Audio Engineering Society

Convention Express Paper 317

Presented at the AES 158th Convention
2025 May 22–24, Warsaw, Poland

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. 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.

On the effect of photogrammetric reconstruction and pinna deformation methods on individual head-related transfer functions

Katharina Pollack[1,2], Piotr Majdak[1], and Lorenzo Picinali[2]

[1]*Acoustics Research Institute, Austrian Academy of Sciences, Vienna*
[2]*Audio Experience Design, Imperial College, London*

Correspondence should be addressed to Katharina Pollack (katharina.pollack@oeaw.ac.at)

ABSTRACT

Individual head-related transfer functions (HRTFs) are instrumental in rendering plausible spatial audio playback over headphones as well as in understanding auditory perception. Nowadays, the numerical calculation of individual HRTFs is achievable even without high-performance computers. However, an unsolved problem is the acquisition of a mesh of the pinnae with a submillimeter accuracy. One approach to this problem is the photogrammetric reconstruction (PR), which estimates a 3D shape from 2D input, e.g., photos. Albeit easy to use, this approach comes with a trade-off in the resulting mesh quality, which subsequently has a substantial impact on the HRTF's quality. In this study, we investigated the effect of PR on HRTF quality as compared to HRTFs calculated from a reference mesh acquired with a high-quality structured-light scanner. Additionally, we applied two pinna deformation methods, which registered a non-individual high-quality pinna to the individual low-quality PR pinna by means of geometric distances. We investigated the potential of these methods to improve the quality of the PR-based pinna meshes. Our evaluation was done in the geometrical, acoustical, and psychoacoustical domains including a behavioral sound-localization experiment. Our results show that neither PR nor PR-improvement methods were able to provide individual HRTFs of sufficient quality, indicating that without extensive manual pre- or postprocessing, PR is prone to smear relevant individual details in the HRTF-relevant pinna regions.

1 Introduction

Numerically calculated HRTFs are a feasible way for dealing with HRTF personalization if compared with acoustic measurements, but a feasible method on how to obtain a high-quality 3D model of the ears without the use of expensive equipment is still missing. Photogrammetric reconstruction from pictures taken with consumer equipment does not yield high enough

individual details, but may provide enough detail to personalize ground truth meshes such that high-quality individualized meshes and, subsequently, individualized HRTFs can be obtained.

Reichinger et al. [1] showed that the PR of a living ear is of low quality with a smartphone camera and no preprocessing (e.g., applying matte spray to reduce subsurface scattering), and yields better results when performed on silicone replicas of ears. Albeit [2] showed that PR performs better with taking pictures than a video, this approach is not very consumer-friendly.

However, the quality of PR and the computational power, especially in image processing, have improved since. Despite the known existing problems inherent to PR (noise, outliers, interpolation) and because of the known benefit of the application of non-rigid registration methods to PR pinna geometries [3, 4], we investigated the effect on HRTFs (numerical and perceptual) of three different methods for obtaining meshes using low-quality input data (i.e. obtained using consumer equipment) from easy-to-DIY PR. In order to investigate the effect of PR, we compared for each listener the PR-based HRTFs with a reference obtained from a structured-light scan.

In previous works, we could show that both the PPM [3] and the NRR algorithm [4] are able to register their respective template pinna to a GT target with a high geometric quality when the PR was executed with a lot of preprocessing, i.e., targeted lighting, shaving vellus hair, applying matte spray and randomly placed sprinkles [5, 4]. However, this investigation follows the question whether the two deformation approaches would be able to register a target ear from a PR with little pre- and postprocessing, while still maintaining a high quality mesh.

2 Methods

Four kinds of meshes were acquired in this investigation. The ground-truth (GT) meshes were taken from the SONICOM database v1.0 after extensive postprocessing. They were translated to the origin of the coordinate center and aligned to the Frankfurt plane. Additionally, loose geometry and the torso part were removed, artefacts of the cap, hair, or the stitching from the scanner were corrected, and the head geometry was smoothed. To this end, the photogrammetrically reconstructed (PR) meshes were obtained by recording a video around each listener's head along a fixed trajectory specified as follows: The initial camera position was in front (azimuth angle = 0°) of the seated listener, moved towards the listener's right ear along the horizontal plane until at an azimuth angle of 90°. Then, the camera was moved closer to the ear and while keeping the whole pinna in the center of the frame, moved along a small sagittal circle around the ear. We continued moving the camera towards the listener's right ear along the horizontal plane behind the listener. The sagittal circle was repeated around the left ear, and the trajectory continued on the horizontal plane until the video recording was stopped in front of the listener. The video was recorded in a lossy format (MP4) without additional lighting or other preprocessing steps. Then, 100 frames of the video were extracted as photos (in the lossy format JPG) which were used for the PR (ReCap Photo v25.0, Autodesk). In order for the PR meshes to have the correct position, orientation, and size, minimal cleanup and an affine registration were done in Blender v4.1.0. In the affine registration, the global position, orientation and size of the PR were aligned to the ones of the GT. Here, the focus was on aligning the pinna *contours* as closely as possible. After cutting out the pinnae, the NRR algorithm [6] was applied directly to both ears using a non-individual template from the ARI database [7]. The PPM v2.0 [3] [1] was aligned manually such that the Blender plugin PointCloudCompare [2] showed a median pointwise Euclidean distance of around 1 mm. The PPM uses the average mesh from the WiDESPREaD database [8] as template.

For all meshes, free-field compensated HRTFs were calculated using Mesh2HRTF v1.1.1, using the SONICOM grid and a frequency resolution of 129 frequency bins equidistantly distributed between 0 and 24k Hz, such that the sampling frequency in the corresponding HRIR matches 48 kHz.

The PR and its improvements PPM and NRR were evaluated in three domains, i.e., the geometrical, acoustical, and psychoacoustical domain. Although the main interest was the pinna and its monaural spectral features that are directly connected to spatial detail, the influence of the PR on listeners' head measurements subsequently had influence on the spectral features in the calculation of the HRTFs, because of the boundary condition that

[1] https://github.com/Any2HRTF/PPM/releases/tag/bezier-ppm-2.0

[2] https://github.com/Any2HRTF/PPM/tree/PointCloudCompare

the head is assumed to be soundhard. In the *geometrical* domain, we measured the head's width, depth, circumference, and interaural distance in Blender using a Boolean filter of the head and a plane that was placed horizontally just above the listener's ears. Additionally, the Jaccard index

$$J(A,B) = \frac{A \cap B}{A \cup B}$$

was calculated, where A and B represent to point clouds, and the Jaccard index is the ratio between the intersection and union of said point clouds. In order to show the effect of the PR and its improvements, J was calculated between the reference and the other three conditions, i.e., the PR, NRR, and PPM.

In the *acoustical* domain, the log-spectral distortion was calculated in order to show the effect of PR and its improvements using the Spatial audio metrics (SAM) Toolbox [3]. Additionally, differences in interaural time and level differences (ITD and ILD, respectively) were calculated in order to show the effect of PR on the reference.

In the *psychoacoustical* domain, we performed a sound-localization experiment. The experiment setup was designed in such a way that significant differences could be found between individual and dummy-head HRTFs in a pre-test. Previous studies showed no significant differences in localization performance between measured and calculated individual HRTFs [9]. In order to factor out the effect of measurement vs calculation, we use calculated HRTFs for the main experiment. There were four conditions in the experiment: (1) photogrammetry, (2) photogrammetry + NRR and (3) photogrammetry + PPM, compared with a (4) high-quality reference (acquired using a structured-light scanner). 9 listeners (3 female) with no self-reported hearing loss from the SONICOM database participated in the experiment. The room was the same in which the SONICOM dataset was acquired, and the whole experiment was deployed on a Meta Quest 2 and headphone playback over a pair of Sennheiser HD599. The stimulus presented in all conditions were three noise bursts, created as follows: a 300 ms white noise signal was windowed with three consecutive Hann windows, each with a length of 100 ms. The sound-source grid was constructed at 42 positions distributed in four rings around a sphere and played back at a distance of 1.2 m, which matched the distance of the SONICOM grid. The four horizontal rings were at elevations of -30, 0, 30, and 60°, consisting of 12, 16, 12, and 2 equidistantly spaced directions starting at an azimuth of 0°. A familiarization phase introduced the listener to the user interface and how to respond to the stimuli given the Quest 2 controller. In this phase, an HRTF of a Kemar head with small ears was used. Most participants used one controller, only a few were using both. For the first 10 of in total 100 trials, the static stimulus was played during which the participant was not allowed to move their head. After the response was given by the participant pressing a button, a visual feedback was given on the position of the sound source including continuous Gaussian noise allowing for dynamic head movement until the participant pressed a button. After the participant realigned their head to the center of the coordinate system, the next trial started. For the remaining 90 trials, neither the sound source was shown nor dynamic movement was allowed, training the listener on the usage of the controllers and the interaction with the user interface without training them on the localization task. The rest of the static experiment was divided into blocks of 100 trials, after which participants were given the opportunity to take a short break. We assured randomization of the order of the four conditions. The total test duration was 300 trials per condition, i.e., 1200 trials, which corresponded to approximately 1.5-2 hours. From the submitted sound-localization positions, we calculated lateral accuracy and precision, quadrant error rate, and polar accuracy and precision (excluding the quadrant error rate) using `localizationerror` from the AMT v1.5.

[3] https://github.com/Katarina-Poole/Spatial-Audio-Metrics

3 Results and Discussion

Figure 1 shows the four mesh acquisition methods. Despite a lot of individual details over all regions of the mesh seem to be lost between the reference and the PR, there is some level of detail preserved in the pinna region. The NRR approach (Fig. 1 (d)) registered the details in the PR well, but showed little similarity to the reference. The PPM approach (Fig. 1 (c)) seems to provide details that seem more similar to the reference, which may be due to the parameter space it spans.

Head measurements from all meshes were taken using Blender v4.1.1. They do not replicate an in-person head measurement, but count the cumulative edge length of the mesh on a ring that sits in the horizontal plane just

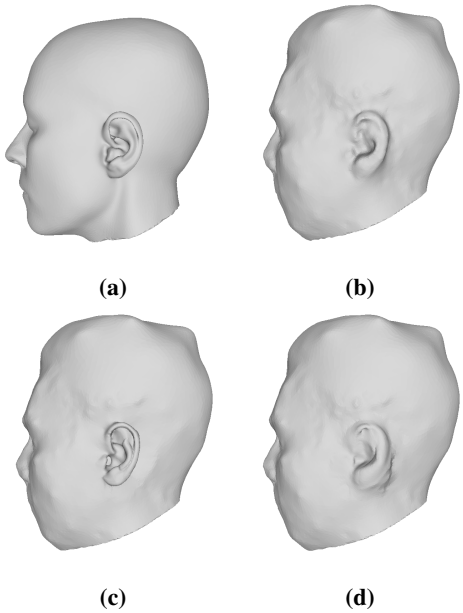


Fig. 1: Meshes of an example listener: **(a)** Ground truth scan, **(b)** PR, **(c)** PPM registered to PR, and **(d)** NRR registered to PR.

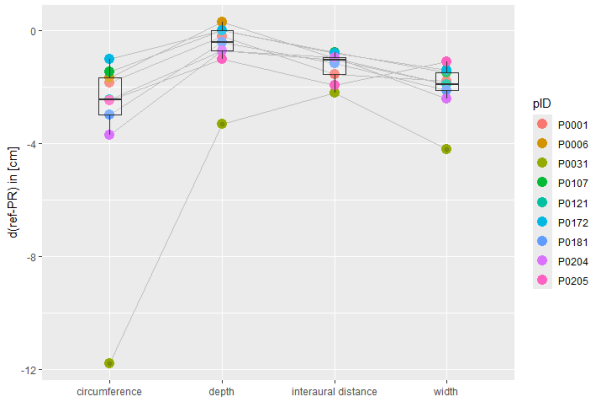


Fig. 2: Difference in various head measurements of participants, reference scan versus photogrammetric reconstruction. Lines indicate same listener for better readability, there is no mathematical or conceptual connection between data points.

on the upper edge of the pinnae (where, e.g., glasses would be placed). Figure 2 shows the difference of head measurements between the reference and the PR $d(ref - PR)$. If the difference is positive, the PR yielded a smaller value than the reference. Overall, the majority of values is negative, suggesting that the PR typically yielded bigger heads than the reference (despite the affine alignment focusing on the pinna contours!). This is not surprising, as the listeners had various hair styles which, typically for PR, get captured as part of the geometry. P0031 had prominent hair that was not preprocessed by, e.g., using a hair band or postprocessed by, e.g., cutting it out in Blender, and the head measurements were far off, likely impacting acoustical and psychoacoustical results.

Next, we looked at local similarities of the resulting pinna meshes. Figure 3 shows the Js between the left ears acquired by NRR and PPM compared to their respective reference point clouds and PRs, and the Js between the PR and the reference point clouds. After testing for normality and homoscedasticity, a one-way ANOVA showed no significant differences between the three acquisition approaches ($p > 0.05$). and a Kruskal-Wallis rank sum test showed significant differences between NRR vs PR and PPM vs PR, NRR vs PR and PR vs reference, and PPM vs PR and PR vs reference ($p < 0.05$).

The comparison to the reference shows that neither NRR nor the PPM yielded significant improvement in the Jaccard index over the PR. When comparing these meshes to the PR, however, both approaches show a significantly higher Jaccard index than the reference. The high Jaccard values for the comparison between NRR and PR and its significantly different Jaccard values indicate that NRR is able to register the PR closer than the PPM. This suggests similar acoustic features in the HRTFs of PR and NRR, but leaves questions to what the PPM-based HRTF will look like as there is no significant improvement between PPM and PR in Fig. 3.

Unfortunately, both approaches reach their limits when using only the information from a PR with this (little) level of detail. Although the PPM was seemingly able to model the ear in marginally higher detail than NRR (see Fig. 1 (c) vs (d)), its parameters of the concave structures were set to their upper limit values in order to minimize the distance to the PR, which likely does not match the reference shape in these regions.

Figure 4 shows the calculated HRTFs from the meshes in Fig. 1 for the same listener in the Median plane. A similar loss in detail can be observed as in the J results,

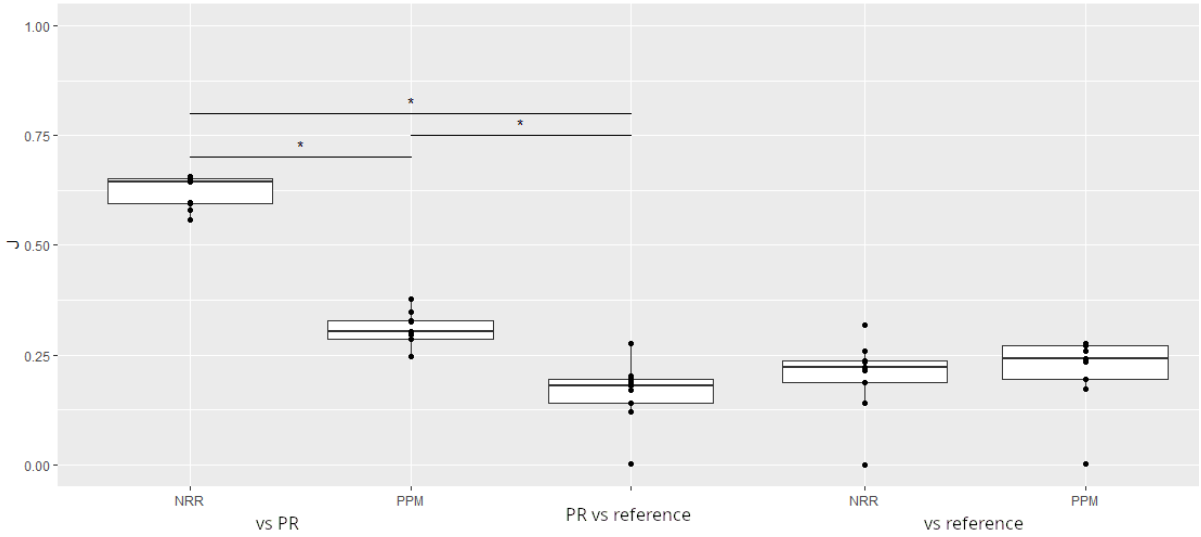


Fig. 3: Similarity of meshes (in term of the Jaccard index, J): From left to right: NRR and PPM as compared to the PR; as well as PR, NRR, and PPM as compared to the reference mesh.

i.e., that the PR HRTF contains a lot less detail than the reference, PPM contains some spectral detail closer to the reference than the PR, and NRR contains similar details as the PR.

Figure 5 shows the log-spectral distortion (LSD) averaged over location and frequency, and comparing the reference to a Kemar, the reference to the PR, and the PR-improvements both to the reference and to the PR. After checking for normality, a Kruskal-Wallis rank sum test yielded significant differences between all but three conditions. These three conditions were the reference compared to the PR, NRR, and PPM. The fact that there was no significant differences found in the LSD between the PR-based meshes and the reference suggests that the PR and its improvements will yield similar performance in a localization experiment.

The LSD as a scalar gives an indication of the magnitude of change in localization errors (Middlebrooks 1999, virtual localization). E.g., for a LSD of approximately 5-6 dB, an increase of up to 10 percentage points can be expected for the quadrant error rate.

Differences in ITDs between the reference and the PR across participants were found at $1.96 \pm 11.18\ \mu s$, which corresponds to approximately 1 sample. Differences in ILDs between the reference and the PR across participants were found at -0.06 ± 1.86 dB. The results in ITD and ILD difference suggest that the differences between the reference and the PR are negligible.

Figure 6 shows lateral accuracy and precision errors. A one-way ANOVA found no significant differences between groups, which was to be expected. The quite high values of lateral precision may be due to the fact that listeners were not trained on the localization task. No significant differences between groups in the lateral domain indicates that the head shape difference had no significant influence on the lateral accuracy or precision.

Figure 7 shows the quadrant error rate (QER). An ANOVA showed significant effect of the condition ($p>0.05$) and the multiple comparison revealed significant differences between PR and the reference, and NRR and the reference, but no significant difference between the PPM and the reference (even when excluding P0181, an outlier in the PPM condition). This indicates that, by means of QER, localization performance using the PPM HRTF was not significantly worse than using a reference HRTF (even when excluding P0181). The chance rate of the experiment was 64 %, which makes P0031 above chance in the NRR and PR condition, and close to chance in the PPM condition.

Figure 8 shows the polar accuracy and precision errors, both excluding the QER [7]. An ANOVA found no

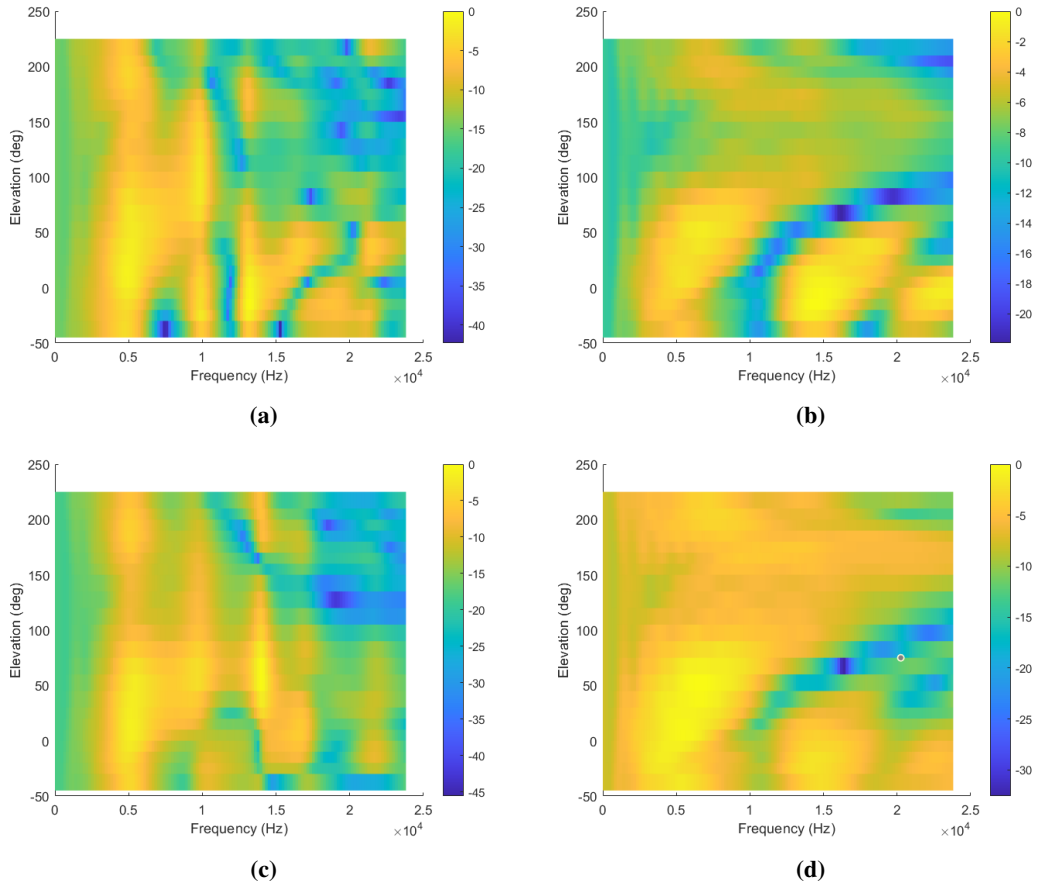


Fig. 4: HRTFs in the median plane of the same listener as shown in Fig. 1 calculated for the following meshes: **(a)** Ground truth scan, **(b)** PR, **(c)** PPM registered to PR, and **(d)** NRR registered to PR.

significant differences between groups for polar accuracy (p=0.68). A Kruskal-Wallis rank sum test revealed no significant differences between groups (p = 0.1) for polar precision. Despite the geometrical and acoustical differences between conditions, there were no significant differences in polar precision, which indicates that there is no effect of PR-based HRTF and the reference HRTF to this localization error metrics. However, there might be a ceiling effect present. One indicator might be that the polar precision levels are close to chance, and another indicator might be that we did not train the listeners on the localization task. The chance rates were 14° and 36°, respectively, thus, our results are potentially affected by ceiling effects.

4 Conclusion

In this article, we investigated the quality of state-of-the-art easy-to-use photogrammetric reconstruction (PR) of pinnae. Additionally, we investigated two improvement methods – a non-rigid registration (NRR) algorithm and a parametric pinna model (PPM) – that have been proven capable of registering a template to a target pinna. We evaluated the quality of the PR and its improvement methods in the geometrical, acoustical, and psychoacoustical domains. In the geometrical and acoustical domain, we showed that NRR was able to register the PR closer than the PPM. However, a lot of spectral detail was missing in the PR, and NRR was not able to maintain the rough shape of the template ear. The PPM showed more spectral detail in the median plane than the PR, which may be due to its little

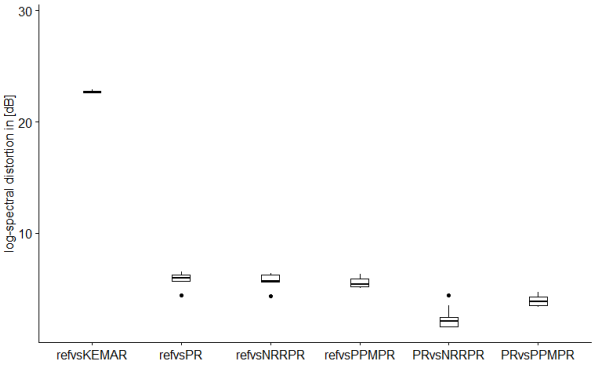


Fig. 5: Log-spectral distortion (LSD) between various DTFs. From left to right: Reference vs Kemar, reference vs PR, reference vs NRR registered to PR, reference vs PPM registered to PR, PR vs NRR registered to PR, as well as PR vs PPM registered to PR.

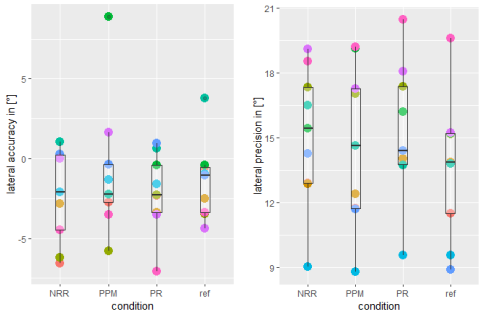


Fig. 6: Lateral (a) accuracy and (b) precision error obtained from the behavioral experiment. Chance rates are 0° and 54°, respectively.

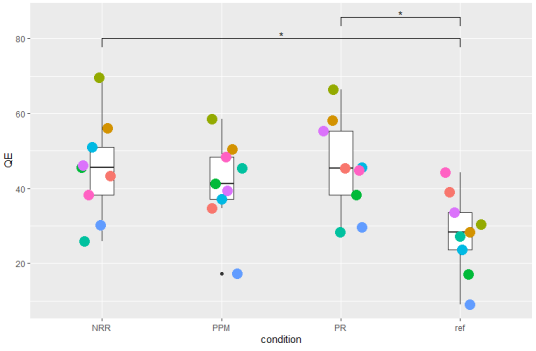


Fig. 7: Quadrant error rate (QER) obtained from the behavioral experiment. Chance rate was 64 %.

geometric similarity to the PR. However, using a PPM-HRTF did not show significantly better localization performance results than using the reference HRTF in the sound-localization experiment, which could have several reasons.

In the sound-localization experiment, the familiarization phase might have confused (naive) listeners as they were exposed to a dummy-head HRTF, and had feedback on their performance for the first 10 trials only. This way, the listeners were trained on the usage of the user interface, not on the localization task. The missing training on localization performance might result in a ceiling effect which can be observed for the polar accuracy and precision errors, as listeners performed close to chance level.

The results across domains suggest that the PR has a significant influence on the individual geometry and, subsequently, the individual HRTFs. Without extensive pre- or postprocessing or without the use of elaborate equipment, even our PR improvements were not able to significantly improve the result in terms of a better sound-localization performance. As it seems, the PR used in our study lacks of capturing geometric details relevant for high-quality individual HRTFs.

5 Acknowledgements

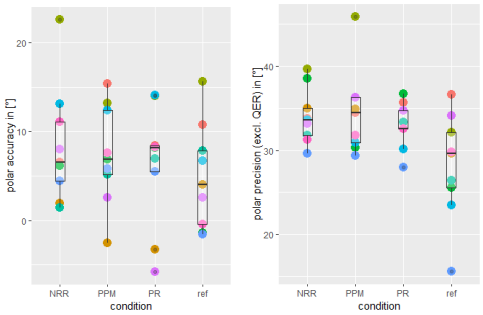
This work was supported by the European Union's Horizon 2020 research and innovation funding programme (grant no. 101017743, project "SONICOM").

References

[1] Reichinger, A., Majdak, P., Sablatnig, R., and Maierhofer, S., "Evaluation of Methods for Optical 3-D Scanning of Human Pinnas," in *2013 International Conference on 3D Vision - 3DV 2013*, pp. 390–397, 2013, doi:10.1109/3DV.2013.58.

[2] Pizzo, M., Viola, E., Solari, F., and Chessa, M., "Evaluation of 3D reconstruction techniques for the blending of real and virtual environments," in *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 360–367, IEEE, 2024.

[3] Pollack, K., Pausch, F., and Majdak, P., "Parametric pinna model for a realistic representation of



(a) **(b)**

Fig. 8: Polar accuracy (a) and precision (b) error obtained from the behavioral experiment. Chance rates are 14° and 36°, respectively.

listener-specific pinna geometry," in *24th International Congress on Acoustics: ICA*, 2022.

[4] Pollack, K., Majdak, P., and Furtado, H., "Combination of photogrammetry and non-rigid pinna registration for the calculation of personalised head-related transfer functions," in *10th Convention of the European Acoustics Association*, European Acoustics Association, 2023.

[5] Pollack, K., Brinkmann, F., Majdak, P., and Kreuzer, W., "Von Fotos zu personalisierter räumlicher Audiowiedergabe," *e & i Elektrotechnik und Informationstechnik*, 138(3), pp. 250–255, 2021.

[6] Hirose, O., "A Bayesian formulation of coherent point drift," *IEEE transactions on pattern analysis and machine intelligence*, 43(7), pp. 2269–2286, 2020.

[7] Majdak, P., Goupell, M. J., and Laback, B., "3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Attention, perception, & psychophysics*, 72(2), pp. 454–469, 2010.

[8] Guezenoc, C. and Seguier, R., "A wide dataset of ear shapes and pinna-related transfer functions generated by random ear drawings," *The Journal of the Acoustical Society of America*, 147(6), pp. 4087–4096, 2020, doi:10.1121/10.0001461.

[9] Ziegelwanger, H., Majdak, P., and Kreuzer, W., "Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization," *The Journal of the Acoustical Society of America*, 138(1), pp. 208–222, 2015.