PERCEPTUALLY MOTIVATED ANALYSIS OF NUMERICALLY SIMULATED HEAD-RELATED TRANSFER FUNCTIONS GENERATED BY VARIOUS 3D SURFACE SCANNING SYSTEMS

Manoj Dinakaran[⋆†], *Fabian Brinkmann*[†], *Stine Harder*[‡], *Robert Pelzer*[†], *Peter Grosche*[⋆], *Rasmus R. Paulsen*[‡] *and Stefan Weinzierl*[†]

[⋆]Huawei Technologies, German Research Center, Riesstraße 25, Munich, Germany.
[†]Technical University Berlin, Audio Communication Group, Einsteinufer 17c, Berlin, Germany.
[‡]Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.

ABSTRACT

Numerical simulations offer a feasible alternative to the direct acoustic measurement of individual head-related transfer functions (HRTFs). For the acquisition of high quality 3D surface scans, as required for these simulations, several approaches exist. In this paper, we systematically analyze the variations between different approaches and evaluate the influence of the accuracy of 3D scans on the resulting simulated HRTFs. To assess this effect, HRTFs were numerically simulated based on 3D scans of the head and pinna of the FABIAN dummy head generated with 6 different methods. These HRTFs were analyzed in terms of interaural time difference, interaural level difference, energetic error in auditory filters and by their modeled localization performance. From the results, it is found that a geometric precision of about 1 mm is needed to maintain accurate localization cues, while a precision of about 4 mm is sufficient to maintain the overall spectral shape.

Index Terms— HRTF, Numerical simulation, Geometric, Localization

1. INTRODUCTION

Individualizing head-related transfer functions (HRTFs) provides an approach to improve the quality of binaural synthesis, e.g. by maintaining the accuracy of localization comparable to the corresponding real sound fields [1]. The most precise approach to obtain individual HRTFs is a direct acoustic measurement. This requires a special setup in the anechoic chamber [2] making it impractical for consumers to use individual HRTFs in binaural applications. Numerical simulations by means of the boundary element method (BEM) allow to simulate HRTFs over the entire audible frequency range [3] on the basis of high quality 3D surface scans of the head, pinna, and torso. In recent years, several approaches have been proposed with a focus on increasing the accuracy of the simulated HRTFs [4, 5, 6] or speeding-up the simulation process [7]. Different techniques for the acquisition of 3D surface scans exist such as MRI scanners, structured light scanners, laser scanners, infrared scanners, stationary scanners, hand held scanners, or by using mobile camera pictures [6]. Each of them provides a different resolution and accuracy which directly affects the quality of the 3D surface scans. The influence of the scanning accuracy on the numerically simulated HRTFs are subject to research.

In this paper we systematically analyze the accuracy of 3D surface scans obtained by different approaches and study their influence on the resulting HRTFs. First, we acquired 3D surface scans of the head and pinna of the FABIAN dummy head by using 6 different methods. An artificial head was chosen because its position remains constant and allows for the evaluation of time consuming scanning methods. Then, the influence of the different scanning methods on the numerically simulated HRTFs was analyzed by means of interaural time difference (ITD), interaural level difference (ILD), energetic error in equivalent rectangular bandwidth (ERB) auditory filters, and their simulated localization performance. To isolate the influence of the scanning method on the HRTF, the different scanning methods were evaluated against a high resolution structured light scan (ground truth) instead of measured HRTFs. However, HRTFs generated with the ground truth showed a very good agreement to it's acoustically measured correspondent in an earlier study [8].

The paper is organized as follows: In section 2, the different scanning methods are explained. In section 3, the alignment and numerical simulation process is detailed, followed by the results and analysis in section 4.

2. ACQUISITION OF 3D SURFACE SCANS

The **GOM ATOS-I** *(GOM-Ref)* is a stationary, high-resolution structured light scanner with a resolution of 0.01 mm from a working distance of 0.45 m to 1.2 m. Multiple scans of FABIAN's head and pinna were captured and aligned using the *ATOS professional* software. More details can be found in Brinkmann *et al.* [8]. The final scans are shown in Fig. 1.

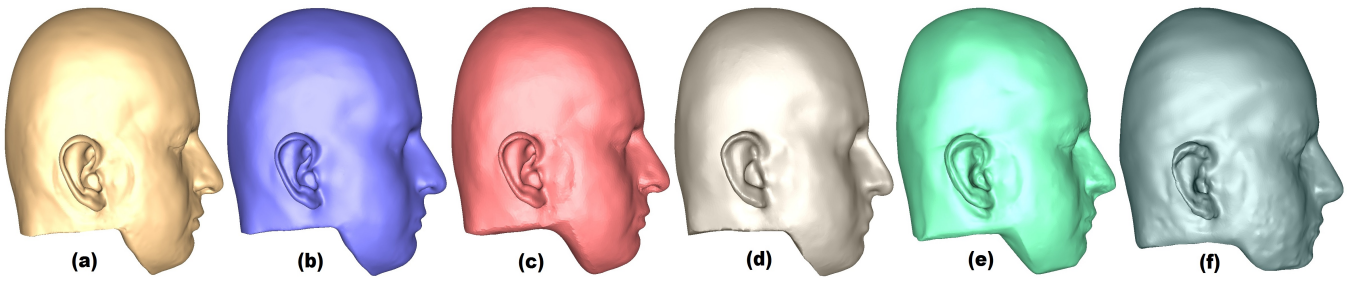

Fig. 1: FABIAN 3D Surface Scans using (a) GOM ATOS-I Scanner *(GOM-Ref)*, (b) Artec Space Spider Scanner *(SPY)*, (c) Canfield Vectra M3 scanner *(CAN)*, (d) Kinect scanner *(KIN)*, (e) Autodesk 123D *(123D)*, and (f) PPT *(PPT)*.

The **Artec Spider** *(SPY)* is a hand-held structured light scanner with 0.05 mm point spacing resolution from a working distance of 0.2 m to 0.3 m [6]. Due to the limited working distance, only FABIAN's pinna was scanned and stitched to a head mesh from the Kinect scanner (see below). For this purpose, 10 to 20 scans from different parts of each pinna were taken from different angles. The separate scans were aligned and stitched, and holes are filled using the *Artec Studio Professional 12* software.

The stationary **Canfield Vectra M3** *(CAN)* scanning system has an accuracy in the range of 0.1 mm at a working distance of 1 m. Sixteen scans are captured from multiple directions to capture all the intricate details of the head and ears. These surfaces are then aligned and stitched together using a custom software package [9] that also fills holes in an anatomically plausible way, resulting in a complete scan. Further details can be found in [10].

The mobile **Microsoft Kinect** *(KIN)* scanner uses an infrared projector and depth sensor to capture the environment as 3D points. The Kinect was set up at ear level height and with a distance of 1 meter from FABIAN, which was placed on a small turnable table. The scan was generated by rotating the table around 360° and acquired using Kinect fusion with the *developer toolkit browser v1.8.0* [11]. Further details of the post processing are presented in Dinakaran *et al.* [12].

Autodesk 123D catch *(123D)* is a free mobile application which allows the user to derive a 3D model from at least five overlapping photos [13]. Here, 16 pictures of FABIAN were taken around 360° using an iPhone 6. In addition, 6 to 10 pictures of each ear were taken from different angles to cover it's details (front, back, above, below and side views). All pictures were uploaded to the Autodesk server via the app's user interface. The processed 3D model *123D* was then obtained from the server. Because 123D catch does not capture the actual size of the scanned objects, the mesh was scaled to best match the size of the reference scan *(GOM-Ref)*.

The **Python Photogrammetry Toolbox** *(PPT)* is an open source tool which has a pipeline to construct a 3D model from a set of pictures. The toolbox uses bundler and patch based multi view stereo software (MVS) from Furukawa *et al.* [14] which performs feature extraction from the photos and generates the point cloud by dense image matching. Here, a set of raw photographs surrounding FABIAN with an approximate resolution of 10° using an iPhone 6 mobile camera were used as the input data, again combined with 6 to 10 images of each pinna. The resulting surface mesh *PPT* [15] was again scaled to best match the size of the reference scan *(GOM-Ref)*.

3. NUMERICAL HRTF SIMULATION

3.1. Alignment and Re-Meshing

In order to compare and analyze the 3D surface scans from different acquisition methods, they were aligned to each other by the following procedure: In the first step, *GOM-Ref* was rotated and translated until the interaural axis (axis through the center of the ear channel entrances) coincided with the y-axis of the coordinate system. In a second step, the mesh was moved until the interaural center fell into the origin of coordinates. Lastly, the upright position was established by a rotation about the y-axis. The remaining FABIAN surface scans were then aligned with respect to *GOM-Ref* using the iterative closest point (ICP) algorithm [16] from the *surface manipulation and transformation toolkit* (SUMATRA) [17].

After the alignment, the meshes were regularized and the number of elements was reduced, which can considerably speed up the processing time of the numerical simulation. For this purpose, an efficient a priori mesh grading algorithm (resulting in non-uniform meshing) was deployed according to Ziegelwanger *et al.* [18]. Because the element size in the graded meshes increases with the distance from the ear, two different models were generated for each scanning method: One for the left pinna (with small mesh elements at the left ear, and large elements at the right), and one for the right pinna. The target lengths used were 1 mm to 10 mm, which resulted in around 20,000 elements per mesh. These Settings showed good results compared to non-regulated meshes [18].

3.2. HRTF simulation

For numerical HRTF simulation, the *Mesh2HRTF* implementation of the 3-dimensional Burton-Miller collocation BEM

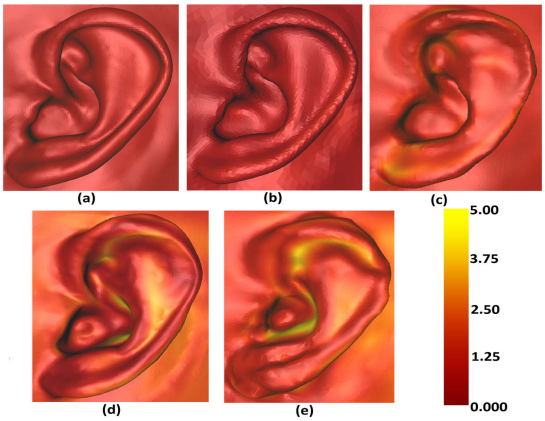

Fig. 2: Geometric difference of (a) *SPY*, (b) *CAN*, (c) *KIN*, (d) *123D* and (e) *PPT* with respect to *GOM-Ref* (in mm).

was used. *Mesh2HRTF* reads geometrical data, calculates the corresponding sound field by numerically solving the wave equation and outputs complex HRTF spectra at discrete frequencies; in our case between 100 Hz and 22 kHz in steps of 100 Hz. Details on *Mesh2HRTF*, including a description and an evaluation of the algorithm, can be found in [19]. The *Mesh2HRTF* input files were created in *Blender* [20]. Assuming reciprocity, the sound field on a desired spatial sampling grid was calculated by assigning a volume velocity to a single element in the mesh located at the entrance to the blocked ear channel [21]. The reciprocal approach is usually chosen to save calculation time and resources by interchanging the positions of loudspeakers and microphones. Hence, the calculation of one active and vibrating element results in the sound pressure information for all nodes of the spatial sampling grid [7]. Afterwards, the complex HRTF spectra were normalized with respect to the surface area of the sound emitting mesh element at the blocked ear channel, it's volume velocity and referenced to a point source in the origin of coordinates by spectral division. This was done with extensions to *Mesh2HRTF* that can be obtained upon request. Finally the 0 Hz bin was set to 1 (0 dB), the single sided spectra were mirrored using the complex conjugate and HRIRs were obtained by inverse Fourier transform. Initially, HRTFs were simulated on a Lebedev grid of degree 1730 at a radius of 1.5 m as implemented in the *SOFiA-Toolbox* [22]. In a second step, the complex spectra were subjected to a spherical harmonics transform of order 35 using *AKtools* [23]. Calculating one HRTF set (left and right ear) took approx. 13 hours using 4 cores of an Intel i7 4 GHz CPU and 32 GB RAM.

4. ANALYSIS

4.1. Geometric differences

Geometric differences with respect to the reference were calculated directly after ICP alignment (before re-meshing) with

Scans	X_1 (μ, σ)	X_2 (μ, σ)	X_3 (*max*)	X_4 (*max*)	X_5 (*max*)	X_6 (*max*)
SPY	0.14 (0.24)	0.17 (0.29)	0.50	0.75	0.75	0.75
CAN	0.66 (0.80)	0.66 (0.54)	0.66	0.90	0.90	0.80
KIN	1.53 (1.28)	1.50 (1.08)	1.50	2.50	1.25	1.25
123D	1.98 (1.42)	2.10 (1.41)	5.00	3.75	2.50	3.75
PPT	1.77 (1.72)	1.68 (1.56)	5.00	5.00	3.75	5.00

Table 1: Geometric differences in mm ($\mu \rightarrow$ Mean, $\sigma \rightarrow$ standard deviation, *max* $\rightarrow$ maximum difference): $X_1 \rightarrow$ Head, $X_2 \rightarrow$ Head without pinna, $X_3 \rightarrow$ Concha, $X_4 \rightarrow$ Antihelical fold, $X_5 \rightarrow$ Antihelix, $X_6 \rightarrow$ Fossa.

Measures	*SPY*	*CAN*	*KIN*	*123D*	*PPT*
PE	$< 0.7°$	$< 0.7°$	6°	11°	12°
QE	< 0.4	< 0.4	4	6	6
S_d	< 0.5	< 0.5	< 1	1-2	1-2

Table 2: PE, QE $\rightarrow$ Increase in polar error (in degree), quadrant error (in %) and $S_d \rightarrow$ Averaged spectral difference (in dB).

SUMATRA (cf. Fig. 2). For *SPY* and *CAN*, differences of 0.75 mm and 1.5 mm occur in the concha and antihelical fold part of the pinna. Mean (μ) and maximum differences (M_x) for the head without pinnae are 0.17 mm and 0.66 mm, and 3.02 mm and 8.3 mm, respectively. For *KIN*, differences of 1.53 mm occur at the entire pinna, and μ (M_x) for the pinnaeless head are 1.5 mm (9.6 mm). For *123D*, geometric differences of 3.75 mm occur at the concha, 2.5 mm difference can be found at the antihelix. In this case, μ and M_x of 2.1 mm and 9.62 mm were found for the head. The largest differences were found for *PPT*, with 5 mm occurring at the cavum concha, and 3.75 mm at the antihelix; 1.68 mm μ and 12 mm M_x were found for the head only for this method (cf. Table. 1).

4.2. Modeled localization performance

Localization performance in elevation along the sagittal median plane was modeled according to Baumgartner *et al.* [24]. For assessing the similarity between HRIRs calculated based on different scanning methods, the decrease in localization performance was calculated as follows: The baseline-performance (reference HRIRs used as template and target for the Baumgartner model) was compared to the cross-performance (reference HRIRs used as template only). To be comparable to Baumgartner, a median listener sensitivity of $S_l = 0.76$ was used. While increases in quadrant error (QE) and polar error (PE) were only marginal for *SPY* and *CAN* (QE<0.4%, PE<0.7°), considerable increases in the range of 4-6% QE and 6-12° PE were observed for the remaining methods (cf. Table. 2).

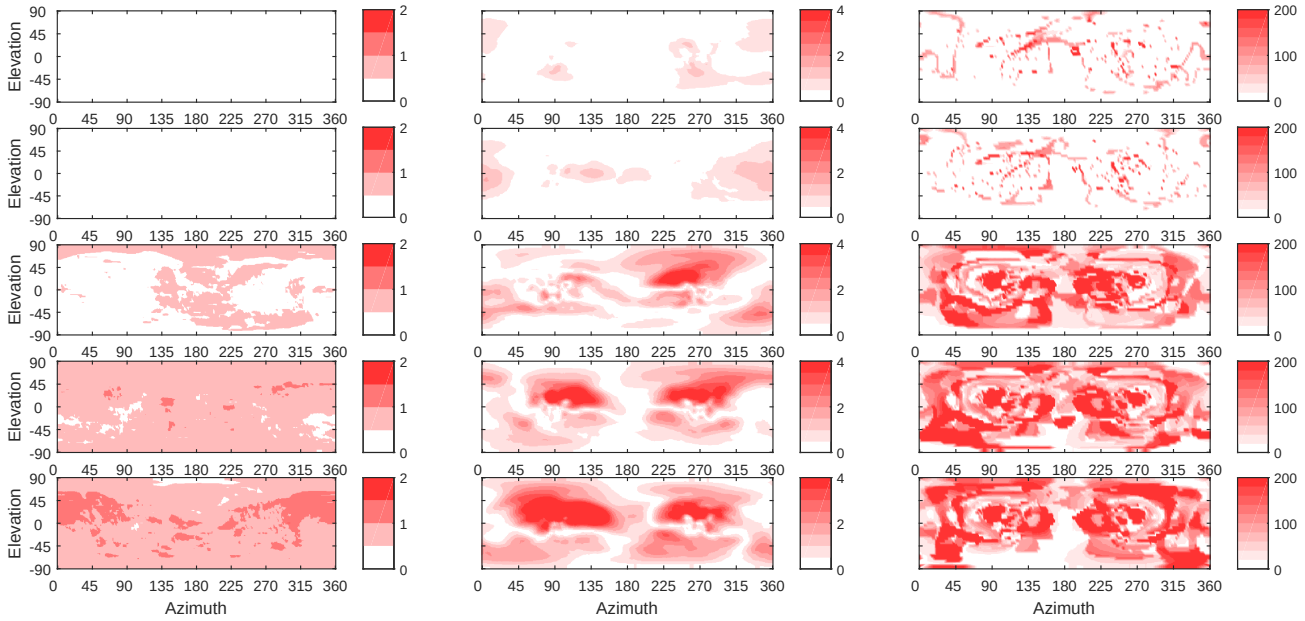

Fig. 3: Absolute differences in the HRTF magnitude spectra averaged across the entire frequency range for both ears (left column) (in dB), absolute differences in ILD (middle) (in dB) and absolute differences in ITD (right) (in μs) with respect to *GOM-Ref*: *SPY* (1st row), *CAN* (2nd row), *KIN* (3rd row), *123D* (4th row) and *PPT* (5th row). The quantization of the colorbar refers to the JND values given in the text.

4.3. Spectral differences

Spectral differences with respect to the reference were calculated in auditory filters according to Brinkmann *et al.* [8], and absolute differences averaged across frequency and ears are shown in Fig. 3 (left column). Differences for *SPY* and *CAN* are smaller than 0.5 dB (treshold found in [25]), and below 1 dB for *KIN*, while maximum differences of 1.5 dB can be observed for *123D* and *PPT* (cf. Table. 2). In general, it appears that the amount of detail in the pinnae fine structure – which is mostly affected by the different scanning methods – has only a minor influence on the overall spectral shape. However, the localization performance considerably suffers from this loss of geometry.

4.4. Differences in ITDs and ILDs

ITDs were estimated by detecting the onsets in the ten times upsampled HRIRs and absolute ITD differences with respect to the reference are shown in Fig. 3 (right column). Differences for *SPY* and *CAN* are smaller than the JND of ± 20 μs [26] for most directions, and rarely exceed ± 60 μs. For the remaining methods, differences are on average in the range of 50 μs to 150 μs, while maximum differences of about 300 μs occur outside the median plane.

Broadband ILDs were calculated based on RMS level differences between the left and right ear, and absolute differences between ILDs with respect to the reference are shown in Fig. 3 (middle column). For *SPY* and *CAN*, differences never exceed 1.5 dB, and are less than 0.5 dB for the majority of source positions, which is below the just noticeable difference (JND) [26]. For the remaining methods, differences increase to 4 dB with maxima at approximately 90° and 270° azimuth, which is expected to be audible.

5. CONCLUSION

HRTFs were numerically simulated based on 3D surface meshes obtained from six different scanning methods. To assess the scanning accuracy that is needed for an accurate HRTF simulation, geometrical differences between various scanning methods were related to perceptively motivated physical error measures for spectral coloration, median plane localization, and interaural time and level differences. Expectedly, a high precision of about 1 mm is needed when capturing the pinnae geometry to assure accurate localization cues. This criterion was met only by the *SPY* and *CAN* scanning methods. However, the overall coloration showed to be below 1 dB, even for geometric errors of up to 4 mm, which occurred for the *KIN* method. The remaining methods (*123D* & *PPT*) showed geometric deviation of up to 5 mm and a slightly larger coloration of up to 1.5 dB. In future works, listening test will be conducted to analyze perceptual differences in terms of coloration, localization and others using simulated (*SPY*) and measured HRTFs.

6. REFERENCES

[1] H. Møller, M. F. Sørensen, C. B. Jensen and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *J. Audio Eng. Soc.* 44(6), pp. 451-469, 1996.

[2] A. Fuß, F. Brinkmann, F. Jürgensohn and S. Weinzierl, "Ein vollsphärisches Multikanalmesssystem zur schnellen Erfassung räumlich hochaufgelöster, individueller kopfbezogener Übertragungsfunktionen," *Fortschritte der Akustik – DAGA 2015*, Nürnberg, Germany, pp. 1114-1117.

[3] Y. Kahana, *Numerical modelling of the head-related transfer function*. Ph.D Thesis, University of Southhampton, United Kingdom, 2000.

[4] T. Huttunen, A. Vanne, S. Harder, R. R. Paulsen, S. King, L. Perry-smith and L. Kärkkäinen, "Rapid generation of personalized HRTFs," *55th International Conference: Spatial Audio*, Helsinki, Finland, 2014.

[5] C. T. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. Van Schaik, A. I. Tew, C. Hetherington and J. Thorpe, "Creating the Sydney York Morphological and Acoustic Recordings of Ears Database," *IEEE Transactions on Multimedia*, 16(1), pp. 37-46, 2014.

[6] A. Reichinger, P. Majdak, R. Sablatnig and S. Maierhofer, "Evaluation of Methods for Optical 3-D Scanning of Human Pinnas," *International Conference on 3D vision*, France, pp. 390-397, 2013.

[7] W. Kreuzer, P. Majdak and Z. Chen, "Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range," *J. Acoust. Soc. Am.* 126(3), pp. 1280-1290, 2009.

[8] F. Brinkmann, A. Lindau, S. Weinzierl, S. van de Par, M. Müller-Trapet, R. Opdam, and M. Vorländer, "A high resolution and full-Spherical head-related transfer function database for different head-above-torso orientations," *J. Audio Eng. Soc*, 65(10), pp. 841-848, 2017.

[9] R. R. Paulsen and R. Larsen, "Anatomically plausible surface alignment and reconstruction," *Theory and Practice of Computer Graphics*, pp. 249-254, 2010.

[10] S. Harder, R. R. Paulsen, M. Larsen, S. Laugesen, M. Mihocic and P. Majdak, "A framework for geometry acquisition, 3-D printing, simulation, and measurement of head-related transfer functions with a focus on hearing-assistive devices," *Journal of Computer-Aided Design*, vol. 75, pp. 39-46, 2016.

[11] Microsoft Cooperation, Kinect for Windows, Developers, 2011. URL: https://developer.microsoft.com/en-us/windows/kinect, (checked Oct. 2017).

[12] M. Dinakaran, P. Grosche, F. Brinkmann and S. Weinzierl, "Extraction of anthropometric measures from 3D-Meshes for the individualization of head-related transfer functions," *140th Audio Engineering Society Convention*, Paris, France, Paper 9579, 2016.

[13] J. G. Fryer and J. H. Chandler, "AutoDesk 123D Catch: How accurate is it?," *Geomatics World*, vol. 2, pp. 28-30, 2013.

[14] Y. Furukawa, B. Curless, S. M. Seitz and R. Szeliski, "Towards internet scale multi view stereo," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1434-1441, 2010.

[15] P. Moulon and A. Bezzi, "Python photogrammetry toolbox: A free solution for three dimensional documentation," *ArcheoFoss*, Napoli, Italy, pp. 1-12, 2011.

[16] P. J. Besl and N. D. Mckay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2), pp. 239-256, 1992.

[17] R. R. Paulsen, "MRF-Surface," URL: http://www2.imm.dtu.dk/image/MRFSurface, (checked Oct. 2017).

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

[19] H. Ziegelwanger, W. Kreuzer and P. Majdak, "Mesh2HRTF: An open-source software package for the numerical calculation of head-related transfer functions," *22nd International Congress on Sound and Vibration*, Florence, Itlay, 2015.

[20] Blender: Open source 3D creation suite. https://www.blender.org/ (checked Oct. 2017).

[21] H. Ziegelwanger, W. Kreuzer and P. Majdak, "Effect of element size and microphone model on the numerically calculated head-related transfer functions," *International Conference on Acoustics AIA-DAGA*, Merano, Italy, pp. 600-603, 2013.

[22] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, "SOFiA Sound Field Analysis Toolbox," *International Conference on Spatial Audio*, pp. 7-15, 2011.

[23] F. Brinkmann, and S. Weinzierl, "AKtools – An Open Software Toolbox for Signal Acquisition, Processing, and Inspection in Acoustics," *142nd Audio Engineering Society Convention*, Berlin, Germany, e-Brief 309, 2017.

[24] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *J. Acoust. Soc. Am*, 136(2), pp. 791-802, 2014.

[25] F. Brinkmann, R. Roden, A. Lindau, and S. Weinzierl, "Audibility and interpolation of head-above-torso orientation in binaural technology," *IEEE J. Sel. Topics Signal Process.*, 9(5), pp. 931-942, 2015.

[26] J. Blauert, *Spatial hearing*, 2nd ed., MIT Press, Camebridge, MA, 1997.