Psychological and Physiological Acoustics
Session 1pPPa: Binaural Hearing and Binaural Techniques I

1pPPa9. Calculation of listener-specific head-related transfer functions: Effect of mesh quality
Harald Ziegelwanger*, Andreas Reichinger and Piotr Majdak

*Corresponding author's address: Acoustics Research Institute, Austrian Academy of Sciences, Wohlebengasse 12-14, Vienna, 1040, Vienna, Austria, harald.ziegelwanger@oeaw.ac.at

The geometry of head and ears defines the listener-specific directional filtering of the incoming sound. The filtering is represented by the head-related transfer functions (HRTFs), which provide spectral features relevant for the localization of sound-sources. HRTFs can be acoustically measured or numerically calculated based on a geometric representation of the listener. While the acoustically measured HRTFs usually provide localization performance similar to that obtained in free-field listening, the performance obtained with numerically simulated HRTFs, however, heavily depends on the quality of the geometric and acoustic model of the listener used for the simulation. In this study, we show how to calculate listener-specific HRTFs with spectral features similar to that from acoustically measured HRTFs for the entire audible frequency range. We review the boundary-element method coupled with the fast-multipole method and we present details on the prerequisites like the geometry-capture technique, acoustical parameters, and the numerical algorithms. Further, the effect of the mesh quality on the HRTFs was investigated by systematically varying the average edge length from 2-5mm. The HRTF amplitude spectra were analyzed and evaluated by visual comparison and in a localization model. The results of the coarser meshes show indications for the required edge length in HRTF calculations.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Head-related transfer functions (HRTFs) describe the direction-dependent filtering of the incoming sound at the entrance of the ear canal due to scattering, shadowing, and reflection caused by the head, torso, and the pinna (Blauert, 1997). Thus, HRTFs depend on the particular details of a listener's geometry. In the frequency range up to 16 kHz, HRTFs contain listener-specific spectral features, which are required for accurate sound source localization in binaural reproduction systems (Möller, 1992). As an alternative to the common acoustic measurement of the HRTFs (Majdak et al., 2007) where small microphones are placed into the ear canals, the HRTF simulation, where HRTFs are numerically calculated from a representation of a listener's geometry, offers potential for a consumer-suitable, non-contact HRTF acquisition method. A lot of research effort has been done on the calculation of HRTFs by the use of the boundary element method (BEM) in the last years (Katz, 2001a, 2001b; Kahana and Nelson, 2006, 2007; Gumerov and Duraiswami, 2009; Kreuzer et al., 2009). However, HRTFs yielding a localization performance as good as that for acoustically measured ones have not been calculated yet. The results of the HRTF simulation heavily depend on the quality of the listener's geometry data, i.e., three-dimensional mesh, which results from the scanning procedure and the post-processing required for the numerical calculations. Particularly, the mesh accuracy and the mesh resolution, i.e., the number of elements in relation to the frequency of interest, are important properties. Also acoustical parameters like the impedance model of hair and skin have been investigated (Treeby et al., 2007a, 2007b, 2007c). However, requirements on these geometry properties and acoustic parameters for HRTF calculation have not been clarified yet.

In this study we aimed at calculating listener-specific HRTFs similar to that obtained with the acoustical measurement procedure and yielding a localization performance as good as that observed in free-field listening. We assumed, that the accuracy of the mesh is vital and the impedance of the hair can be neglected. Thus, in order to reach this goal it was required to work with the best possible mesh quality and accuracy, regardless of the required effort. While a method suitable for possible future consumers should be as fast, convenient, and pleasant as possible, these restrictions do not hold for this test case. Capturing the human pinna was of special interest. Spectral features, which are indispensable for accurate sound localization, are caused by the individual shape of a listener's pinna, in particular, by its folds, undercuts, and deep cavities. However, these geometric details exacerbate the scanning procedure. Parts of some pinnae are not even visible from the outside, which makes surface scanning methods (e.g., laser scanner) unsuitable. Thus, we propose a geometry capturing procedure, which overcomes these difficulties, evaluate the calculated HRTFs and investigate the effect of the mesh resolution on sound source localization.

METHODS

HRTF Measurement

HRTFs were acoustically measured for 1550 directions in a semi-anechoic chamber with the blocked ear-canal method (Möller et al., 1995). The subject was seated in the center of a circular loudspeaker array (see Fig. 1a). To reduce the intensity of reflections the array construction was covered with acoustic damping material. HRTFs were measured with an exponential frequency sweep (0.05 to 20 kHz). The total time required to measure the HRTFs was decreased by applying the multiple exponential sweep method (Majdak et al., 2007). In-ear-microphones (Sennheiser KE-4-211-2), which were placed in the subject's ear canals, were connected via amplifiers (RDL FP-MP1) to the digital audio converters. The subject was rotated horizontally to measure HRTFs at several elevations at once for one azimuth by playing the sweeps and recording the acoustic signals with the microphones. The horizontal and vertical range was 360° and -30° to 80°, respectively. The horizontal resolution was 2.5° and 5° for directions inside and outside of ±45°, respectively. The vertical resolution was 5°. The position and orientation of the head was tracked during the measurement to observe and reduce head movements.

Equipment transfer functions were derived from a reference measurement in which the in-ear microphones were placed in the center of the array and impulse responses were measured for all loudspeakers. The equipment transfer functions were cepstrally smoothed and their phase spectrum was set to the minimum phase (Oppenheim et al., 1999). The resulting minimum-phase equipment transfer functions were removed from the HRTFs by spectral division. Then, directional transfer functions (DTFs) were calculated using a method similar to the procedure of Middlebrooks and Green (1990). Finally, the directional impulse responses, i.e., the inverse Fourier transforms of the DTFs, were windowed with an asymmetric Tukey window (fade in of 0.25 ms and fade out of 1 ms) to a 5.33-ms duration (256 Samples at sampling rate of 48 kHz). The resulting DTFs are shown in Fig. 3 (AC).
FIGURE 1. (a) Subject during the acoustic HRTF measurements. (b) Subject during the visual scan of the head. (c) Reference mesh of the subjects head with approximately 2.5 million triangular elements.

Geometry Acquisition

Pinna

In order to achieve the most detailed representation of the pinna, a CT scanner was used. Since CT is a volumetric scanning method, the scan is not affected by undercuts and occlusions. The medical scanning-devices typically have too poor resolution and a too high noise level for our purposes. In addition, all CT scanners expose the subject to harmful x-ray radiation. The required scan quality, however, can be achieved with an industrial high-energy CT scanner, where the amount of radiation is not limited by medical concerns. To prevent the subjects from the radiation we made impressions of the ears and scanned them instead of the subjects. Note that other volumetric scanning methods like the magnet-resonance tomography might have been used, however, these methods usually yield global distortions.

The impressions were made of silicone. In order to prevent deformation of the pinna and to capture possibly all of the surface, we used an addition-crosslinked precision silicone with a very low viscosity (viscosity: 2200 MPa·s, shore hardness: A 22). The subjects sealed their ear canals with silicone ear protectors\(^1\) to prevent the silicone from reaching the ear drum. The silicone was pressed-in as deep as possible without being inconvenient for the subject, at least some millimeters past the ear canal opening. The subjects lied on the side, their hair was covered with Vaseline, a cut paper cup was placed over their ears, sealed with modeling clay, and the freshly prepared silicone was slowly filled in with a constant thin flow (see Fig. 2a). The ear was moved to release any large air bubbles captured in cavities. This way, only very small air bubbles remained in the silicone. The silicone material takes about half an hour to harden. After this time, the impression was carefully pulled off the ear, and set aside a day to fully harden. Torn hair was removed from the impression, and it was cleaned from rests of modeling clay. Thin unstable parts were cut resulting in a sufficiently rigid impression.

By using the impressions, the hair on the ears was fully ignored, and only the skin surface was impressed. The hair was either pressed away, or formed very narrow canals that did not affect the CT-scan results. The local surface was very accurate, and some effort was put to also reduce global deformations. The silicone material initially had a very high viscosity leaving the ear in its natural position. After hardening it became hard but elastic (see Fig. 2b).

The impressions were scanned with an industrial CT scanner (\(\text{vctome}^\text{TM} \times \text{c}^2\)) using 1000 x-ray images spaced in 0.36° intervals (tube settings: 120 kV, 260 \(\mu\)A, average of three images with 333-ms exposure time each). During the scans, we avoided any stress on the material. The impressions were placed horizontally on the scanner with its flat side on the platform. No other attachments were required. The scanning platform moved abruptly, but the scanner was set to wait 333-ms before the scanning in order to let material oscillations decay.

The resulting volumetric data sets had a cubic voxel size of 50 to 60 \(\mu\)m, totaling to approx. 2 Gigavoxels with 16-bit integer density values. The transition between air and silicone had a width of typically 3 voxels (see Fig. 2c). The isosurface separating air and silicone were extracted using auto-thresholding (VGStudio Max, Volume Graphics GmbH\(^2\)). This method is a proprietary version of the marching cubes algorithm (Lorensen and Cline, 1987) that extracts the isosurface with sub-voxel precision and curvature-dependent increase in resolution. The effective accuracy can be estimated to be in the range of two to four voxels, i.e., 0.1 to 0.2 mm. From the resulting mesh, we manually removed the surface part produced by the cup wall and the top surface.
Despite the careful handling of the silicone, air bubbles were inevitable. Most of the bubbles occurred inside the silicone and have been removed by keeping the largest connected surface. However, several bubbles were very close to the surface, leaving a too thin layer of silicone for the CT resolution to find the correct isosurface. In addition, some bubbles actually were on the surface. At these locations, the ear geometry could not be recovered. Most of the small bubbles have been removed based on a curvature threshold that detects the sharp connections between the bubbles and the ear surface (algebraic point set surface, APSS, MeshLab®). The remaining bubbles that were not removed with this procedure, have been manually cut and all holes have been filled (Automatic filling method in curvature mode, Geomagic, Geomagic Studio®).

**Head**

The head was scanned by a laser scanner (ZScanner 700CX®, ZCorp) projecting two orthogonal laser lines which are manually swept over the scanned surfaces (see Fig. 1b). This scanner features a built-in optical tracking system based on reflective points that are placed directly on the scanned object. Therefore, the scanned parts are always registered relative to the head, ensuring that a globally consistent and accurate model is created, even when the head is moving. We used the highest resolution settings and a tightly set working volume around the head. In order to cope with hair that cannot be scanned, the subjects wore a custom tailored cap with holes for the ears. The mesh, exported from the scanning software, was apparently extracted from an internal volumetric representation with approximately 0.5-mm voxel size and with holes where data could not be captured. These holes where filled with the curvature-based filling algorithms (Geomagic), and the neck was cut and closed.

**HRTF Calculation**

Free-field HRTFs were calculated using the collocation boundary element method (BEM) with constant elements. To avoid irregular frequencies we used the approach of Burton and Miller (1971). To calculate HRTFs up to 20 kHz in an appropriate time we coupled the BEM with the fast-multipole method (Kreuzer et al., 2009). Calculation time was further decreased by applying the principle of reciprocity (Morse and Ingard, 1987), where the receiver is exchanged with the transmitter. Computation was done on a Linux cluster containing eight machines with Intel i7-3820 processors running at 3.6 GHz and 64 Gbyte of RAM each.

BEM requires a watertight mesh of the head scan and the two pinnae. The pinna meshes were aligned to the pinnae of the head mesh (Manual Registration, Geomagic) using manually placed tie points, and refined (Best Fit Alignment, Geomagic) using only that part of the pinna as alignment reference, which was visible from the side. This was the most reliable part since it was best captured by the surface scanning method of the head, and thus it ensures that the acoustically relevant parts are oriented most accurately. The scan results differed noticeably in the joint region of the head behind the pinna. This was mostly caused by the cap used during the head scan, which slightly enlarged the head. Also a slight rotation of the pinna between both scans might have contributed to the differences. We gave precedence to the pinna scan, since it captured the hard-to-see region behind the ear most detailed and remodeled the head scan to fit seamlessly. To this end, we removed the pinna region of the head scan, and stitched the pinna scans. The seam-region was re-sculpted and smoothed (Voxel-Sculpting, 3D-Coat®) in a
resolution of below 0.5 mm voxel-size, taking care to only modify the head mesh. Figure 1c shows the final high accuracy reference mesh, which consisted of approximately 2.5 million triangles. In the following, this mesh is referred to as the base mesh.

Because of memory size, the number of elements used in the BEM was limited to approximately 120,000 elements. Thus, the resolution of the reference mesh was reduced in three steps. First the whole mesh was re-meshed to 1 mm side-length on average ensuring approximately equilateral triangles (Remesh, Geomagic). Then everything except the pinna region was re-meshed to 1.5 mm, the pinna region expanded by 4 polygon rows and the rest further re-meshed to 2.5 mm. The resulting mesh (see Fig. 3a BASE) had approximately 1 mm side-length at the pinna region, 2.5 mm at the head and a transition region with 1.5 mm side-length that was additionally relaxed to reduce degenerate non-equilateral triangles at the region seams. Therefore, most of the polygon budget (approx. 45%) was used for the pinna region, where it is acoustically most relevant, and was reduced continuously to the acoustically less important head region. For acoustic BEM simulations the average edge length of the underlying mesh should lie between a one-tenth and one-sixth of the observed wavelength (Marburg, 2002). Thus, for the numerical HRTF calculation up to 20 kHz an average edge length between 1.7 and 2.8 mm is preferred. Our base mesh fulfills these requirements for the hole boundary surface. To investigate the effect of increasing the average edge length, the surface was uniformly re-meshed to an average edge length of 2, 3, 4, and 5 mm (Openflipper³, Möbius and Kobbelt, 2012).

The microphone was modeled as sound source by applying a corresponding velocity boundary condition at the blocked entrance of the ear canal. For the rest of the head and pinna the normal velocity was set to zero, i.e., the boundary surface was modeled fully reflective. The impedance of the hair was neglected. The complex sound pressure was calculated at 200 sampling points in the frequency domain, which were linearly spaced between 100 and 20,000 Hz. HRTFs were calculated for 16,022 directions at a distance of 1.2 m. The resolution was 2° for the lateral and the polar angle. After the inverse Fourier transform, we obtained the head-related impulse responses (HHRIs), which were re-sampled at a sampling rate of 48 kHz. Finally, DTFs were calculated following the same procedure as for the acoustically measured HRTFs.

RESULTS

First, the calculated DTFs for the base mesh were evaluated by visual comparison of the resulting amplitude spectra to the acoustically measured data. Figure 3c shows the amplitude spectra in the median plane for the acoustically measured (AC) and calculated (BASE, 2 to 5 mm) DTFs of subject's left ear. Note, that the acoustically measured DTFs were band-limited at 18 kHz and the plot is affected by the coarser spatial sampling in the acoustic measurement setup. The dark parts in the amplitude spectra represent notches, which are the relevant features for sound source localization in the median-plane. They are caused by interference of superposed pinna reflections at the entrance of the ear canal. The fine horizontal lines, which look like torso reflections (we have not included the torso in the mesh) might be caused by the abrupt cutting of the mesh at the bottom end of the neck without any smoothing. Overall, the acoustically measured DTFs (AC in Fig. 3) and DTFs calculated for the base mesh (BASE in Fig. 3) seem to be similar.

A more psychoacoustic-related way to evaluate the calculated DTFs is to analyze the sound localization in a localization experiment where a subject listening to sounds filtered with the calculated DTFs responses to the perceived direction. The localization performance can be described by the amount of front-back confusion, i.e., the quadrant error rate, and the local error, i.e., the local polar RMS error (Middlebrooks, 1999). In our study, the localization performance has been modeled in a sagittal-plane localization model (Baumgartner et al., 2013), which considers listening to sounds filtered with non-individualized DTFs like our calculated ones. The model results are two-fold: 1) response probability of pointing to a direction, and 2) the localization performance parameters, i.e., quadrant error and polar RMS error. Figure 3d shows the response probability to a direction for given polar angles of a virtual sound source, in a simulated localization experiment (baumgartner2013 from the AMToolbox³; (Søndergaard and Majdak, 2013)) with the measured and calculated DTFs. The localization model considered a frequency range from 5 to 18 kHz. For the measured DTFs the model predicted a polar error of 29° and a quadrant error rate of 9.3 %. For the base-mesh DTFs, the model predicted increases in the polar and quadrant error of 1° and 1.7 %, respectively. We assume that such small differences can be neglected. Thus, we consider the base-mesh DTFs as perceptually valid.

Further, the HRTFs calculated using meshes with less elements were analyzed. The details on the average edge length, number of elements, and computation time are shown in Table 1. The calculated HRTFs are shown in Fig. 3b and 3c (2 mm to 5 mm). Note the less distinct notches between 5 and 15 kHz for larger average edge lengths of the
FIGURE 3. (a) Pinna meshes. (b) DTF amplitude spectra for the interaural horizontal plane (brightness: magnitude in dB). (c) DTF amplitude spectra for the median plane. (d) Response probability matrix (brightness: response probability). AC: Acoustically measured DTFs. BASE: Base mesh (pinna: 1-mm edge length; head: 2.5 mm).
TABLE 1. Mesh parameters, computation time required to calculate HRTFs, and the modeled localization performance.

<table>
<thead>
<tr>
<th>Average Edge Length (mm)</th>
<th>Number of Elements</th>
<th>Computation Time (hours)</th>
<th>Polar Error (°)</th>
<th>Quadrant Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pinna Head</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acoustic Measurement</td>
<td>2 488 720</td>
<td>-</td>
<td>29</td>
<td>9.3</td>
</tr>
<tr>
<td>0.4</td>
<td>2 488 720</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>111 456</td>
<td>19</td>
<td>30</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>87 706</td>
<td>14</td>
<td>32</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>44 190</td>
<td>8</td>
<td>33</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>25 442</td>
<td>6</td>
<td>38</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>16 464</td>
<td>4.5</td>
<td>40</td>
<td>17</td>
</tr>
</tbody>
</table>

meshes (Fig. 3c). Also, the modeled response probabilities (Fig. 3d) suggest a more blurry localization with larger edge lengths, an effect further revealed by the localization performance parameters (Table 1). Generally, the localization performance monotonously decreased with increasing average edge length, showing indications for the required average edge length in future HRTF calculations.

CONCLUSIONS

In this study, we investigated a method for capturing the geometry of the human head and pinna, and for calculating HRTFs, aiming at satisfying requirements for perceptually valid HRTFs. The BEM was coupled with the fast-multipole method to calculate free-field DTFs for frequencies up to 20 kHz. With the smallest average edge length computed, measured and calculated DTFs were similar. The results of a sagittal-plane localization model suggest a similar localization performance with the measured and calculated DTFs. Further results on meshes with a larger average edge length indicate that the requirements on the mesh quality can be lowered.

1 http://www.cecem.de/gehoerschutz.html, last viewed on 18.1.2013
3 http://www.volumeographics.com, last viewed on 18.1.2013
4 http://meshlab.sourceforge.net, last viewed on 18.1.2013
5 http://www.geomagic.com, last viewed on 18.1.2013
7 http://3d-coat.com/, last viewed on 18.1.2013
8 http://www.openflipper.org, last viewed on 18.1.2013
9 http://sourceforge.net/projects/amtoolbox/, last viewed on 18.1.2013

ACKNOWLEDGMENTS

We thank Christa Rothner for the cap design, Florian Rist (Institute of Arts and Design, Vienna University of Technology) for the ear impressions, Daniel Habe (Österreichisches Gießerei-Institut, Leoben) for the CT Scans, and Matthias Kucera (Ludwig Boltzmann Institute for Archaeological Prospection and Virtual Archaeology, Vienna) for the head scans. We are also grateful to our subjects for their patience during the longsome measurements. This study was supported by the Austrian Science Fund (FWF, P 24124-N13).
REFERENCES


