environments, pp 1–5, 2018

[8] Schönberger, J. L. and Frahm, J. M., "Structure-from-Motion Revisited", Conference on Computer Vision and Pattern Recognition, pp. 4104–4113, 2016

[9] Schönberger, J. L., Zheng, E., Pollefeys, M. and Frahm, J. M., "Pixelwise View Selection for Unstructured Multi-View Stereo", European Conference on Computer Vision, pp. 1–15, 2016

[10] Ziegelwanger, H., Kreuzer, W., and Majdak, P., "MESH2HRTF: an Open-Source Software Package for the Numerical Calculation of Head-Related Transfer Functions", Proceedings of The 22nd International Congress on Sound and Vibration, pp. 1–8, 2015

[11] Ziegelwanger, H., Kreuzer, W. and Majdak, P., "A priori mesh grading for the numerical calculation of the head-related transfer functions", Applied Acoustics, 114, pp. 99–110, 2016

[12] Ziegelwanger, H. and Majdak, P., "Modeling the broadband time-of-arrival of the head-related transfer functions for binaural audio", 134th Audio Engineering Society Convention, pp. 553–559, 2013

[13] ARI artificial database, last visited 07/21 `http://sofacoustics.org/data/database/ari%20(artificial)/`

[14] Ari database, last visited 07/21 `http://sofacoustics.org/data/database/ari/`