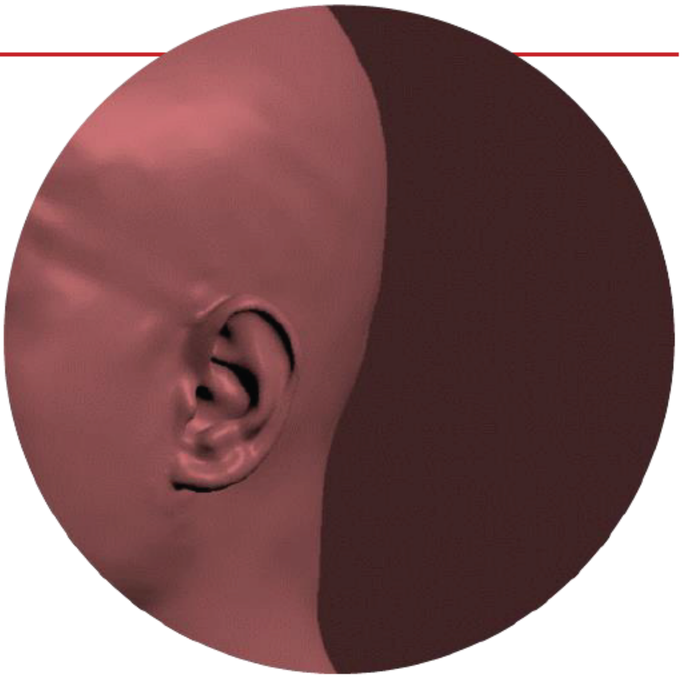
Master Thesis

Perceptually motivated analysis of head-related transfer functions for individualization

Robert Pelzer



Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt gegenüber der Fakultät I der Technischen Universität Berlin, dass die vorliegende, dieser Erklärung angefügte Arbeit selbstständig und nur unter Zuhilfenahme der im Literaturverzeichnis genannten Quellen und Hilfsmittel angefertigt wurde. Alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind kenntlich gemacht. Ich reiche die Arbeit erstmals als Prüfungsleistung ein. Ich versichere, dass diese Arbeit oder wesentliche Teile dieser Arbeit nicht bereits dem Leistungserwerb in einer anderen Lehrveranstaltung zugrunde lagen.

Mit meiner Unterschrift bestätige ich, dass ich über fachübliche Zitierregeln unterrichtet worden bin und diese verstanden habe. Die im betroffenen Fachgebiet üblichen Zitiervorschriften sind eingehalten worden. Eine Überprüfung der Arbeit auf Plagiate mithilfe elektronischer Hilfsmittel darf vorgenommen werden.

Berlin, den 3. Juli 2018

Robert Pelzer

Master Thesis

Perceptually motivated analysis of head-related transfer functions for individualization

Robert Pelzer

Submitted by
Robert Pelzer
[redacted]

For the academic degree
Master of Science (M.Sc.)

Submitted to
Technische Universität Berlin
Fakultät I - Geisteswissenschaften
Institut für Sprache und Kommunikation
Fachgebiet Audiokommunikation
http://www.ak.tu-berlin.de

Submission date
July 3, 2018

Supervision
Prof. Dr. Stefan Weinzierl
Fabian Brinkmann
Manoj Dinakaran

Date of the defense
August 7, 2018

Abstract

Head-related transfer functions (HRTFs) provide all the necessary information a listener needs to perceive and localize sounds in binaural synthesis. HRTFs can very accurately be obtained using acoustic measurements or numerical simulations with 3D meshes but both methods are time-consuming and need extravagant technical equipment. Therefore, HRTF individualization based on anthropometric features (AFs) of the head and ears is a desirable ambition.

In this work, a comprehensive database of HRTFs for 93 subjects was created. The database includes HRTFs obtained by acoustical in-ear measurements in an anechoic chamber and numerically modeled HRTFs which were created using 3D meshes and BEM calculation. To generate 3D meshes of all subjects a Kinect scanner was used to scan the heads and torsos whereas an Artec spider scanner was used to create high-resolution pinna scans. Scans for each subject were post-processed and merged to generate models appropriate for BEM calculation. For the reproducibility of some work steps, a semi-automatic workflow was used.

Most AFs were automatically extracted using an approach by Dinakaran et al. (2016). Pinna rotation and flare angle were semi-automatically detected via Python scripts. A correlation of AFs was calculated showing the deep linkage between various features.

Cross-validation was used to compare measured and modeled HRTFs showing few individual deviances but an overall good fit on average.

A listening test was carried out with 42 subjects using dynamic binaural synthesis. In the first part of the listening test subjects evaluated different HRTFs against their own HRTF. In the second part of the listening test, subjects rated differences between their own modeled against their own measured HRTF. The ratings during the listening test were given in respect to different perceptual qualities.

Statistical analysis using a hierarchical mixed-effects model was used to predict HRTF ratings by AFs. Parameter tuning, including feature interaction and squared features helped to account for the complicated interdependency of AFs.

Finally, effect sizes were estimated for each AF by determining the predictor importance for each feature. Hence, key features for HRTF individualization could be identified the top 3 key features being *pinna rotation angle*, *cavum concha width* and *cymba concha height*. This showed that the pinna-related features are prior to head and torso-related features.

In a second statistical approach, support vector machine regression models were used to recommend the most appropriate HRTF from the database based on a person's AFs.

Zusammenfassung

Kopfbezogene Übertragungsfunktionen (HRTFs) liefern für die Binauralsynthese alle erforderlichen Informationen, die ein Hörer benötigt, um Klänge wahrzunehmen und zu lokalisieren. HRTFs können sehr präzise durch akustische Messungen oder durch numerische Simulationen mit 3D-Modellen generiert werden, jedoch sind beide Methoden zeitaufwendig und erfordern anspruchsvolle technische Messmittel. Eine HRTF-Individualisierung, die auf anthropometrischen Merkmalen des Kopfes und der Ohren basiert, ist daher ein erstrebenswertes Ziel.

In dieser Arbeit wurde eine umfassende Datenbank, bestehend aus 93 HRTFs erstellt. Die Datenbank enthält sowohl HRTFs, die durch akustische Messungen im Gehörgang der Probanden erstellt wurden als auch HRTFs, die auf Basis von 3D-Modellen durch BEM-Berechnung generiert wurden. Um 3D-Modelle aller Probanden zu erstellen, wurde ein Kinect-Scanner zum Erfassen der Köpfe und Torsos verwendet. Ein Artec-Spider-Scanner wurde zur Erstellung detailreicher Ohrmuschelscans verwendet.

Ein Großteil anthropometrischer Merkmale wurde automatisch mithilfe einer von Dinakaran et al. (2016) entwickelten Methode extrahiert. Die Rotations- und Neigungswinkel der Ohrmuschel wurde semi-automatisch mithilfe von Python-Skripten ermittelt. Eine Korrelation anthropometrischer Merkmale wurde berechnet, um wechselseitige Beziehungen verschiedener Merkmale aufzuzeigen.

Eine Kreuzvalidierung wurde angewandt, um gemessene und berechnete HRTFs zu vergleichen. Dabei wurden zwar einige individuelle Abweichungen sichtbar, aber es zeigte sich ein im Durchschnitt insgesamt gute Übereinstimmung.

Ein Hörversuch mit dynamischer Binauralsynthese wurde mit 42 Probanden durchgeführt. Im ersten Teil des Versuchs verglichen die Probanden einige fremde HRTFs mit der eigenen. Im zweiten Teil des Hörversuchs beurteilten die Probanden Unterschiede zwischen der eigenen berechneten und gemessenen HRTF. Die Bewertung erfolgte jeweils im Hinblick auf verschiedene perzeptive Qualitäten.

Ein hierarchisches Mixed Model wurde anschließend verwendet, um HRTF-Bewertungen mittels anthropometrischer Merkmale vorherzusagen. Quadrierungen sowie Interaktionen von Prädiktoren trugen dabei der komplizierten Interdependenz anthropometrischer Merkmale Rechnung.

Schließlich wurden die Effektstärken für jedes Merkmal geschätzt wodurch Schlüsselmerkmale für die HRTF-Individualisierung identifiziert werden konnten. Die drei wichtigsten Merkmale sind: Rotationswinkel der Ohrmuschel, Breite der Cavum Concha und Höhe der Cymba Concha. Hiermit wurde verdeutlicht, dass ohrmuschelbezogene Merkmale ausschlaggebender als Merkmale des Kopfes oder des Rumpfes sind.

In einem zweiten statischen Ansatz wurden Support Vector Machine Regressionsmodelle verwendet, um basierend auf anthropometrischen Merkmalen die passendste HRTF aus der Datenbank für eine Person zu bestimmen.

Acknowledgements

I would like to thank Stefan Weinzierl for this interesting topic and Fabian Brinkmann for a great job of supervising, especially when navigation through dense undergrowth was needed. I also want to thank Manoj Dinakaran for his help with 3D-mesh processing and his spontaneous help during the listening test.
Thanks to Michael Horn, who provided help with the custom version of SoundScape Renderer.
I also want to thank Steffen Lepa for his advice and help in statistical matters.

I want to thank all participants, who provided the data for this work by sitting through HRTF measurements and 3D scanning. Special thanks you, who even came back for the listening test.

I especially thank Tirza for having my back and for her support. I also thank my kids for being patient with me.

Table of Contents

List of abbreviations

AEL	*Average Edge Length*
AF	*Anthropometric Feature*
ERB	*Equivalent Rectangular Bandwidth*
FFT	*Fast Fourier Transform*
FHP	*Frankfurt Horizontal Plane*
HpTF	*Headphone Transfer Function*
HRIR	*Head-Related Impulse Response*
HRTF	*Head-related Transfer Function*
ICP	*Iterative Closest Point*
ILD	*Interaural Level Difference*
ITD	*Interaural Time Difference*
IV	*Independent Variables*
ML-FMM	*Multi-Level Fast Multipole Method*
NHP	*Natural Head Position*
OSC	*Open Sound Control*
PCA	*Principle Component Analysis*
PD	*Pure Data*
PE	*Polar Root-Mean-Square Error*
PEQ	*Parametric Equalization*
QE	*Quadrant Error Rate*
SE	*Standard Error of the Mean*
SH	*Spherical Harmonics*
SHT	*Spherical Harmonics Transformation*
SL-FMM	*Single-Level Fast-Multipole Method*
SSR	*Residual Sum of Squares, SoundScape Renderer*
SST	*Total Sum of Squares*
SVM-R	*Support Vector Machine Regression*
χ^2_{Diff}	*Chi-Square Difference*

1 Introduction

1.1 Motivation

When we hear sounds, this happens because the ear converts small pressure changes that cause the eardrum to vibrate into sound information, which in turn is perceived and processed by the brain. Binaural synthesis in virtual acoustics aims to reproduce the signals at the eardrums. The idea behind binaural synthesis is, that if a technical system produces the same sound pressures a real source would produce at the listener's ear canal entrances and eardrums, then the recipient perceives a virtual source that cannot be distinguished from a real source (Møller et al., 1996). However, for two different recipients listening to the same sound source, the information at the eardrums would not be identical because the signals that arrive at the eardrums are distorted and shaped by the listeners head, the torso and the pinnae (outer ears) (Møller, 1992). These individual characteristics have been analyzed by our brains throughout our whole lives to self-train and fine-tune our very accurate auditory localization performance in correspondence to the visual system (Hartmann, 1999). The human brain has learnt to decode interaural level differences (ILD) and interaural time differences (ITD) on both ears to determine localization of a sound source in the lateral dimension (Blauert, 1997). Even though, small head movements are used to distinguish a source coming from front or back (Møller, 1992), ITD and ILD are not sufficient for precise vertical localization of sound sources on the median plane which are equidistant from both ears. Due to their shape, the pinnae are responsible for a complex combination of diffractions, resonances and reflections which are direction and distance dependent. For example, reflections caused by the ear canal entrance funnel, the cavum concha, are needed for localization of sources on the median plane (Hebrank and Wright, 1974). Through these spectral cues, spatial information of a sound field is coded into spectral and temporal attributes (Blauert, 1997).

The influence of the head, the torso and the pinnae on acoustic signals at the eardrum has been described as head-related transfer function (HRTF). HRTFs are directional transfer functions that are defined as the ratio of sound pressure at the blocked ear canal to the sound pressure at the center of the head, the interaural center with the head of the listener absent (Møller, 1992). Therefore, in binaural synthesis HRTFs provide all the necessary information that the listener needs to perceive and to localize sounds (Nicol, 2010).

Very accurate ways of measuring HRTFs are by either placing microphones in the ear canals and measuring acoustic signals for the complete audible frequency spectrum for all directions as in (Masiero et al., 2012) or by performing numerical simulations using individual 3D meshes as in (Katz, 2001a). However, both methods require elaborate equipment, specific skills to perform the measurements and can be time-consuming and uncomfortable for the listener. Hence, it seems appealing to do HRFT individualization based on anthropometric features (AFs) of the head and ears. Key features which influence the HRTFs could either be used to choose a best fit from an HRFT dataset or to numerically create a HRFT. However, anthropometric key features to be idiosyncratic still must be studied and found. This work wants to contribute to that question and find relevant AFs of pinna and head for HRTF individualization.

1.2 State of the art

1.2.1 Numerical HRTF simulations

Creating virtual sound sources using HRTFs is not something new and has been studied since the early 90s. Over two decades ago Bronkhorst (1995) showed that virtual sound sources can almost be as accurately localized as real sound sources.

So far, the standard approach to acquire HRTFs has been in an acoustical manner by placing a small microphone in the entry of the subject's ear-canal who is sitting on a chair in an anechoic environment. However, using acoustic measurements can be a difficult and resource-demanding procedure and oftentimes is uncomfortable for the listener who must sit through multiple frequency sweeps.

HRTFs can be numerically calculated using the boundary element method (BEM) in which the Helmholtz equation which describes an acoustic wave in a domain is being transformed into a boundary integral equation. The boundary surface is being discretized to solve this integral equation numerically. This calculation of the HRTF is made on the assumption that only the surface characteristics of the head are relevant and propagation through the head is ignored (Katz, 2001a). Furthermore, the human skin was shown to be acoustically rigid, while hair is responsible for absorptions (Katz, 2001b). Sufficient frequency resolution was shown to be 100Hz. A minimum of six elements on the 3D mesh was suggested per period of an acoustic wave which limits the shortest possible wavelength to six element edges (Katz, 2001a).

Using numerical BEM calculations for HRTFs has been tested and evaluated in several studies and showed to yield a sound localization performance similar or better to acoustically measured HRTFs (Ziegelwanger et al., 2015b, Ziegelwanger et al., 2013)

This work uses open-source software *mesh2HRTF* [1] (Ziegelwanger et al., 2015a) which numerically calculates HRTFs by an implementation of the 3-dimensional Burton-Miller collocation BEM which is coupled with the multi-level fast multipole method (ML-

[1] https://sourceforge.net/projects/mesh2hrtf/ (accessed June 20, 2018)

FMM) and therefore saves calculation resources. In more general terms *mesh2HRTF* reads geometrical data, calculates the corresponding sound field and outputs HRTFs.

Calculation time and resources can additionally be saved by calculating HRTFs using a reciprocal approach. This requires the virtual microphone which is placed at the entrance of the ear canal and the virtual loudspeaker to be interchanged. Hence, the calculation of one active and vibrating element results in the sound pressure information for all nodes (Kreuzer et al., 2009).

A third way to save calculation resources is achieved by a using a remeshing algorithm. Ziegelwanger et al. (2016) showed that mesh grading prior to HRFT calculation reduces computation time which would otherwise be in the order of days per subject to around ten percent. Their work also showed that HRTF calculation for graded meshes returned similar or even better results than for high-resolution uniform meshes regarding numerical accuracy and the predicted sound-localization performance.

1.2.2 Influence of anthropometric features on HRTFs

In HRTFs the lower frequency regions are primarily shaped by head diffractions and torso reflections (Algazi et al., 2001a). The pinnae only start contributing to the HRTF at around 3.5kHz (Katz, 2001a). Figure 1 and Figure 2 exemplary show characteristic peaks and notches as the first pinna notch at around 7 to 9kHz, depending on the position. Figure 1 shows magnitude spectra for four different positions on the horizontal plane and in Figure 2 all azimuth angles on the horizontal plane are shown. A head-related impulse response (HRIR) in the time domain is shown in Figure 3 also on the horizontal plane.

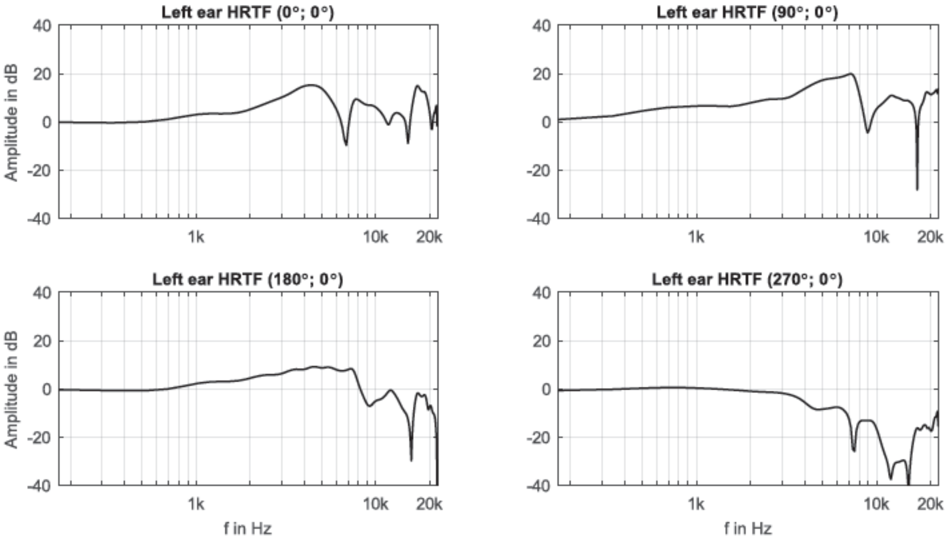


Figure 1: HRTF of subject 7 (modeled) on horizontal plane for the left ear with source from 4 different locations. Pinna notches and peaks can be identified. The first pinna notch is visible at around 7kHz for a source coming from front and at around 9kHz four a source pointing at the left ear.

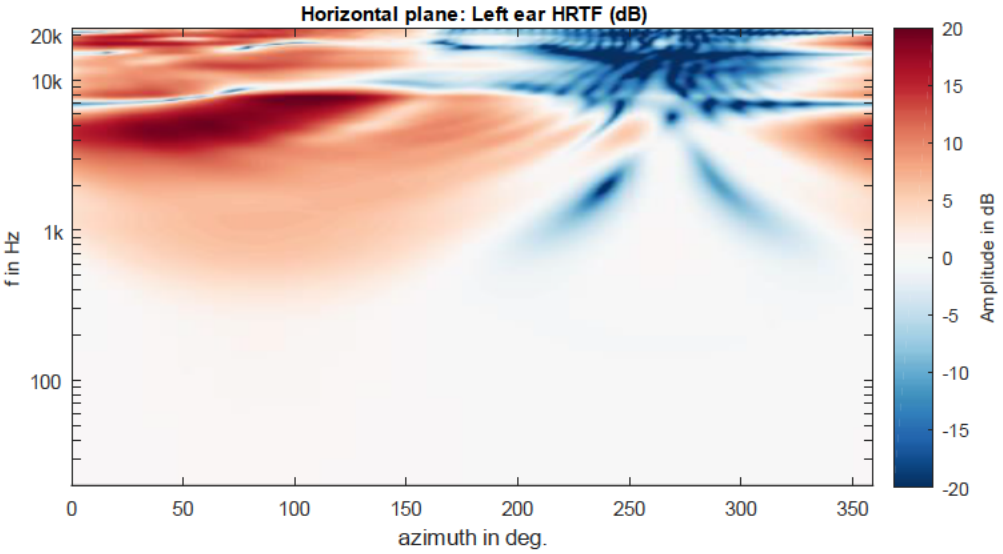


Figure 2: HRTF of subject 7 on the horizontal plane. The first pinna notch is the vertical blue structure between 7 and 9kHz.

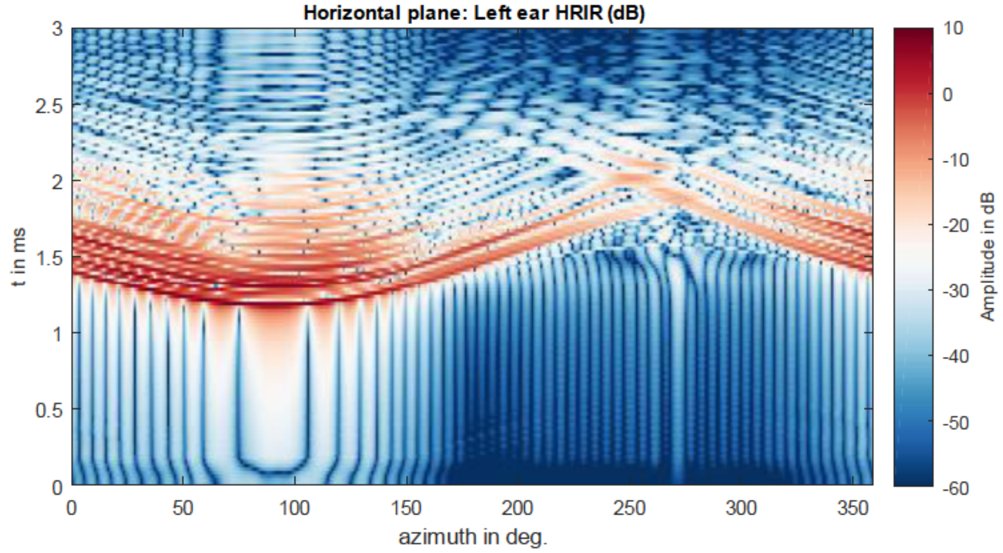


Figure 3: HRIR of subject 7 on the horizontal plane

As mentioned above, ILD and ITD provide primary cues for azimuth localization for sources coming from left and right. It seems obvious that ITD is strongly correlated to head width, i.e. the inter-tragus distance (Algazi et al., 2001b). The cavum concha seems to play a major role for localization on the sagittal plane as mentioned earlier. A basic model of interference and delay paths caused by the cavum concha is introduced by Batteau (1967). The already mentioned paper by Hebrank and Wright (1974) describes how cavum concha reflections add to median-plane localization. Lopez-Poveda and Meddis (1996) even developed a physical model of the human concha which accounted for some of the reflection and diffraction but only approximations could be achieved due to the complicated and interdependent spectral behavior of the pinna.

The analysis of peaks and notches of the HRTF and their origin was taken further by Takemoto et al. (2012), who compared HRTFs and pinna-related transfer functions and who studied pressure nodes and anti-nodes. Particular peaks and notches were selected from the HRTF which delivered frequencies and source positions for following examinations of pressure distribution patterns on the pinna using sinusoidal waves. They found that the pinna generates the basic peak-notch patterns and that the basic pattern was common between different subjects, however, with differing fine structures. The study also showed that peak frequencies are not sensitive to changes in the source elevation angle, whereas notches are. Takemoto et al. (2012) also showed that among subjects the frequency and the location of nodes and anti-nodes was different. They concluded that the relationship between the individual pinna geometry and the frequencies and amplitudes of spectral peaks and notches still could not be clarified because of the complexity of the pinna shape.

So far, the findings on the attempt to clearly assign specific AFs to peaks and notches of HRTFs suggest that the scattering and the diffraction of incident sound waves by the pinna is a complex process that is difficult to describe. The statistical relationship between HRTFs and anthropometric parameters seems to be highly complicated and cannot be described by the linear combination of a few pinna-related parameters (Xie, 2013).

1.2.3 HRTF individualization

By definition, HRTFs express a linkage between the listener's morphology and his or her hearing. It follows, that similarity in anatomy leads to similarity in HRTFs (Middlebrooks, 1999a). However, individualization of HRTFs by means of anthropometric key features does not necessarily include an understanding of why a certain feature contributes to the HRTF. Middlebrooks (1999a) used frequency scaling and hypothesized that for different participants different anatomical dimensions would result in similar spectral features at different frequencies. Middlebrooks found that the scaling method is successful in aligning major spectral features along the frequency axis but that it does have no effect on idiosyncratic features such as the detailed shapes of peaks and notches. It was also shown that localization errors could be reduced by scaling another person's HRTF to the listeners dimension. However, this procedure which is based on the assumption of different subjects having similar anatomical shapes with differing features can only be seen as a rough approximation (Xie, 2013).

Algazi et al. (2001b) measured HRTFs for 45 subjects including 27 anthropometric head, torso and pinna features. These measurements are widely used and known as the CIPIC HRFT database. A set of AFs was defined, and identifiers were assigned which are used in this work as well (see Table 2). In their work Algazi et al. also looked for AFs correlating to the frequency of the first pinna notch (f_{pn}). They found the cavum concha height to be the best predictor but the pinna angles and the fossa height also to be among top predictors for f_{pn}.

Other approaches tried to achieve HRTF customization by taking anthropometric features into consideration. Zotkin et al. (2002) used a course closest neighbor approach and the CIPIC database to show that localization performance and liking significantly increases by just measuring the anthropometric features of subjects and comparing them to the measures of the database.

To further improve HRTF individualization, several approaches tried to identify anthropometric top predictors on the basis of which a matching HRTF from a data base could be selected. Zotkin et al. (2003) introduced HRTF personalization using a set of seven AFs and an additional low-frequency head-and-torso model. Again, CIPIC HRTF database was used und the study concluded that the incorporation of the head-and-torso model enhances localization performance, whereas the personalization based on AFs does not always perform well.

Liu and Zhong (2016) reduced Zotkin's list of seven features to four features (namely cavum concha height, pinna height, pinna rotation angle and pinna flare angle). Their matching method included the analysis of correlation results of pinna parameters and spectral distortion comparison, also using CIPIC database. The authors finally proposed a customization method which was tested in a localization performance test. The results suggested that the study's method is supposedly prior to Zotkin's.

Xu et al. (2007) also based their research on the CIPIC database and studied the morphological influence on HRTFs by correlation analysis and principle component analysis (PCA). The authors suggested that three factors explain most of the HRTF amplitude, namely a factor consisting of shoulder-related measures and two factors representing rotation and flare angles.

Some studies have based their research on self-generated data which seems important to compare and generalize findings. Fels and Vorländer (2009) varied different AFs for pinna and torso via CAD and BEM calculation which allowed them to examine isolated effects. Results showed that the pinna shape does not influence ITD but has a clear influence on ILD. Their results also showed that the distance between ear and shoulder and the head depth have a huge influence on HRTFs. Among pinna features, their findings suggest that the cavum concha depth and width have the greatest influence and that also the rotation angle of the pinna plays an important role. This underlines the influence of the cavum concha on HRTFs once more and suggests that also the pinna rotation angle should be considered.

Bomhardt (2017) used a database of acoustically measured HRTFs and 3D ear models (Bomhardt et al., 2016) to examine the influence of AFs on HRTFs by applying the previous mentioned scaling method (Middlebrooks, 1999a) in which the frequency vector was scaled with an optimal scaling factor to minimize inter-subject spectral differences for two different HRTFs. Bomhardt analyzed the relationship between AFs and the scaling factor and found the cymba concha height to have the strongest positive influence. Bomhardt's work (2017) also included individualization of HRTFs by PCA using AFs. Individualization results were evaluated by testing the localization performance which revealed that individual HRTFs work slightly better for subjects than individualized HRTFs.

Ghorbal and Auclair (2017) showed that the aim to identify meaningful AFs for HRTF individualization has led to various results from several studies which overlap in some cases and are contradicting in others. The authors also stated that a consensus has not been found yet and try to contribute to that goal by varying a set of AFs through deformation of meshes and BEM calculation. The study focused on AFs which were identified from previous studies including cavum concha height and width, cymba concha height, fossa height, pinna height and width as well as pinna rotation and flare angle. The authors identified cavum concha width as the parameter with the highest effect, followed by fossa height.

It can be summarized that HRTF individualization based on AFs has been analyzed and examined in several studies. However, besides finding a consensus about the cavum concha in accordance with earlier mentioned reflection and attenuation theories, still no appropriate set of key features has been identified. Also, only few studies consider the subjects' perception of HRTF individualization results and if doing so, usually only localization performance is being examined.

1.3 Outline

This work wants to contribute to HRTF individualization by using perceptual analysis to identify anthropometric key features. Choosing a perceptual approach rather than performing spectral analysis, as seen in previous works, has the advantage of going straight to the core of the question.

First a new HRTF dataset will be created consisting of acoustically measured and BEM calculated (modeled) HRTFs for 93 subjects. In sections 2.1 to 2.6, 3D-meshes which are needed for BEM calculation will be created, covering head and pinnae. For mesh alignment and mesh preparation before BEM calculation, various Python scripts will be used to create a semi-automatic work flow in sections 2.5 and 2.7. Under 2.11 AFs will be measured automatically using an algorithm by Dinakaran et al. (2016) while some features will be semi-automatically determined using a Python script.

In 2.12 a listening test will be designed and conducted with 42 subjects who will compare other subjects' HRTFs against their own HRTF. In a second part of the listening test, subjects will additionally compare their own modeled version against their own measured version of HRTF. The listening test will be carried out including dynamic binaural synthesis. Individual headphone equalization filters will be created for each subject under 2.12.3. The rating of HRTFs during the listening test will happen in respect to various audio qualities, i.e. SAQI items (Lindau et al., 2014).

Comparison of acoustically measured and BEM modeled HRTFs will be carried out in several ways. Cross-validation under 3.1 will visually and by means of different error measures compare measured and modeled HRTFs and state average results. Secondly, listening test results for the evaluation of similarity between measured and modeled will be analyzed under 3.14. Additionally, a comparison of rating means will be carried out under 3.3.

Statistical analysis will focus on the central question of this work: How can HRTF individualization work based on anthropometry? Therefore, two different approaches are being followed. One approach focuses on the identification of anthropometric key features while the other approach focuses on the recommendation of suiting HRTFs from a dataset.

In sections 3.5 and 3.8 to 3.11, multilevel mixed-effects regression models will be used to determine regression coefficients for each SAQI item. Effect sizes will be determined for each AF under 3.12. This will allow a ranking of AFs by predictor importance and therefore deliver a list of key features. For the second approach under 3.13, a support vector machine regression will be used to build a recommender which can moderately predict the fit of HRTFs from the dataset for a subject based on anthropometrics.

2 Methods

2.1 Preliminary Work

For this work 3D surface meshes of listeners' head, torso and pinnae were generated for 93 subjects. The subjects were between 13 and 61 years old, the age median being 34. From the 10 female and 83 male subjects there were 54 who were affiliated with TU Berlin, 32 joining from Sennheiser and 7 subjects coming from Huawei Technologies in Munich. All subjects except of two had listening test experience. Some had previously taken part in more than 20 listening tests, the median being 8 listening tests.

The head and upper torso area was scanned with a low-cost scanner, the Kinect sensor[2] as in (Dinakaran et al., 2016). The Kinect uses RGB camera and depth sensors to capture 3D environments with a resolution of 0.5 mm accuracy (Yang et al., 2015). During this procedure the subjects were seated on a swivel chair and were instructed to look straight forward, focusing a point in the far distance. The subjects' hair was covered using a swim cap to create a smooth surface for 3D scanning and to reduce the influence of hair on the scans. The Kinect 3D scanner, and the Kinect fusion with the developer toolkit browser v1.8.0[3] were used to generate 3D surface meshes by slowly rotating the subjects by 360 degrees. The Kinect sensor was set up about a meter away from the subjects and was set to its maximum resolution of 748 voxels per meter. The Kinect scans where then manually post-processed in *Meshlab*[4]*,* an open source system for processing and editing 3D meshes. During this post-processing step, first, unsuitable parts in the mesh such as surroundings, unwanted parts of the torso, and irregular parts close to holes were removed and secondly, holes were filled. In this way meshes representing the head and the facial shape sufficiently were created.

The pinnae were scanned as in as in (Dinakaran et al., 2018). A much higher resolution was accomplished via an Artec Space Spider scanner[5] which uses blue structured light technology to achieve 0.05 mm point spacing accuracy. This means that the Artec Space Spider captures 10 times more detailed meshes than the Kinect sensor. To create a single mesh again multiple scans were captured at multiple angles.

In addition to the 3D scans the subjects' HRTFs were measured acoustically by placing miniature electret condenser microphones in the ear canals of subjects as in (Lindau and Brinkmann, 2012). The measurements took place inside the low-reflection room of TU

[2] https://www.microsoft.com/en-us/store/d/kinect-sensor-for-xbox-one/91hq5578vksc (accessed June 20, 2018)

[3] https://developer.microsoft.com/en-us/windows/kinect (accessed June 20, 2018)

[4] www.meshlab.net/ (accessed June 20, 2018)

[5] https://www.artec3d.com/3d-scanner/artec-spider (accessed June 20, 2018)

Berlin by using a fully spherical multi-channel measurement system (Fuß et al., 2015) shown in Figure 4.

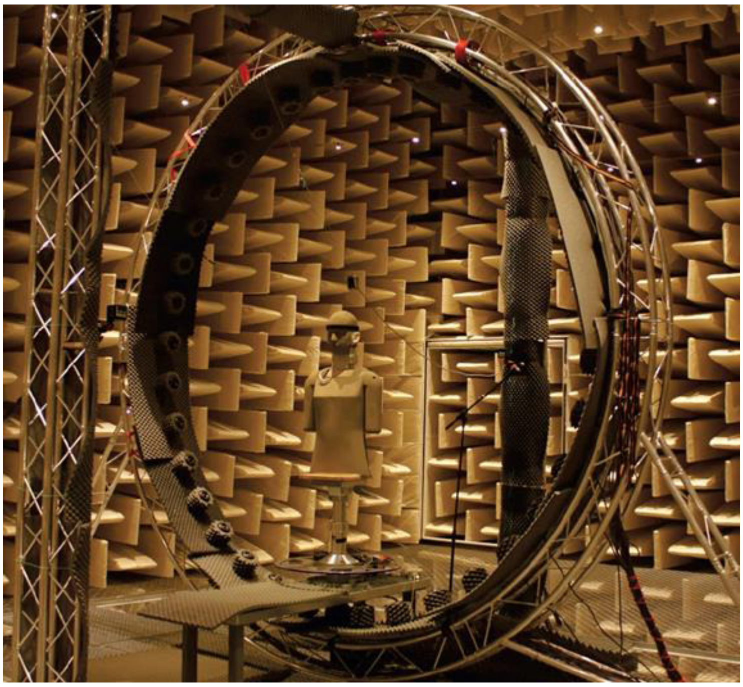


Figure 4: Multi-channel measurement system inside the low-reflection room of TU Berlin. The Photo was taken from (Fuß et al., 2015)

Also, for each subject a photo was taken in which the position of the microphone in the ear canal was marked. An example with FABIAN (Lindau and Weinzierl, 2006) can be seen in Figure 5.

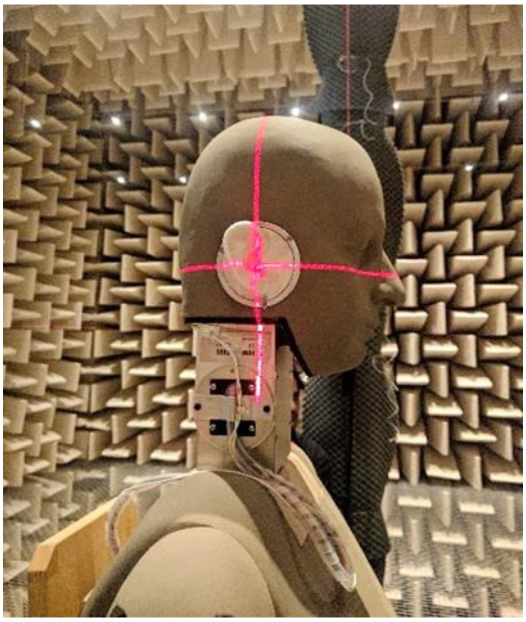


Figure 5: FABIAN head in the low-reflection room of TU Berlin with marked position of microphone in the ear canal

To cancel out the on-axis frequency responses of microphones, amplifiers and speakers a reference microphone was placed in the position of the subject, with the subject absent. HRTFs were then obtained via spectral division.

The above mentioned preliminary work was conducted by Fabian Brinkmann from TU Berlin and by Manoj Dinakaran from Huawei Technologies.

2.2 High-resolution pinna meshes

The high-resolution scans of pinnae generated by the Artec scanner had to be post-processed and cleaned to produce usable 3D meshes. The Artec scanner creates a file format that can only be post-processed and opened with Artec Studio[6], Artec's data processing software which can be purchased in form of an annual subscription plan. For this work an *Artec Studio 12 Professional* license was used to import raw scanner data consisting of all the unaligned meshes for the current project. An example is shown in Figure 6. In this work the number of single scans per ear ranged from 9 to around 25 scans.

Most of the RAM and CPU exhaustive steps of post-processing were accomplished on a PC at TU Berlin with a dual-boot system consisting of Ubuntu 14 and Windows 7. The PC was supplied with 32 GB of RAM and an 8-Core Intel i7-4790K processor with 4 GHz.

During the procedures described below the weakest link was the RAM size, even though the available memory seemed sufficient at first glance. This issue became especially relevant in BEM calculations but also during the phase of post-processing of Artec scans. For that reason, once the most applicable post-processing procedure was found, the command history was disabled in Artec Studio 12 under *File/Settings/Performance*. This made corrections in the form of undoing a command impossible but saved a lot of computation time and made the following steps faster and smoother.

To generate high-resolution pinna meshes the following procedure delivered the best and quickest results:

First the meshes were auto-aligned using the auto alignment tool. This roughly aligned the meshes so the basic details of the pinna could be recognized (Figure 7). Single scans which did not align properly during this process were usually deleted because generally the number of scans was more than sufficient to create an acceptable outcome. In most cases 6 scans would have already been enough to create high-resolution pinna meshes but of course the area the scans captured also played a role.

In a second step the erase tool was used to delete outer areas which either did not contribute to the pinna or consisted of noisy parts caused by hair on the edge of the swim cap (Figure 8). This reduced the file and RAM size and made the following steps faster

[6] https://www.artec3d.com/de/3d-software/artec-studio (accessed June 20, 2018)

and more accurate. A part of the plane representing the shape of the head around the pinna was left because it was useful during the further processing of alignment and combining meshes.

The third step consisted of the fine alignment called *global registration* (Figure 9). *Global registration* converts all one-frame surfaces to a single coordinate system using information on the mutual position of each surface pair. To do so, it selects a set of special geometry points on each frame, followed by a search for pair matches between points on different frames.[7] For the algorithm to perform correctly the initial approximation that was achieved by automatic pre-alignment was necessary. The settings for the step were changed to "geometry" because of the rich geometry and poor texture of the pinna. The settings for "minimal distance" and "iterations" were left in default values of 50 mm and 2000. After *global registration* the maximal error which is shown in the workspace on the right side has usually decreased to 0.1mm. Meshes with errors above 0.2 mm were deleted if they did not contain fundamental and unique information. However, if a scan held unique information that no other scan captured (for example the area behind the ear) the project was re-assessed, and more cleaning and editing was done before redoing *global registration.*

Artec Studio online documentation[8] suggests performing *outlier removal* following *global registration* to remove more noise in the mesh. In the case of this work it was found that this can be a time-consuming task taking up to a few hours because *outlier removal* is based on a statistical algorithm that calculates the mean distances between every surface point and a certain number of adjacent points. Therefore, the elimination of 3D-noise was done in a later step and *global registration* was followed by *fusion,* a step which creates a polygonal 3D model by solidifying and melting the captured frames (Figure 10). To do this a "sharp fusion" was carried out with the option "watertight" selected to close any holes in the mesh. Because *outlier removal* was not done prior to the fusion small objects disconnected to the main mesh were usually visible. This remaining noise now in the form of small disconnected particles could be cleaned by applying a "small object filter" with settings set to default. A trade-off between detail and further calculation time was chosen by setting the resolution for "sharp fusion" to 0.5mm. Because further mesh simplification is applied for BEM simulation later, this measure seemed reasonable. The final mesh was than exported choosing the Stanford triangle (.ply) format.

[7] http://docs.artec-group.com/as/12/en/process.html#global-registration (accessed June 20, 2018)

[8] http://docs.artec-group.com/as/12/en/process.html#eliminating-3d-noise-outlier-removal (accessed June 20, 2018)

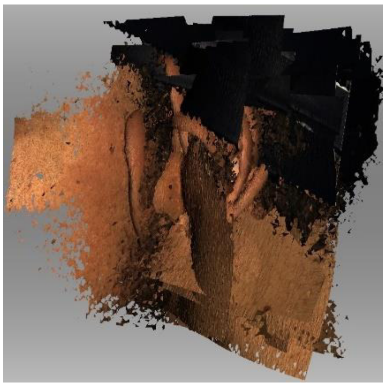


Figure 6 Artec Studio – raw and unaligned meshes

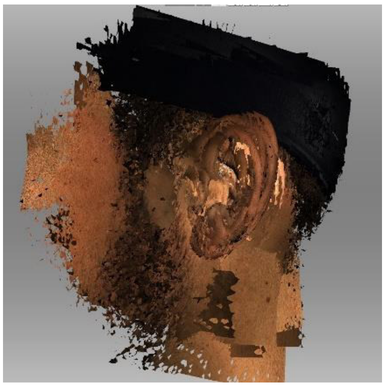


Figure 7: Artec Studio – auto-aligned mesh

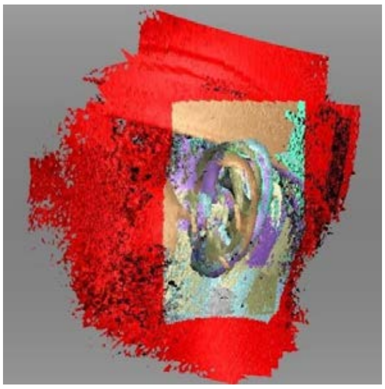


Figure 8: Artec Studio – red area will be deleted using the erase tool.

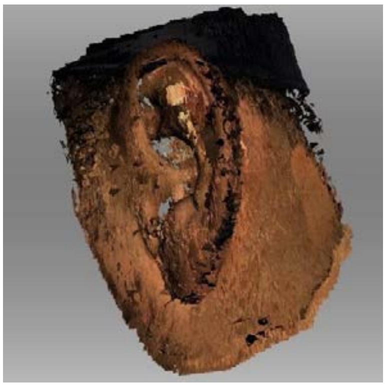
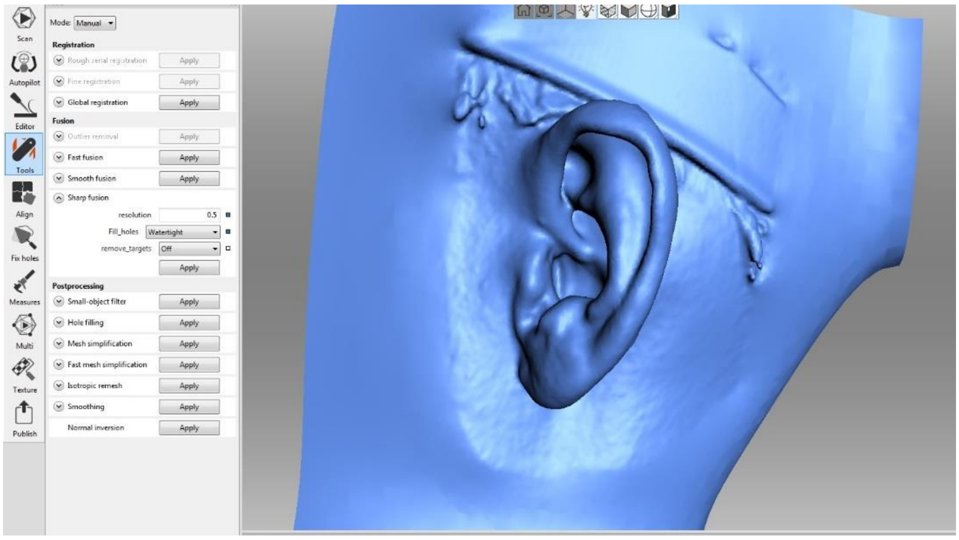


Figure 9: Artec Studio – mesh after global registration. Fine alignment reduces the error.



Mode: Manual

resolution 0.5 ■

Fill_holes Watertight ■

remove_targets Off □

Figure 10: Artec Studio – final mesh after sharp fusion process and small object filter application

2.3 Closing ear canals

As mentioned earlier, HRTFs are directional transfer functions that contain information of sound events at a defined location, namely the blocked ear canal (Møller, 1992). Many of the high-resolution pinna scans that were taken with the Artec scanner still showed the opening of the ear canal. To reach comparable results for all subjects, the ear canals were closed in such a way that the bottom of the cavum concha as the deepest spot closed the ear canal with a flat surface, as if a microphone would be placed in the ear canal (Figure 11 and Figure 12). This procedure was done by Manoj Dinakaran.

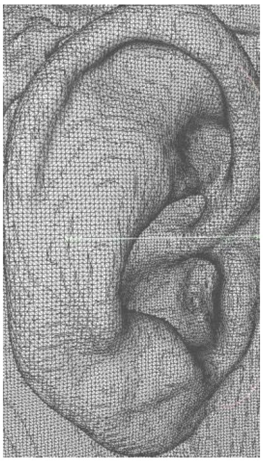


Figure 11: a subject's ear with remains of ear canal entrance

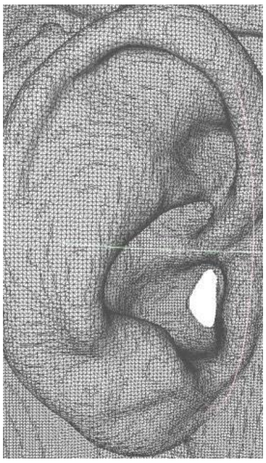


Figure 12: creating a blocked ear canal closed by adding a flat surface

Side note:

During the mesh processing procedure it was found that mesh centering as described under 2.5 is easier to perform before ear canal openings are being deleted. Therefore, this closing of ear canals was carried out after the centering step.

2.4 Merging pinna and head

When both, the Kinect scan of the whole head and the high-resolution pinna were post-processed both had to be aligned to each other and merged into a single mesh. This was done using the Geomagic Studio's point-based glue[9]. Figure 13 shows a pinna in two different resolutions: one mesh created with the Kinect and the other one taken with the Artec scanner.

[9] http://www.geomagic.com/en/products-landingpages/re-designx-wrap (accessed February 21, 2018)

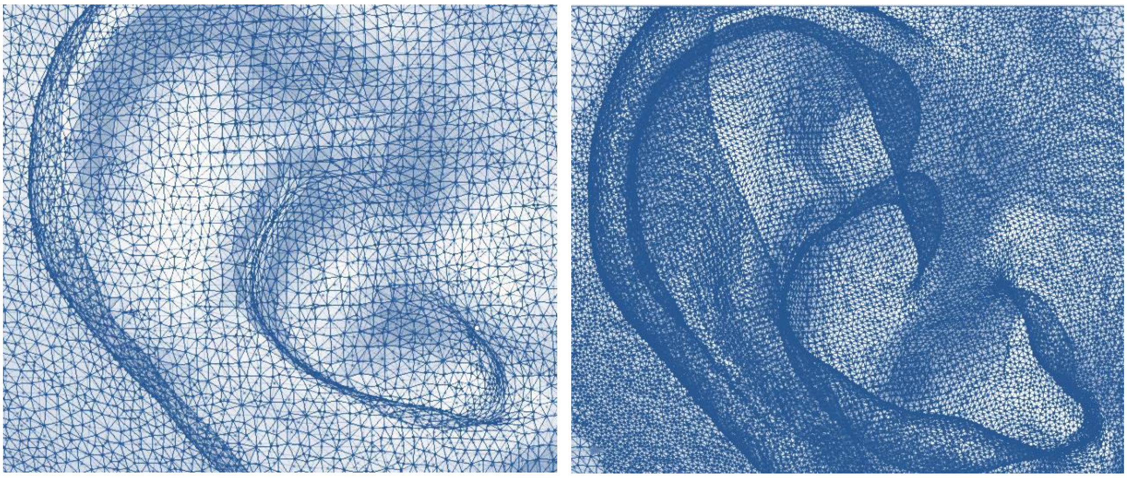


Figure 13: pinna of subject 7 in different resolutions: Kinect mesh with a lower resolution on the left and an Artec scan with a much higher resolution on the right

Additionally, the torso was deleted at defined position at the end of the neck leaving most part of the neck. Reproducibility for this task was achieved by selecting two points on the neck automatically. Similar to Dinakaran et al. (2016) an outline of the 3D mesh in side view was extracted as shown in Figure 14. On the front of the neck the point is determined by the minimum position in x-direction and on the back of the neck the gradient on z-axis serves to calculate the selected point.

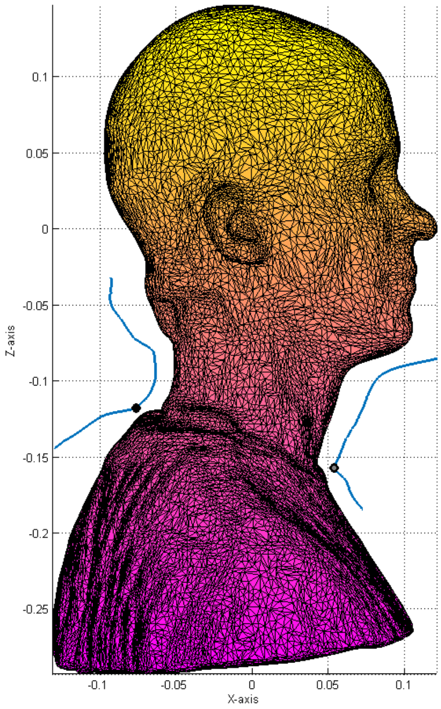


Figure 14: Side view of 3D surface mesh with the extracted outlines for determining defined points for a neck end

Deleting torsos from the meshes had several reasons. One was that the torso was not captured fully and the amount that was captured during the scan with the Kinect sensor also varies depending on the subject's height. Secondly, only the frontal position was captured, hence for multidirectional HRTF simulation including a torso, other head directions would have to be captured as well. Thirdly, the torso adds many elements to the mesh which means that BEM calculation time would increase dramatically.
This working step was completed by Manoj Dinakaran.

2.5 Centering and Natural Head Position

Before starting BEM calculations, it was vital that the subjects' meshes were being aligned identically. This was important for reproducibility and for the following steps because during alignment the designation of ear canal positions took place indirectly. The meshes were aligned to the interaural center, the ear canal being the y-axis with the left auricular point on the positive half-axis and the nose pointing in positive x-direction. The z-axis pointed upwards. In HRTFs, by definition, the interaural center is the coordinate origin (Møller, 1992). Aligning the meshes as said also followed the approach of Ziegelwanger et al. (2015).

Because manual alignment of meshes to their interaural center can be a fiddly task set out for irreproducible results this work followed a semi-automatic approach. Meshes were imported into *Blender*[10], an open source 3D modelling software equipped with a Python API. A Python script for semi-automatic alignment was written for which the selection of three points in the mesh is required. The first point to be selected was the center of the left ear canal entrance, the second point the right ear canal entrance and the third point a position in the center of the face, e.g. the nose tip or preferably a spot between the eyes. Choosing the ear canal entrances as reference points, special anthropometric occurrences such as one ear being vertically displaced from the other or being further back did not change the defined alignment of the mesh. In other words, the interaural axis was the same as the y-axis and were determined by the first two selected points.

The execution of the script moved the interaural axis of the mesh on the y-axis in such a way that the third point determining the center of the head was in the coordinate origin of the y-axis. A more detailed explanation can be found in the header of the script[11] itself.

The described procedure corrected sideways tilts, head rotation and general mesh disposition in a comfortable way by selecting only three reference points. However, it should be noted that especially deciding on the center of the ear canal entrances can be a some-

[10] https://www.blender.org/ (accessed June 20, 2018)

[11] Digital_Appendix\11 Scripts\Python_Blender\head_centering.py

what ambivalent task. Sometimes a clear ear canal entrance could not be identified because the mesh did not cover the entrance. In these cases, a thorough evaluation and comparison of both sides was done and the Artec scans of the pinnae were examined to gain additional information on where to place the position marks.

The head centering script did not correct any head rotations around the interaural axis. However, a natural head position (NHP) is of great importance for later calculations because it determines the exact position that an unelevated, frontal signal is coming from. In other words, should the subject's true NHP be something else than in the mesh the signal heard through the simulated HRTF might feel horizontally off. In cephalometry, the study and measurement of the head, a common way to determine whether a head or skull is in its NHP is using the *Frankfurt Horizontal Plane* (FHP). At the craniometric conference in Frankfurt am Main in 1882 the plane through left and right porion, which is the point of the human skull at the upper margin of the ear canal and the left orbitale, which is the bony socket of the eye, was decided to best represent the horizontal plane parallel to the surface of the earth (Moorrees, 1994). However, on one hand the meshes and pictures in this work do not contain mere bone or skull structures and on the other hand further research showed that the FHP is somewhat outdated in terms of representing NHP because there seems to be a discrepancy of a few degrees (Ramírez et al., 2013).

NHP is defined as a reproducible position of the head obtained when the subject is in a relaxed position, sitting or standing, looking into the horizon or into an external reference point at eye level (Ramírez et al., 2013). The way that 3D scanning with the Kinect sensor took place fulfills these criteria, so that NHP was achieved for the subjects in most cases. After running the head centering script, the naturality of the head position was manually examined. For approximately 30 % of the subjects a manual correction of the NHP was applied by rotating the mesh around y-axis for 5 to 7 degrees.

Finally, the mesh was exported again in Stanford triangle (.ply) format with "vertex colors" being the only checked option.

2.6 Remeshing

Before BEM calculation could start, a re-meshing of the 3D-models was an important measure to reduce the number of elements in the meshes by gradually increasing the element sizes. Ziegelwanger et al. (2016) introduced a mesh-grading algorithm available as an implemented plugin for the open-source tool *Openflipper*[12] which was used to remesh the meshes. The re-meshing process created two different model versions per subject, one with the left ear in high resolution and one with the right ear, respectively. The

[12] https://www.openflipper.org/ (accessed June 19, 2018)

algorithm gradually re-meshed the meshes starting at the ear canal, i.e. the origin of the y-axis at the corresponding pinna. Hence, the previous step of centering meshes is vital because ear canal positions were defined.

In this work a target length of 1 mm to 14 mm was used which reduced the number of elements from approximately 45000 to 15000 in most cases. Meshes with an average edge length (AEL) of 1mm were shown to lead to a sound localization performance similar or better to acoustically measured HRTFs (Ziegelwanger et al., 2015b). The main reason to choose a 1mm to 14 mm grading was to reduce the number of elements and computation time as much as possible. A closer look on computation time corresponding to the number of elements will be taken under 2.8.

Re-meshing deforms the contralateral ear and the shape of the face as can be seen in Figure 15.

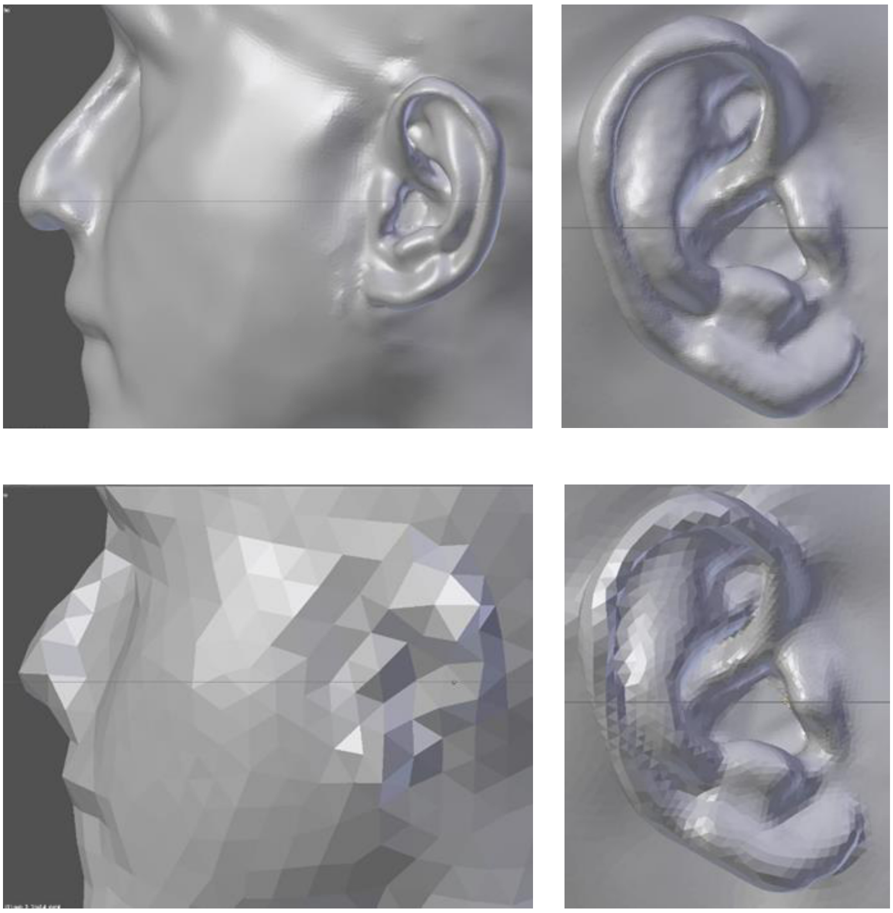


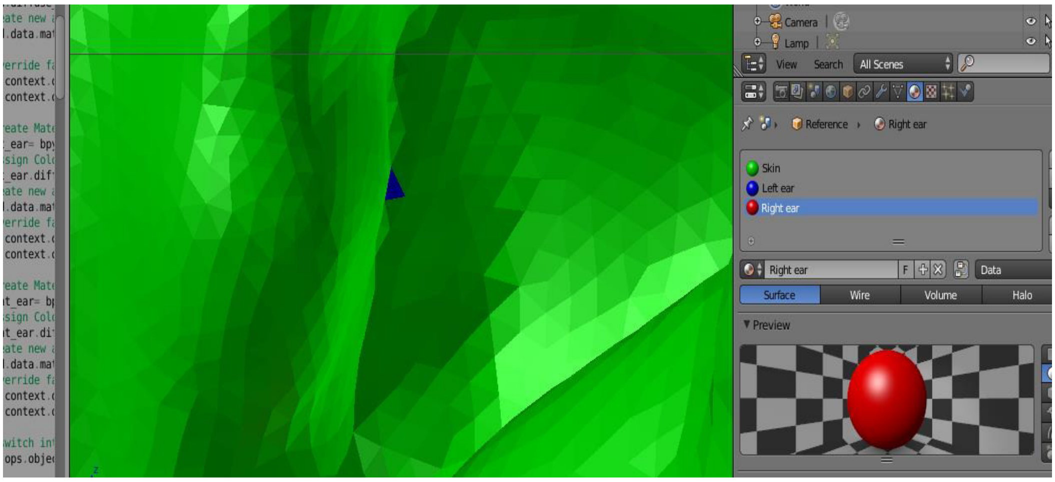
Figure 15: original and re-meshed meshes with applied 1 to 14 mm mesh-grading algorithm for subject 2. The left side shows the contralateral side before and after re-meshing. The right side shows the right pinna prepared for BEM calculation on the bottom and before re-meshing on the top.

2.7 Preparation of mesh2HRFT input files

Before BEM calculation using *mesh2HRFT* could start, input files had to be generated. A plugin that comes with *mesh2HRTF* and had to be installed in *Blender* generated *mesh2HRFT* input file format[13]. For installation the file “export_mesh2hrtf.py” had to be installed manually and after installation the option “NumCalc” was added in the selectable export options.

Once a certain mesh was imported into Blender several steps had to be be carried out. First, a definition of the microphone element needed for the reciprocal approach in which the roles of microphone and loudspeaker are interchanged by defining a vibrating element at the blocked ear canal, as described earlier had to be carried out. The ideal size for the virtual microphone was found to be 1mm in radius (Ziegelwanger et al., 2015b). This criterion was fulfilled through selection of only one single element with an AEL of 1mm as mentioned before.

Defining a microphone element at the corresponding ear in the ear canal happened by selecting an element and assigning a material to that element which had to be called “Left_ear” and “Right_ear,” respectively. Another material called “Skin” had to be assigned to the rest of the mesh. Finally, the whole object name had to be changed to “Reference”. Figure 16 shows a screenshot of the material assignment and the selected left ear element.



```
eate new a
].data.mat

erride fa
 context.c
 context.c

eate Mate
_ear= bpy
sign Colo
_ear.dif
eate new a
].data.mat
erride fa
 context.c
 context.c

eate Mate
t_ear= bp
sign Colo
t_ear.di
eate new a
].data.mat
erride fa
 context.c
 context.c

witch int
ops.objec
```

Figure 16: Material assignment in Blender using a Python script. The selected ear canal element is marked blue.

To simplify the working process of material assignment and renaming the object, a Python script[14] was created which carried out material assignment and renaming the object automatically. For selecting the left and right ear canals, elements lying on y-axis where searched. For the re-meshed and somewhat deformed contralateral ear, the correct

[13] https://sourceforge.net/p/mesh2hrtf/wiki/Mesh2Input/ (accessed June 20, 2018)

[14] Digital_Appendix\11 Scripts\Python_Blender\material_and_assignment.py

element could be found by choosing the outermost element on the y-axis. For the ipsilateral ear canal which is often covered by the tragus in y-direction (see Figure 16) the innermost element on y-axis of the corresponding direction had to be chosen to prevent the tragus from being defined as the ear canal. To determine which side to handle as ipsilateral side the name of the original mesh was analyzed by the script. Of course, a condition which had to be fulfilled was the appearance of the concerning side in the mesh name. Therefore, in this work re-meshed left meshes were unanimously called “sub_xx_1to14_left.ply”.

After the input file was created using “NumCalc” export several specific settings were selected. First, the concerning ear, for which calculations should be executed had to be determined. Also, an evaluation grid had to be defined. For this work a Lebedev grid (Lebedev, 1977) served as the evaluation grid because it has a higher efficiency than an equiangular Gauss quadrature with a similar number of nodes (Bernschütz, 2016). The Lebedev grid was generated using the SOfiA-Toolbox (Bernschütz, 2011) and had a radius of 1,5 m and 1730 nodes which allowed for a maximal spherical harmonics (SH) order of 35. (After BEM calculation the discrete values will be transformed into SHs. See section 2.10.)

The maximum frequency was set to 22kHz using a step size of 100Hz. Another important setting for the very practical reason of calculation time was the number of used cores. Depending on the number of elements of the concerning mesh, the number of cores varied between 4 and 8. Further explanation can be found under 2.8. Table 1 summarizes the used settings.

Table 1: Selected settings for Blender NumCalc export and mesh2HRTF input

Name of the setting	***Selected value***
Title	Head-Related Transfer Function
Ear	left, right
Pictures	unchecked
Point Source	Irrelevant for reciprocal approach
Reciprocal	Checked
Speed of sound c	343 m/s
Density of air ⍰⍰	1.1839 kg/m^3
Unit	m
Evaluation Grid 1	Lebedev
Near field calculation	Unchecked
Frequency step	100 Hz
Maximal Frequency	22000 Hz
Frequency dependency	Unchecked
Method	BEM
CPU fist and CPU last	1
Number of used cores	4 - 8

Side note: Some changes to Blender plugin “export_mesh2hrtf.py”

Before installing the mentioned Blender plugin, a few changes were made on the file “export_mesh2hrtf.py”[15]. Firstly, several of the above shown settings were implemented as default values. Secondly, a bug that prevented the right ear from being properly calculated was fixed. Thirdly, an export of the microphone area, i.e. the size of the selected element in the ear canal was implemented by doing trigonometrical calculations for said element. This information was used later to calibrate the modeled sound pressure after BEM calculation, when the generated data is being imported into MATLAB (see section 2.9).

2.8 BEM calculations

Once input files for *mesh2HRTF* were created, BEM calculations could start on the prepared Linux machine. The first tests showed that for some meshes the calculated HRFTs had distortions in the higher frequency area starting at around 17 kHz, as can be seen in Figure 17. Sometimes these errors even distorted the rest of the spectrum in lower frequency regions. Because these distortions did not appear for all meshes some tests with modifications of an affected mesh were done to encircle the error. The changes included flattening out the nose and the contralateral ear because it was assumed that sharp edges due to re-meshing were responsible for these distortions. However, neither of these changes effected the distortions in the higher frequency area. Rectification for the affected meshes could only be achieved by neither using the multi-level fast-multipole method (ML-FMM) nor the single-level fast-multipole method (SL-FMM). Instead, the standard BEM method was selected during preparation of input files as described under 2.7.

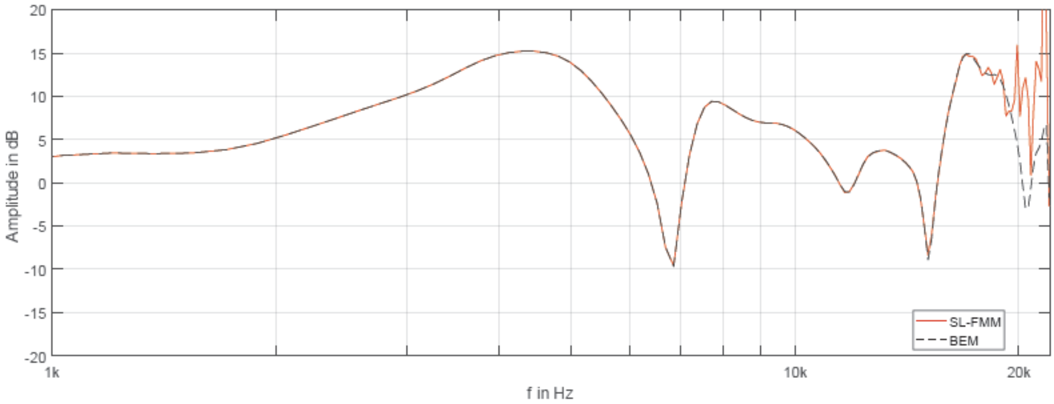


Figure 17: HRTF for subject 7 at 0° on the horizontal plane. SL-FMM shows distortions in the higher frequency area.

[15] Digital_Appendix\11 Scripts\mesh2hrtf-0.1.2\Mesh2Input

SL-FMM and ML-FMM are elegant means to save calculation resources by building clusters that combine certain mesh areas and produce results with similar accuracy with a substantially reduced memory requirement (Kreuzer et al., 2009). Unable to use SL-FMM and ML-FMM during this work had a major effect on the calculation time for the 180 different meshes (2 ears for 90 subjects that were included).

Note: The reported bug responsibly for distortions above 14kHz when using SL-FMM and ML-FMM has been fixed by Wolfgang Kreuzer in the meanwhile.

As stated before, the weakest link during BEM calculation was the available memory. A RAM size of 32 GB seems sufficient for most tasks but soon became insufficient for larger meshes. For BEM calculation the amount of memory needed was around 16.5 bytes per element. Equation 1 was used to estimate the amount of memory needed during BEM calculation which is dependent on mesh size and the number of used CPU cores n_{cores}. The number of elements of the mesh $elements_{mesh}$ appears as a squared term in the equation due to the way that BEM calculation works, using square matrix calculation as described in (Ziegelwanger et al., 2015a). A system memory usage of approximately 0.5 GB was observed. If the needed memory exceeded 32 GB for a certain mesh, the number of used CPU cores had to be reduced. This prevented memory from being outsourced to the hard disk drive which would have resulted in a drastic slowdown of the whole process.

$$needed\ memory\ [GB] = \frac{elements_{mesh}^2 \times 16.5\ bytes}{1024^3} \times n_{cores} + system\_offset$$

Equation 1: estimation of needed memory for BEM calculation based on the number of mesh elements and the number of used CPU cores

The above equation shows that meshes of up to around 16.000 could still be processed with all 8 available CPU cores. For some meshes of subjects with large heads only four CPU cores could be used for calculation. It follows that for meshes with many elements calculation time is increased in two ways: through the higher number of elements BEM calculation has to be completed for and through fewer possible parallel processes. The observed and estimated time for BEM calculation of a single element was 9.5s on a single core. Calculation time of a mesh could therefore be estimated using Equation 2.

$$calculation\ time\ [s] = \frac{elements_{mesh}\ \times\ 9.5s}{n_{cores}}$$

Equation 2: estimation of needed BEM calculation time dependent on the number of mesh elements and used cores

Around five hours were needed for BEM calculation of a single mesh, however, in some cases much longer due to the above stated reasons.

After preparation of input files, the files were loaded on the Linux machine. A bash script[16] was used to successively go through all input file folders of all meshes and simultaneously start *NumCalc* instances, i.e. *mesh2HRTF* executables for the number of assigned CPU cores.

2.9 MATLAB import

After completion of BEM calculation, the results for each mesh were imported by executing a MATLAB script called "Output2HRTF.m". An individual version of this script is automatically being created during input file preparation (see 2.7) and can be found in each folder of every calculated mesh. The script contains different information like the number of used CPU cores and the implemented microphone area, i.e. the size of the element that was chosen in the ear canal. This information is handed over to a function called "Output2HRTF_Main"[17] which then collects the computed data, i.e. the complex sound pressure that was calculated by *NumCalc* on every field point of the spherical sampling grid (evaluation grid). Through an earlier mentioned reciprocal approach, the sound pressure *p* at the blocked ear canals is obtained. The complex pressure *ps* of the point source in the origin serving as reference is calculated using Equation 3 for a monopole and using the reciprocal approach. The equation is originally defined in Equation 6.71 in (Williams, 1999) and includes the volume flow Q_s, the radius *r*, the density of air ρ_0 and the frequency *f*.

$$ps = \frac{-i\,\rho_0\,2\pi f}{4\pi}\,Q_s\,\frac{e^{i\,2\pi f\,r}}{r}$$

Equation 3: sound pressure for a monopole as in Eq. 6.71 in Williams (1999)

The volume flow is defined as a product of velocity and the area of the vibrating element (see section 6.7.11 in (Williams, 1999)). In this case the vibrating element is the size of the microphone element in the ear canal entrance.

Following the definition of HRTFs given by (Møller, 1992) as mentioned earlier, "Output2HRTF_Main" uses the complex sound pressure at the blocked ear canal *p* and the reference pressure in the origin *ps* to calculate the HRTF according to Equation 4.

$$HRTF = \frac{p}{ps}$$

Equation 4: HRTF definition as in (Møller, 1992)

[16] Digital_Appendix\11 Scripts\bash_script\run_all_cores

[17] Digital_Appendix\11 Scripts\mesh2hrtf-0.1.2\Output2HRTF\Output2HRTF_Main

The division by the reference pressure containing information about the velocity and the size of the vibrating element leads to a normalization of the HRTF to 0 dB at low frequencies, following the original definition.

Prior to use, the consideration of the actual microphone area was implemented in the "Output2HRTF_Main" function. In (Ziegelwanger et al., 2015b) the velocity had been fixed to 0.1mm/s.

Finally, 220 frequency bins (100 Hz – 22 kHz) were calculated for 1730 field points. The data was then saved in the SOFA format (AES Standards Comittee, 2015). To use the SOFA as output format, a SOFA application-programming interface[18] (API) for MATLAB is needed.

2.10 HRTF post-processing

Several signal processing steps were carried out to generate final HRIRs. For reproducibility this was done using a script[19]. Most of the signal processing steps in MATLAB were done with functions from AKtools (Brinkmann and Weinzierl, 2017). First the SOFA-file that was generated in the previous step was imported and transformed to an AKtools-common format. This provided the complex spectrum for each node and frequency bin, except for the first frequency bin at 0 Hz which was not included in the calculation. This purely real part of spectrum was added manually to be 1 (0 dB).

The first step was followed by a complex conjugation which was necessary because of an unwanted phase rotation the cause of which could not be identified.

Next, the signal is extended to a both sided spectrum and transformed back into time domain by inversing the Fast Fourier Transform (FFT).

It was stated earlier that by definition the HRTF provides information of sound pressure at the ear canals in respect to a reference in the center of the head with the head absent (Møller, 1992). This division happens in the frequency domain and hence leads to a subtraction in the time domain and the occurrence of a group delay. The signal must be circulated by a few samples to prevent some samples from being on the negative time axis. Therefore, a circular shift of 60 samples was applied.

In the next step, the impulse responses were windowed to reduce the sample size to 256 and get rid of the unnecessary endings. Also, a fade-in of 10 samples and a fade-out of 20 samples was applied.

BEM calculation using mesh2HRFT provided information at the nodes of the defined Lebedev sampling grid. Spherical harmonics transformation (SHT) could now be applied to compute the coefficients of the basis function's expansion. This expansion then allows for an interpolation of the function anywhere on the sphere (Nicol, 2010). A SHT was

[18] https://github.com/sofacoustics/API_MO (accessed June 25, 2018)

[19] Digital_Appendix\11 Scripts\Matlab\save_sh_coeffs.m

applied with prior FFT. The SHT was carried out with sampling weights as defined by (Rafaely, 2015) in Eq 3.2. The order of SHT is defined by the used sampling grid. As mentioned in 2.7 the applied Lebedev sampling grid allowed a maximal SH order of 35. Complex valued SHs where used and the SH coefficients where then saved in one file per subject containing both, left and right ear as well as the frequencies vector and some additional meta data.

2.11 Anthropometric features measurement

This works goal is to identify anthropometric key features which determine the perceptual similarity of another person's HRTF to one's own. Therefore, anthropometric features (AFs) had to be either measured or automatically extracted. Dinakaran et al. (2016) introduced an algorithm for the automatic extraction of a comprehensive set of AFs from 3D meshes. 27 AFs have been identified and defined in the widely used CIPIC HRFT database (Algazi et al., 2001b) and have been shown to be relevant for correspondences and correlations to HRTF features. Out of that list the mentioned algorithm can extract 14 features automatically. This includes 8 head and neck-elated features and 6 pinna-related features for left and right. In addition to that 5 more general AFs of the torso and the body such as *height* and *torso width* were measured manually during the procedure of 3D scanning. Figure 18 shows a few important pinna features on a 3D mesh. Figure 19 shows the original definition of pinna features by (Algazi et al., 2001b).

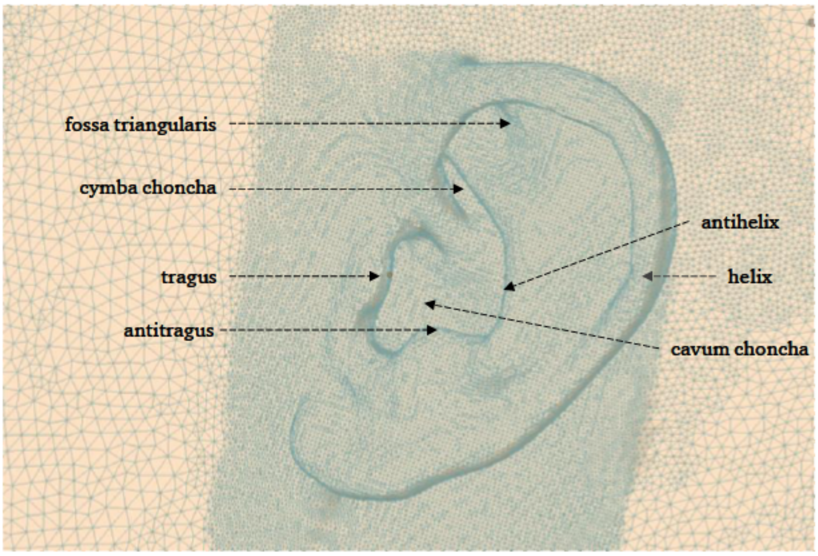


Figure 18: 3D mesh of left pinna of subject 70 with some relevant pinna features

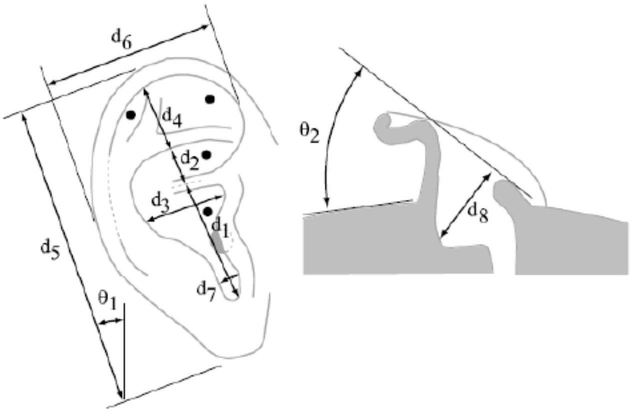


Figure 19: Definition of pinna measures, taken from Algazi et al. (2001b)

An overview of all anthropometric features that were extracted and measured and used for statistical analysis during this work can be found in Table 2. The extracted and measured AFs for all subjects can be found in the digital appendix[20]

Table 2: anthropometric features, identifiers and measurement methods

Anthropometric feature	***Identifier***	***Method of measurement***
head width	x1	Automatic feature extraction (Dinakaran et al., 2016)
head height	x2	
head depth	x3	
pinna offset down	x4	
pinna offset back	x5	
neck width	x6	
neck height	x7	
neck depth	x8	
left cavum concha height	L_d1	
left cymba concha height	L_d2	
left cavum concha width	L_d3	
left fossa height	L_d4	
left pinna height	L_d5	
left pinna width	L_d6	
right cavum concha height	R_d1	
right cymba concha height	R_d2	
right cavum concha width	R_d3	
right fossa height	R_d4	
right pinna height	R_d5	
right pinna width	R_d6	
torso top width	x9	Manual measurement
shoulder width	x12	
height	x14	
head circumference	x16	
shoulder circumference	x17	
left pinna rotation angle	L_θ_1	Semi-automatic, scripted measurement in Blender
left pinna flare angle	L_θ_2	
right pinna rotation angle	R_θ_1	
right pinna flare angle	R_θ_2	

[20] Digital_Appendix\1 Documents\Anthropometrics.xlsx

The required input for the algorithm of Dinakaran et al. (2016) are 3D meshes of the subjects including shoulders that are centered to the interaural axis. Because in this work the subjects' shoulders were unnecessary and were therefore deleted from the mesh earlier during data processing an iterative closest point (ICP) alignment of unaligned meshes with shoulders to aligned meshes without shoulder was accomplished using the alignment tool in *MeshLab*[21]*.* The point based gluing method was used in which 4 similar points on each mesh had to be selected.

ICP alignment and anthropometric feature extraction with following cross validation for evaluation using the mentioned algorithm was carried out by Manoj Dinakaran.

However, the introduced automatic extraction method did not include AFs such as pinna rotation angle θ_1 and pinna flare angle θ_2. These features were semi-automatically determined by means of a Python script, again using the *Blender* API. The script requires the manual selection of four points on the pinna to calculate rotation and flare angle. For reproducibility it is vital that the selected points are well defined.

In the already mentioned work of Algazi et al., (2001) the definition of the flare and rotation angle, i.e. the reference dimensions are somewhat vague. For example, in Figure 19 which shows the often-cited definition of AFs given by Algazi et al. the rotation angle is defined as the angle between the vertical and the pinna length *d5.* However, it is imaginable that the pinna length could also be measured differently by for example choosing the longest distance between two points at the external auricle. Another example is th flare angle which is defined as the angle between a line from tragus to antihelix with reference to a plane tangent to the head around the pinna. It stays ambiguous whether the line touches the antihelix in the same vertical position as the tragus or if the most salient point of the antihelix is being used. But when automatically or manually extracting these measures from 3D meshes a clear definition of reference points is central.

In physiology the rotation of the ear is defined as the angulation of the longitudinal axis of the external auricle (Hall, 2007). In the 3D mesh that is the angle between the z-axis (vertical axis) and the axis through the minimum and maximum point on the helix and ear lobe in z-direction as shown in Figure 20.

The flare angle θ_2 is dependent on axis that defines the the tragus-helix distance (Xie, 2013). In this work tragus-helix distance is defined as the distance between the point on the tragus with the largest absolute y-coordinate to the highest absolute y-level of the helix at the same height (z-level) as the selected point on the tragus. θ_2 then is the angle between the axis that defines the tragus-helix distance and the sagittal plane, i.e. the x-axis.

[21] http://www.meshlab.net/ (accessed June 20, 2018)

After the selection of the four points the script was executed starting the trigonometrical calculation θ_1 and θ_2. More details can be found in the script[22] itself.

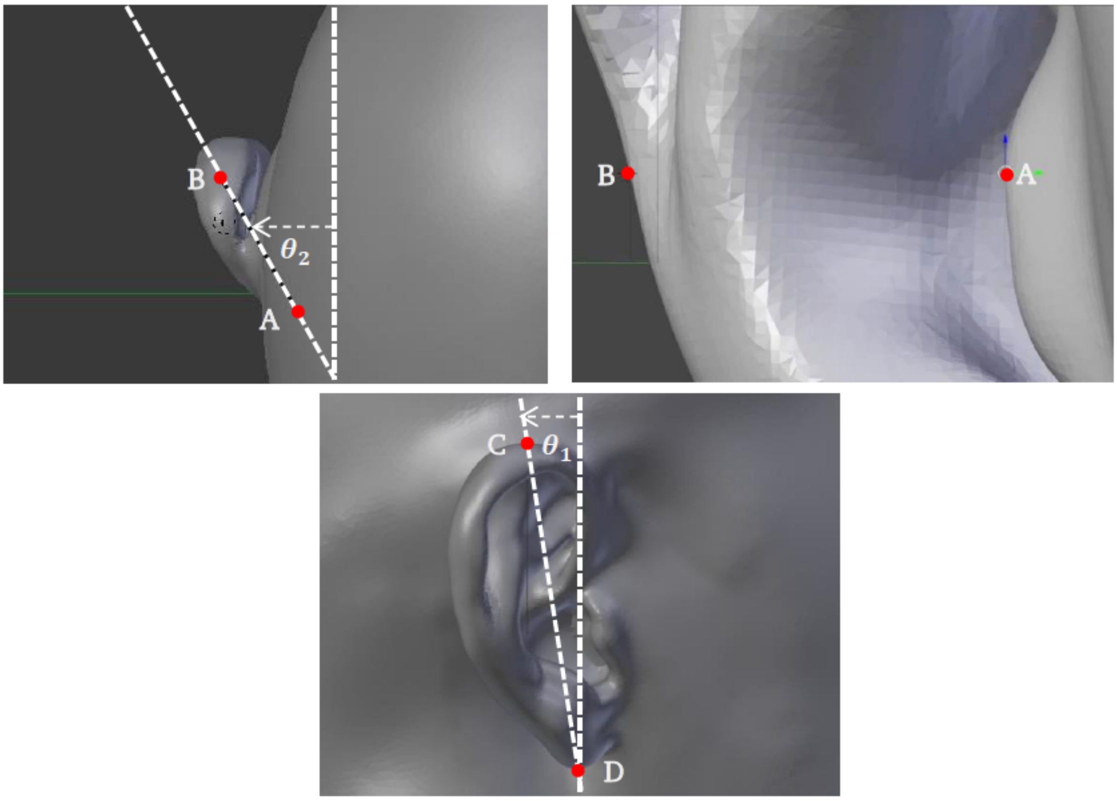


Figure 20: the drawings show the relevant points for semi-automatic measurements of pinna rotation angle θ_1 and flare angle θ_2. The pictures on the top show the selected points A and B on tragus and helix for measurement of $\theta_{2.}$ Points A and B always have the same z-coordinates. The bottom shows the longitudinal axis for determination of θ_1

[22] Digital_Appendix\11 Scripts\Python_Blender\angles.py

2.12 Listening test

2.12.1 Test design

To identify anthropometric key features that make HRTFs idiosyncratic and play a significant role in the perceptual evaluation of HRTFs, a listening test was designed in which the participants had to rate another person's HRTF against a reference, this being their own HRTF. In the first part of the listening test perceptual evaluations were carried out four times under four different aspects for each HRTF. Four SAQI items from the Spatial Audio Quality Inventory (Lindau et al., 2014) were chosen because these qualities are assumed to be easily identified for most binaural applications. The listening test used dynamic binaural synthesis and was therefore conducted with a Polhemus patriot head tracker, allowing head movement of the participants to increase localization performance (Møller, 1992). The head tracker was positioned on top of the Sennheiser HD 800S headphone which was used during the listening test.

Each of the 43 participants rated four different SAQI qualities separately for 15 different HRTFs. Condition variation was included by two different source positions. Two different positions were used to create a measuring repetition for each rating with a slight variation. In the first case the sound source was coming from a frontal position with no elevation. In the shifted case the sound source was elevated for 15 degrees and shifted to the left for 30 degrees.

Also, each subject rated the 15 HRTFs for both, the measured and the modeled HRTF data set. In the first part of the listening test this lead to 240 ratings each participant had to carry out. The definition of the used SAQI items can be found in the digital appendix[23].

The participants had access to the SAQI definition prior to the listening test and the definitions were again shown before entering a certain SAQI section. *Difference* and *Coloration* were changed slightly from the original definition to fit the purpose of this work.

The used stimuli shown in Table 3 were selected because they were assumed to be the most appropriate to evaluate corresponding SAQI features. For *Difference* and *Externalization,* a short reverberation free speech stimulus[24] was used which contained an 8s long, German sentence taken from a poem and spoken by a female speaker. Speech contains fine nuances and has more dynamics than noise and therefore seem to be adequate to test the said SAQI items. Pink noise which contains all frequencies of the audible spectrum was chosen to evaluate changes in *Coloration.* For *Source position* a pulsed pink noise stimulus was chosen. The pules were 0.5 sec long with 0.3 sec pauses and 0.02 sec fade-in and fade-outs similar to the localization experiment of (Majdak et al., 2010). The pink noise stimuli were both 5 sec long and all stimuli were played in a loop.

[23] Digital_Appendix\8 Listening_Test\Documents\SAQI_Features.pdf

[24] Digital_Appendix\8 Listening_Test\audio

Table 3: SAQI items for part 1 of the listening test

SAQI item	***definition***	***scale***	***stimulus***
Difference	Existence of a noticeable difference.	unipolar (3 = high difference or change, 0 = no difference)	speech
Coloration	Existence of any coloration or difference in timbre.		pink noise
Source position	Existence of a source position change		pulsed pink noise
Externalization	Describes the distinctness with which a sound source is perceived within or outside the head regardless of their distance.	bipolar (3 = more externalized, -3 more internalized)	speech

For the first part of the listening test the order of the listening test was designed to be random with certain group consistencies. This means, that first a SAQI item was selected randomly, for which all four groups of condition variations were rated *(measured_front, measured_side, modeled_front, modeled_side).* These groups were also selected randomly in order and each group contained the specific 15 HRTFs the participant was supposed to rate. Whisper (Ciba et al., 2009), an open source tool for performing listening tests was used in this work. The 15 HRTFs were presented on two consecutive pages, the first containing 8 HRTFs and the second page containing 7. The HRFTs were distributed randomly across the pages. Figure 21 shows a rating page in which the A-, B- and Stop-buttons can be seen. The buttons played the stimuli with different HRTFs, while A was always the reference.

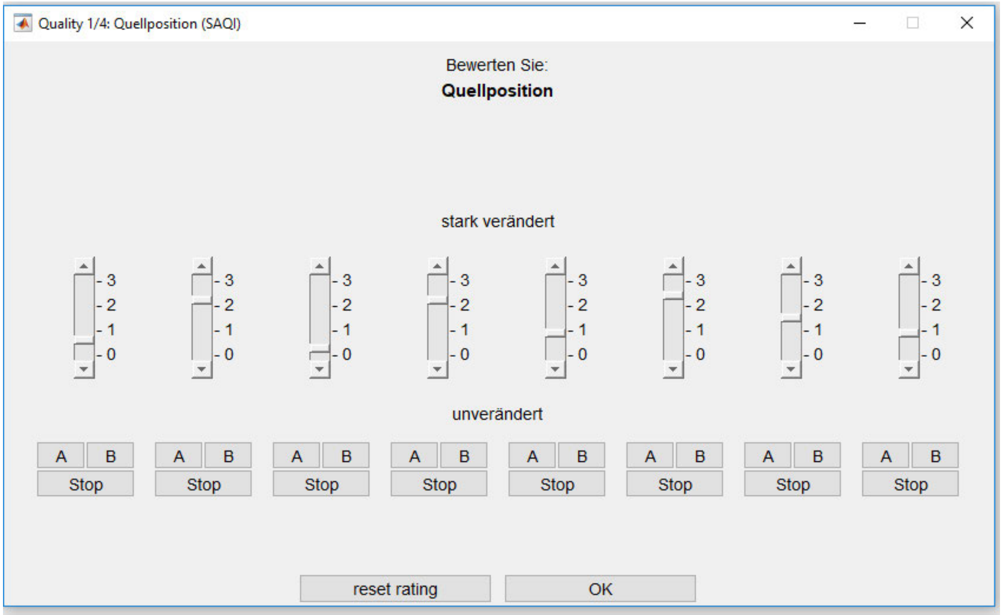


Figure 21: Whisper rating page for SAQI item *source position* used in the listening test

In the second part of the listening test, the similarity test, each subject rated their own measured HRTF against their own modeled HRTF in an A-B comparison for 26 SAQI items. For his part all audio qualities from the inventory were selected excluding the ones which refer to reverberation or did not apply for other reasons. The qualities were examined using different stimuli. A pulsed pink noise was again chosen for directional or localization-related qualities. Table 4 contains a full list of the used qualities for this section. A full definition of the SAQI items can be found in (Lindau et al., 2014).

Table 4: SAQI items, used stimuli and scales which were included in the second part of the listening test

SAQI item (perceptual quality)	***Scale***	***Stimulus***
Difference	Unipolar (3/0)	Speech
Tone color bright-dark	Bipolar (3/-3)	Pink noise
High-frequency tone color		
Mid-frequency tone color		
Low-frequency tone color		
Sharpness		
Comb filter coloration		
Metallic tone color		
Tonalness		
Source expansion		
Loudness		
Horizontal direction	Manual entry of angles	Pulsed pink noise
Vertical direction		
Front-back position	Dichotomous (confused / not confused)	
Width	Bipolar (3/-3)	
Height		
Localizability		
Distance		Speech
Externalization		
Spatial disintegration		
Speech intelligibility		
Naturalness		
Degree-of-Liking		
Crispness		Drums
Dynamic range		
Dynamic compression effects		

2.12.2 Head related impulse responses

As mentioned above, the listening test was carried out using dynamic binaural synthesis to present the played stimuli in a real-time spatial audio environment. Open-source software *SoundScape Renderer* (Geier et al., 2008) was used for that purpose. The HRTFs had to be transformed to a spatial audio file format, the already mentioned SOFA format which could then be loaded into the custom version of the *SoundScape Renderer* (SSR).

Converting HRTFs into a spatial audio format means that they are being transformed back into time domain. To generate these head-related impulse responses (HRIR) an inverse SHT was carried out. To generate smaller sized files for the SSR and save processing time, SOFA files were created only for directions necessary for the listening test. The azimuth ranged from -42° to 42°, while 0° represents the front. Elevation ranged from -16° to 16° with 0° representing the horizontal plane. The range for azimuth and elevation was chosen in respect to the highest values occurring during sound localization (Thurlow et al., 1967). This way extensive head movement was possible but unnecessary directions were not included.

As mentioned earlier, two versions were included in the listening test for every HRIR, one with a source coming from the front and one in which the source was shifted 30° to the left and 15° up.

It should also be noted that no room impulse response was added to the HRIRs. Leaving aside room reflections created a somewhat unnatural result, but the goal of this work was to identify HRTF-related perceptual differences without any effects stemming from room reflections. The use of HRTFs without room reflections allows a more universal examination of perceptual differences.

Finally, a filter was applied only to measured HRIRs to equalize the frequency roll-off caused by SH order truncation due to the limited number of sensors on the microphone array. Ben-Hur et al. (2017) introduced a digital filter which equalizes the frequency spectrum of a low spatial order signal. In numerical BEM simulation a much higher spatial resolution can be achieved. The coarser spatial resolution of the microphone array results in attenuation of higher frequencies during SHT which is why an equalization of SH order 15 to order 35 is applied following equation 12 in (Ben-Hur et al., 2017). Equalization is applied for averaged results for all directions simultaneously. The effect of this equalization filter can be seen in Figure 22.

For the above-mentioned procedure of creating HRIRs from HRTFs a MATLAB script[25] was used for reproducibility.

[25] Digital_Appendix\11 Scripts\Matlab\create_HRIRs.m

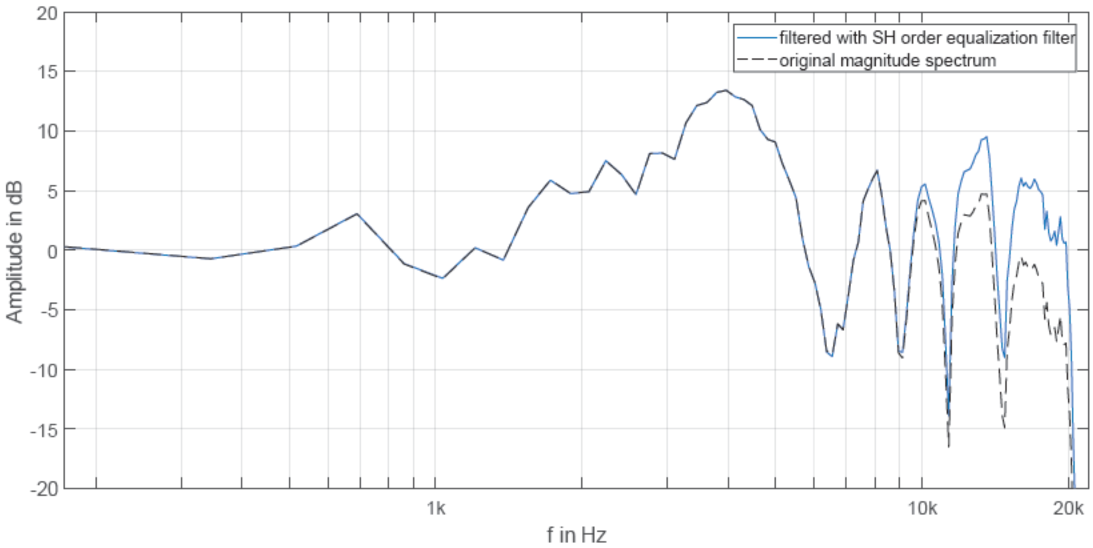


Figure 22: Equalization filter for SH order truncation. Original and filtered HRTF from subject 22 in the frontal position.

2.12.3 Headphone filters

During initial acoustic HRFT measurements also headphone transfer functions (HpTF) were measured by placing microphones in the subjects' ear canals who wore headphones (Sennheiser HD800S). A frequency sweep was used to generate impulse responses which are both, headphone and pinna-related. Dynamic binaural synthesis requires the equalization of these HpTFs to negate the headphone's and the pinna's one-directional influences and minimize timbre colorations induced by spectral cues (Nicol, 2010).

Headphone filters were designed for each subject to inverse the influence of HpTFs for both, left and right ear using an AKtools script[26]. The script allows a regularized compensation as introduced in (Lindau and Brinkmann, 2012). Individual compensation is attained by creating filters which inverse HpTFs, however leaving out pinna notches which typically occur above 5 kHz. The regularization of compensation filters happens manually by adding 1 to 3 parametric, peaking notch filters (PEQ) to defined frequency regions in which the inversion efforts are being limited. This preservation of notches is perceptually advantageous (Lindau and Brinkmann, 2012).

After identification of regions to apply parametric peaking notch filters for regularized compensation, the created filter was applied to the HpTF and an auditory filter bank of equivalent rectangular bandwidth (ERB) filters was used to analyze the perceptual deviation between compensated HpTFs and a target function (a bandpass as seen in Figure 24). The goal was to approximate the target function on one hand while limiting ERB errors to a maximum gain of 2 dB on the other hand (ERB error attenuation is not as perceptually critical as ERB error gain which can cause ringing artefacts). In some cases,

[26] Digital_Appendix\7 Headphone\Headphone filters_HD800S\AKregulatedInversionRP.m

these criteria could be fulfilled easily while in others a readjustment of PEQs was necessary to get closer to the desired outcome.

Figure 23 shows the input channel of averaged data, the filter and the regularization channel. Figure 25 shows the resulting ERG.

After the HpTFs were compensated, SOFA files were written as input files for the SSR for the left and the right ear.

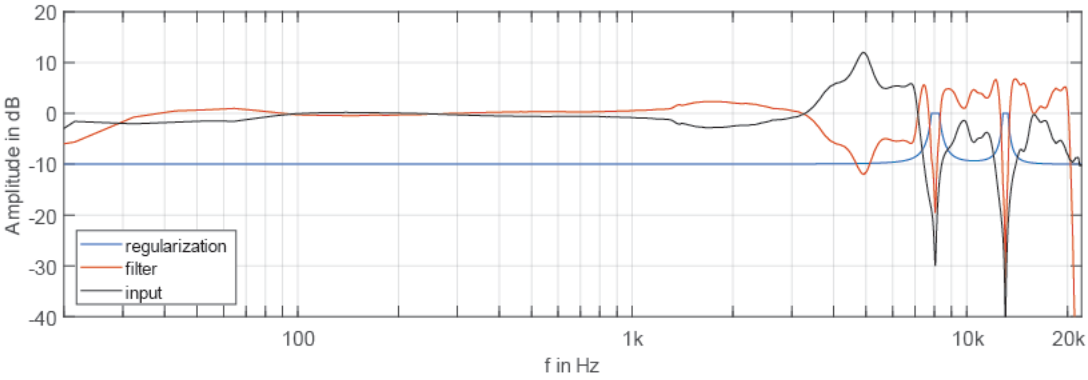


Figure 23: input, filter and regularization channel; channel 2 of subject 40

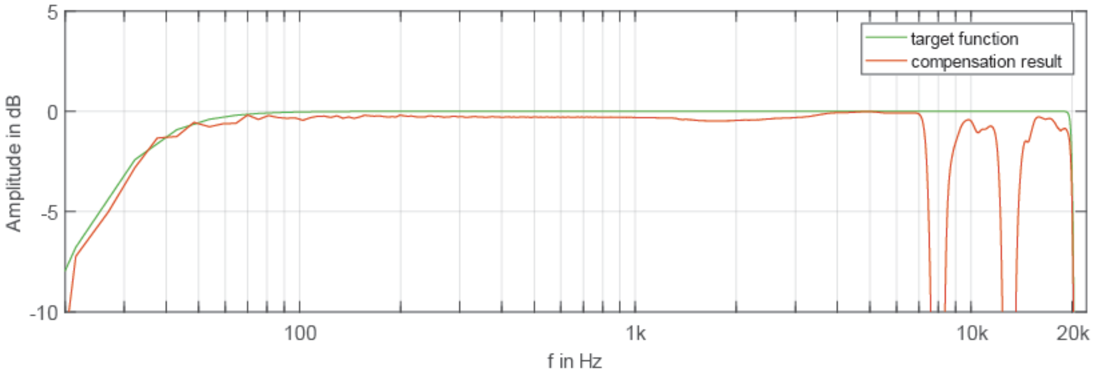


Figure 24: target function and compensation result; channel 2 of subject 40

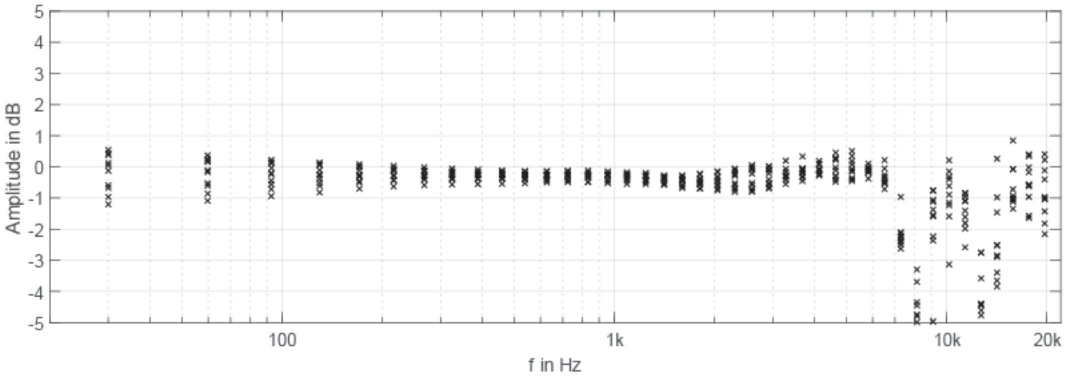


Figure 25: ERB error – perceptual deviation of compensated HpTF to target function; channel 2 of subject 40

2.12.4 Listening test instructions

Starting with the listening test the subjects filled out a form[27] with some basic information including their general hearing ability, health condition and listening test experience. The subjects also signed an agreement which stated that they had been sufficiently informed about the test, agreeing to take part in it and agreeing to the anonymous publication of results. After the test they could also leave comments on the fill out.

Next, they were informed about the two separate sections of the listening test. The head tracker was then calibrated instructing the subjects to take a natural position looking straight forward.

The subjects were informed that a various number of stimuli were being presented to them and that they should rate them only in respect to the current SAQI quality. They were told to always rate stimulus B in respect to stimulus A. Stimulus A was always their own HRTF, however, which was not communicated. The subjects were then instructed to create a ranking while evaluating the stimuli on each page, the highest perceived differences being rated the highest and vice versa (as can be seen in Figure 21). They were told that they could change and adjust ratings on each page as much as needed, always being able to listen to all stimuli repeatedly to create a logical ranking order. To familiarize the subjects with the evaluation process, a short training was added prior to the listening test presenting one page of stimuli which had to be rated and ranked.

The subjects were then told about the second part of the listening test in which they would rate a single stimulus B to a reference A in respect to various perceptual qualities, however not knowing what exactly they rated. During this part, stimulus A was always the subject's measured HRTF. In a short training the subjects could familiarize themselves with all occurring scale types. Finally, remaining questions were answered, and the listening test started.

2.12.5 Listening test setup

Two instances of SSRs were running, one with the subject's headphone filter correction HRIR and another one containing all 15 HRIRs that the current subject should evaluate. The SSR loaded 64 HRIR SOFA-files simultaneously (four condition variations for each of the 15 HRTFs in addition to the subject's own HRTF also in four condition variations). This way, no adjustments or changes on the SSR were necessary during the listening test.

The subjects navigated through the listening test on a laptop with *Whisper* which also saved the evaluation results. When pressing buttons "A", "B" and "Stop" *Open Sound Control* (OSC) signals were sent to the Linux machine on which the SSR instances were running.

[27] Digital_Appendix\8 Listening_Test\Documents\Versuchsunterlagen_HRTF.docx

Pure Data (PD)[28], an open source visual programming language was also running on the Linux machine and received the OSC signals sent from *Whisper*. The OSC signals included information about the current SAQI quality, play or stop commands, the played HRTF number, the position of the sound source (front or side) and the data set type (measured and modeled). A PD patch[29] was created to process this information and select stimuli according to the current SAQI feature which were then routed to one of the 64 inputs of the SSR. Figure 26 shows a schematic illustrating the basic setup of the listening test.

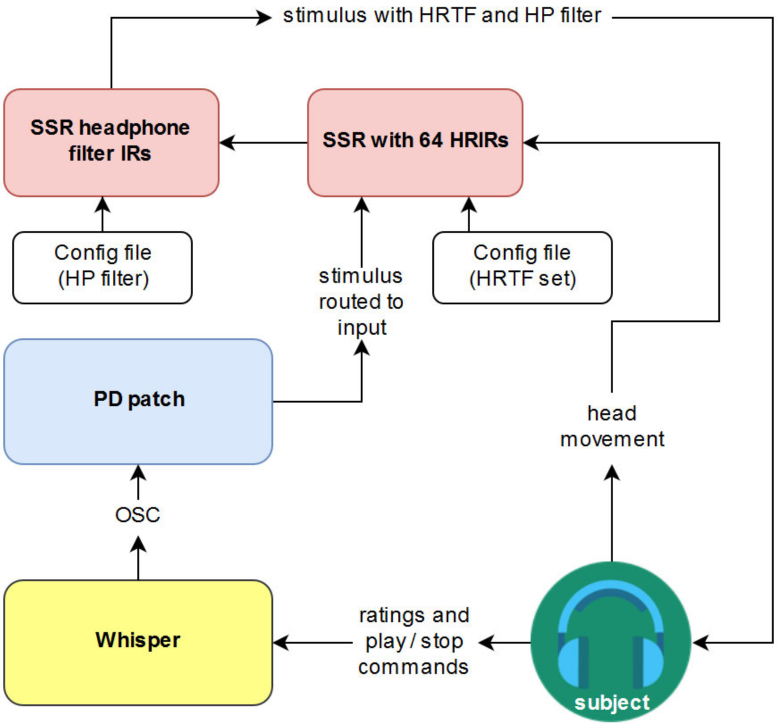


Figure 26: schematic of listening test setup

JACK audio connection kit[30] was used to route audio between different programs on the Linux machine. PD had 64 outputs which were connected to the SSR instance with 64 HRIRs. The stimulus was routed only to the one HRTF which was currently played. The output of this SSR instance was connected to the second SSR with headphone filters, the outputs of which were connected to the main audio output.

The audio output of the Linux machine sent the audio signal to a Lake People Phone-Amp G109 pre-amplifier to which the HD 800S headphones were connected. Prior to the listening test the volume was manually adjusted on the pre-amp and in the PD patch for each stimulus separately.

[28] https://puredata.info/ (accessed June 28, 2018)

[29] Digital_Appendix\8 Listening_Test\hrtf_indi.pd

[30] http://jackaudio.org/ (accessed June 28, 2018)

The head-tracker which provided the SSR with head movement information was mounted to the headphone and was connected to the Linux machine via a serial/parallel USB adapter.

2.12.6 Configuration files

Out of the whole data set 90 HRTFs were selected for the listening test excluding a few cases in which acoustical measurements were faulty. To ensure that each of the 90 HRTFs would be rated equally often, conditions were created using a MATLAB script[31] which created 60 unique conditions, each containing 15 HRTFs to be rated (the best-case scenario of participants joining the listening test was assumed to be 60). For each condition several "allowed" subject IDs existed, namely every subject ID not included in the condition. This prevented subjects from rating their own HRTF.

Configuration files[32] for the SSR were created in asd-format from these 60 conditions. Each configuration file contained the set of 15 HRTF to rate in all four condition variations including the subjects HRTFs as a reference. Because the order in which subjects would take part in the listening test was unknown, configuration files for each condition had to be created for all "allowed" subjects. Configuration files for the SSR included name and path of the HRIRs to load. Also, the SOFA-files range (azimuth_range= 318 to 42; elevation_range= -16 to 16) was included in the configuration file to prevent a source position bug from occurring in the used custom SSR version. Additionally, configuration files[33] for the headphone filter SSR instance were created to load the matching filters for each subject.

A fitting condition had to be chosen for each subject at the beginning of the listening test. The setup on the Linux machine could then be started by running a bash-script[34] and entering subject ID and condition number. The bash script opened both SSR instances with said configuration files, the PD patch and the JACK client with predefined connections[35].

[31] Digital_Appendix\11 Scripts\Matlab\Conditions\generateConditions.m

[32] Digital_Appendix\8 Listening_Test\configs\HRIR_configs

[33] Digital_Appendix\8 Listening_Test\configs\HPfilter_configs

[34] Digital_Appendix\8 Listening_Test\Versuch_Start.sh

[35] Digital_Appendix\8 Listening_Test\configs\jackConnections.xml

2.13 Statistical analysis

2.13.1 Cross-validation of measured and modeled HRTFs

A cross-validation of measured and modeled HRTFs was carried out similar to (Brinkmann et al., 2017) to analyze resulting differences between the different methods. Cross-validation between measured and modeled HRTFs was suggested by Turku et al. (2008) due to a lack of true external reference. In other words, only relative differences can be high lightened because it stays uncertain, in which direction to interpret differences.

Cross validation was done by inverse SHT, creating HRIRs for a sampling grid (with full range for the azimuth angle, elevation angle ranging from -30° to 90° and with 5° of resolution). The previously mentioned SH order equalization filtering which eliminates frequency roll-offs of measured HRTFs was also applied. ERB errors were then calculated between modeled and measured HRTFs and averaged across ears and all directions or ears and frequencies. Also, a pooled representation of ERB errors of all subjects was done.

Additionally, the approach of Baumgartner et al. (2014) to model sound-source localization in sagittal planes was used to compute localization errors. Baumgartner's model uses HRTFs and is based on the comparison of internal sound representation with a template. Monaural and binaural perceptual factors are considered, and the model finally results in a probabilistic prediction of polar angle responses. The model also calculates polar root-mean-spare errors (PE) and the quadrant error rate (QE) as defined by (Middlebrooks, 1999b). The PE is defined as the root-mean-square average of all polar errors (localization errors on the median plane) that are less than 90° in magnitude. Polar errors greater than 90° are scored as quadrant errors, because quadrant confusion on the median plane can occur in that case. The quadrant error rate QE is therefore the percentage in which quadrant errors occur. The model is used via a MATLAB function (baumgartner2014.m)[36].

The cross-validation calculation was done by Fabian Brinkmann using MATLAB scripts[37].

2.13.2 Comparison of rating means with t-tests

During the first part of the listening test, participants rated a set of 15 HRTF against their own HRTF in both versions, measured and modeled. The rating means for measured and for modeled HRTFs were analyzed using a paired sample t-test for each of the four

[36] http://amtoolbox.sourceforge.net/amt-0.9.9/doc/models/baumgartner2014.php (accessed May 30, 2018)

[37] Digital_Appendix\6 crossValidation\a_crossValidationAllSubjects.m, b_crossValidationAllSubjects.m

SAQI items. The assumption of normality of differences between the scores (i.e. the ratings) was tested and confirmed. Effect sizes *r* for *t*-statistics *t* were calculated with degrees of freedom *df* according to Equation 5 (Field, 2009).

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

Equation 5: effect size *r* for *t*-statistic (Field, 2009)

2.13.3 Correlation of Anthropometric Features

For later analysis and identification of key predictors it is also helpful to know whether and how AFs are correlated. That a variety of AFs are correlating suggests itself because it seems logical that a large *pinna width* also means that there is a large *cavum concha width*. A correlation was therefore calculated using SPSS between all AFs. Left and right values of AFs were combined into one mean value. Pearson's r and the one-tailed significance level for all AFs is reported in Table 18.

2.13.4 Parameter tuning and interactions

The following two sections explain how regression models were built to predict listening test results with AFs. Section 2.13.4 explains how the input parameters, namely the AFs were tuned and transformed to create the best results using multiple regression. In section 2.13.5 regression models are extended by random effects to create multilevel mixed-effects models. Section 2.13.6 covers the estimation of effect sizes of the independent variables (AFs) in the created multilevel models.

Statistical analysis was done separately for the two different parts of listening test results. In the first part, each participant rated four SAQI features, comparing his or her own HRTF against 15 different HRTFs for two directions and in both modes, modeled and measured, as was mentioned earlier. The ratings of all subjects were collected using a MATLAB script[38] and transformed into a *long* data structure[39] which then held all the ratings, each line representing a single data point.

Analysis of the data was carried out regarding the focus of this work, which was to find a relationship of the listeners' AFs to the perceived quality of another person's HRTF. It therefore suggests itself to somehow create a relationship between the listener's anthropometric data (listener's AF) and the anthropometrics of the person who's HRTF was rated (owner's AF). By doing this, a set of independent variables (IV) is created that can be used

[38] Digital_Appendix\11 Scripts\Matlab\get_ratings_long_interaction.m

[39] Digital_Appendix\9 Listening_Test_Data\Main_Test_LongFormat

and tested as predictors in further analysis. In a first step it was therefore tested which kind of relationship between the listener's AF and the owner's AF is most beneficial. Each two corresponding AFs were transformed in several ways, including ratio, difference, a cubic or a logarithmic relationship as well as some other relationships. A full list of tested transformations can be found in Table 6. A multivariate linear regression was performed using SPSS to find the relationship of AF pairs yielding the highest R^2, i.e. the set of IVs explaining the most variance of the given ratings. Furthermore, multivariate regression was done twice, once with all predictors entered and once with a stepwise backwards method. The stepwise backwards method excludes certain predictors automatically to test the contribution of each predictor by looking at the significance value of the t-test for each one (Field, 2009). Table 6 under 3.2 contains a full list of every tested anthropometric pair relationship with corresponding R^2 values for all SAQI qualities. The simple difference between anthropometric feature pairs (*diff-features*) yielded the best overall results and was chosen as the transformation mean to create a relationship between the listener's AF and the owner's AF.

As can be seen in Table 6, the coefficient of determination R^2 is still rather low, which means that so far, a small part of the existing variance can be explained by the included predictors. In a next step interactions and squared differences were introduced to the model. This was done, because it was assumed that AFs might not only have an effect on their own, but a combination of certain variables might even be more influential. Additionally, squared diff-features were introduced to the model to add non-linear effects. Before diff-features were transformed to squares and interactions were added, diff-features were grand mean centered by subtracting the mean of all scores from every single score for each variable separately. This way the estimation precision of the model increases (Field, 2009).

By adding squared diff-features and interactions to the model, the number of predictors increases to a great extent. Now, instead of the included predictors consisting of 29 anthropometric diff-features alone the new model could be calculated with 1276 predictors. It should be noted that of course double or triple interactions of AFs could have been included as well as higher polynomial orders. However, the limiting factor was the number of available data and the resulting degrees of freedom, which will be discussed further under 4.3.

The regression was then carried out for each SAQI feature in a partly stepwise manner with a block wise entry using SPSS. The first block consisted of anthropometric diff-features and was added to the model using forced entry (known as Enter in SPSS). This way all main effects were forced into the model. All following predictors were added stepwise forward. Table 5 shows a brief overview of the order of the added predictor blocks and the method of predictors and interactions that were added into the regression model.

Table 5: order and methods of entry of predictors and interactions into the regression model

Entry position	*Predictors*	*Entry method*
1	$diff\_features$	Forced entry
2	$diff\_features^2$	Stepwise forward
3	$diff\_features \times diff\_features$	
4	$diff\_features \times diff\_features^2$	
5	$diff\_features^2 \times diff\_features^2$	

Regressions where then calculated for each SAQI feature and for both modes, modeled and measured.

In this work the MSE was calculated using the residual sum of squares (SSR) divided by the number of observations n as shown in Equation 6. SSR is calculated using the predicted values y_{pred} and the true determined response values y_{true} (Fahrmeir et al., 2016) as and can be found in the numerator of Equation 6. Often in regression the divisor for SSR are the degrees of freedom, however this work uses a more basic approach for later comparison with other models and machine learning approaches.

$$MSE = \frac{\sum(y_{pred} - y_{true})^2}{n}$$

Equation 6: definition of mean square error MSE

2.13.5 Multilevel mixed-effects models

In the listening test the subjects had to evaluate SAQI features like *Coloration*. However, no anchor was included in the test, defining what a weak or a strong *Coloration* is. Hence, ratings can be highly subjective. It was therefore assumed that that data has a hierarchical structure (multilevel structure) and consists of nested data, each participant's rating representing one level (Hox et al., 2010). A random intercept was added to the model to account for subjective rating differences of each subject. This lead to a multilevel mixed-effects model (mixed model) containing both, random effects and fixed effects (consisting of predictors which remained in regression model during the previous step).

The mixed model was calculated in MATLAB. For each model the excel file[40] containing regression coefficients which were calculated during the previous step was loaded to select a list of predictors. First a model without random intercept was calculated. Next, a mixed model with random intercept was calculated using the setting "maximum likelihood" to estimate the model parameters which also enables a later comparison of models

[40] Digital_Appendix\10 Statistical_Models\Regression_Models

(Peugh, 2010). Random effects where assumed to be independent, so the covariance pattern was set to “full”, which corresponds to the standard setting “VC” in SPSS (Field, 2009).

A chi-square difference test (χ^2_{Diff}), also called likelihood ratio test using the deviance values of two models (-2 times the log likelihood) was performed to test the influence of the added random intercept (Peugh, 2010). The full model in Equation 7 means the model with more predictors, which in the mentioned case would be the model with added random intercept.

$$\chi^2_{diff} = (-2logL_{ReducedModel}) - (-2logL_{FullModel}) = deviance_{ReducedModel} - deviance_{FullModel}$$

Equation 7: Deviance difference for likelihood ratio test as in (Peugh, 2010)

χ^2_{Diff}, i.e. the deviance difference could then be referenced to a chi-square distribution test with the corresponding change in degrees of freedom (Peugh, 2010). When comparing two similar models, one with and one without random intercept the change in degrees of freedom between them is 1. The chi-square test (one-sided) then calculated a p-value for $\chi^2_{\text{Diff}}(1)$. When significant ($p<.05$), the added predictor or predictors enhanced the model.

Next, non-significant predictors ($p >.05$) in the mixed model were identified and removed from the model which was then calculated again creating a reduced model. However, this was only carried out for interaction terms because main effects are ought to remain in the model. Another likelihood ratio test tested if the reduced models with viewer predictors yielded a significant enhancement ($p <.05$). Secondly, the Akaike information criterion (AIC) which is another model estimator was used to compare the reduced to the full model. The AIC value is expected to be smaller in a better model (Field, 2009). The reduced model was only chosen as the final model when both mentioned criteria where fulfilled: a significant result of the likelihood ratio test and a lower AIC value. Usually, the reduced model again contained some non-significant predictors. An even further reduction did not yield enhanced models in terms of the above stated criteria, so one reduction was carried out at the most.

In statistics the coefficient of determination R^2 is used to state information about how much of the total existing variance can be explained by a model, so R^2 is the proportion of variance in the dependent variable which can be explained by the independent variables (Field, 2009). In linear regression, the calculation of R^2 is straight forward whereas with multilevel mixed-effects models it seems to be a challenge to report the amount of variance explained in a consistent and comparable way (LaHuis et al., 2014). This work uses the approach provided by Nakagawa and Schielzeth (2013) in which the goodness-of-fit is expressed by two different values: $R^2_{\text{conditional}}$ and R^2_{marginal}. In this approach $R^2_{\text{conditional}}$ expresses the variance explained by the whole model, including fixed and random effects while R^2_{marginal} expresses the variance explained by fixed effects only. Conditional and marginal R^2 are calculated as shown in Equation 8 and Equation 9 as defined in (Nakagawa and Schielzeth, 2013) for a linear multilevel model with one random factor. For the calculation of both versions of R^2 the variance of the random effect σ_r^2, the residual variance σ_ε^2 and the variance of fixed effects σ_f^2 are needed. The estimation of σ_f^2 can be achieved by calculating a vector of fitted values by fixed effects alone which leads to the same result as multiplying the design matrix of the fixed effects with the vector of fixed effect estimates. σ_f^2 then is the calculated variance of this vector of fitted estimates (Nakagawa and Schielzeth, 2013).

$$R^2_{marginal} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_r^2 + \sigma_\varepsilon^2}$$

Equation 8: marginal R^2 representing the variance explained by fixed effects only

$$R^2_{conditional} = \frac{\sigma_f^2 + \sigma_r^2}{\sigma_f^2 + \sigma_r^2 + \sigma_\varepsilon^2}$$

Equation 9: conditional R^2 representing the variance explained by fixed and random effects

The MSE is calculated as stated under 2.13.4 and a comprehensive summary of model criteria for all calculated models is given in Table 9 under 3.5.1.

Lastly the model with was re-calculated with uncentered IVs to create a final model with correct estimates, i.e. regression coefficients. The final models can be found in the digital appendix[41].

2.13.6 Effects sizes

Because it is this work's goal to not only develop means to predict or recommend a certain degree of fit when choosing HRTFs from a database based on AFs but also to identify key features, effect strengths of the main effects (mere AFs, without interactions) must be analyzed. Effect strengths of coefficients in linear regression are delivered by standardized coefficients (betas). However, in multilevel mixed-effects multilevel models no standardized coefficients are reported which leaves the option of calculating the model on manual z-standardized IVs. This way, all predictors in the model are supplied with an estimate which is equivalent to their effect strength. However, in final models a lot of interactions can be found which hinders a direct evaluation of the strength of main effects. Therefore, another approach was chosen to estimate effect sizes in the form of predictor

[41] Digital_Appendix\10 Statistical_Models\Mixed_Models

importance. For each main effect all its occurrences were removed from the model but interactions with other predictors remained and R^2 was calculated again for the model without a certain effect. When for example the effect size of the cavum concha height (d1) was estimated, left and right occurrences (*L_d1, R_d1*) were deleted as well as any squared values. Also, interactions with self (such as L_d1 with squared R_d1) were removed. Next, the calculated R^2 without a certain main effect was subtracted from the R^2 of the full model and the difference was attributed to the corresponding effect. Finally, all remaining interactions were removed one by one and the established R^2 difference was divided up and added to both main effects contributing to the interaction. Conditional R^2 was used for this analysis. This way, the importance of predictors could be evaluated by analyzing how much the R^2 of the model increases (R^2_{increase}) when a certain main effect is added to the model with all other effects already present.

Multilevel mixed-effect models and the predictor importance (R^2_{increase}) of main effects were calculated using a script[42] for reproducibility.

2.13.7 Support vector machine regression

The above described statistical approaches aim to create models explaining a maximum amount of variance and help isolate key predictors. If the focus was creating a recommender model which selects the best fitting HRTF from a dataset based on AFs, a measure of choice could also be using a machine learning approach. A person's AFs could be measured and set into relationship to all AFs in the data set by simply calculating the difference as described under 2.13.4. A recommender system could then be used to predict perceived *Difference* or *Coloration* for each HRTF in the data set and the HRTF with the minimal predicted rating could be chosen as a best fit. In this work a support vector machine regression (SVM-R) was used to build recommender models using a MATLAB script[43].

Machine learning tools such as SVM-R can fit a dependent variable to predictor variables non-linearly, which is advantageous in the case of interactions and non-linear relationships. However, to test how the model performs on "unknown" data, the first step was to split the dataset into a training and a testing set, while 75 % was used for training and 25 % for testing. This training and testing set ratio is the default setting in Python's *scikit-learn* machine learning library[44]. A script[45] was used to be able to create the same training set again later. Using SVM-R this step is vital, because very low values for the *MSE* can be achieved, which, however might be due to overfitting.

[42] Digital_Appendix\11 Scripts\Matlab\Mixed.m

[43] Digital_Appendix\11 Scripts\SVM_regression.m

[44] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html (accessed June 29, 2018)

[45] Digital_Appendix\11 Scripts\Matlab\testing_training.m

To build an SVM-R model, several hyperparameters can be varied, namely *kernel function*, *kernel scale* (called *sigma* in *scikit-learn*) the *box constraint* (called *C* in *scikit-learn*), and *epsilon.* The *kernel functions* that were tested were a gaussian-kernel, a rbf-kernel (radial basis function kernel) and a polynomial kernel. The polynomial kernel was also tested with different polynomial orders. To test various settings of hyperparameters, k-fold cross-validation was used. With *k= 10,* the data set is split into 10 parts randomly while 9 parts are being individually fitted to one remaining validation part. Then a different one of the 10 parts is assigned as the validation set and the process is repeated until each part has been used as the validation part once. A prediction error is estimated for each fit and the prediction errors are combined to a final cross-validation error (Hastie et al., 2009).

In this work, to minimize the cross-validation error instead of a grid search a Bayesian Optimization Algorithm is used in MATLAB to search for optimal hyperparameters by running through a maximum amount of 500 different settings. Prior to this optimization different *kernel functions* were tested. In all cases, the *gaussian* kernel yielded the best results, which is why further optimization was done with a *gaussian* kernel.

The final model was then tested using held-out data, i.e. the training set. The coefficient of determination R^2 was calculated using a standard approach for linear regression models. These values must be handled with caution because, as previously stated with reference to multilevel mixed-effects models (see section 2.13.5), R^2 is defined as the ratio of variance explained by the model to the total amount of variance. This definition led to an individual adaption of R^2 for mixed models and an adaption might as well be needed in respect to SVM-R models. When comparing different approaches such as mixed models with hierarchical structure and SVM-R models, it might be best to focus on the MSE instead. However, it seems that the standard definition of R^2 for linear regression models is also being used in standard machine learning libraries (see the definition for R^2 called "score" in the scikit-learn library[46] for Python). The definition of R^2 can be found in various statistical textbooks, such as (Field, 2009) and uses the residual sum of squares (SSR) and the total sum of squares (SST).

The calculation of SST includes the *true_mean* which is the mean of all response values y_{true}. SST is then defined as Equation 10.

$$SST = \sum(y_{true} - true\_mean)^2$$

Equation 10: total sum of squares SST

[46] http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html (accessed June 29, 2018)

R^2 is then calculated as shown in Equation 11.

$$R^2 = 1 - \frac{SSR}{SST}$$

Equation 11: coefficient of determination R^2

An SVM-R model was calculated for each SAQI item and each dataset. A comprehensive summary of calculated values for R^2s and MSEs with corresponding kernel parameters is given in Table 17 under 3.13.

2.13.8 Descriptive statistical analysis of the similarity test

The second part of the listening test consisted of an A-B comparison part in which the subjects had to rate their own measured against their own modeled HRTF for various SAQI qualities (see section 2.12.1 for more details). *Whisper*, the tool the listening test was conducted with, automatically normalizes all ratings between 0 to 1 for unipolar scales and between -1 to 1 for bipolar scales. Also, degrees (as for horizontal and vertical direction) are normalized from -1 to 1 in which sources that are shifted 90 to the left would generate scores of 0.5 and sources shifted 45° to the right would generate scores of -0.25.

All ratings were summarized into one file, using a MATLAB script[47]. The mean and the standard error of the mean (SE) were calculated for the ratings and are presented under 3.14. SE is calculated by dividing the standard deviations by the root of the number of observations.

[47] Digital_Appendix\11 Scripts\Matlab\get_ratings.m

3 Results

3.1 Cross-validation

During cross-validation of measured and modeled HRTFs ERB errors and localization errors based on Baumgartner's model were calculated. Also, basic visual comparison of magnitude spectra was done, and numerous plots were generated for each subject's HRTF set which can be found in the digital appendix[48].

Figure 27 shows an exemplary comparison of magnitude spectra for different positions on the horizontal plane for the left ear of subject 9. The general trend of the magnitude spectra is similar for both HRTF versions, but clear differences also become obvious. Below 5kHz the curves basically overlap while the modeled HRFT looks like a smoothed version of the measured one. Above that some magnitude differences appear. Sometimes peaks or notches are higher or lower in one or the other method. In the 0° position, the magnitude spectrum does not fall as deep at 8kHz for the modeled curve but instead the little peak at around 9 kHz seems to be overestimated in the modeled HRTF. For the dominant peak at around 15khz a similar occurrence exists oppositely.

Sometimes the HRTFs look like they are offset on the frequency axis. This behavior is even more dominant for other subjects, as can be seen in Figure 28 showing the magnitude spectra of subject 3. Here, the magnitude spectrum at 0° shows this effect very clearly in which, starting at around 6.5kHz, the trend of the modeled HRTF seems to be shifted upwards for appr. 1.5 kHz. Figure 28 is a very drastic example of this behavior. More slightly variants of which, however, can be noted in several cases. Reasons for this frequency shift could be errors during the acoustical measurement such as a slight movement of the microphones in the ear canals which might have gotten out of place or a slight movement of the subject during measurement. Another reason for differences could be the fact that no torso was included in BEM calculation which has been shown to have an influence on the lower frequency areas of HRTFs (Algazi et al., 2001a). Also hair and clothing was not considered, both have absorbing effects (Katz, 2001b). The semi-automatic selection of the ear canal also yields a source of error because the selected element might not exactly represent the ear canal entrance or the position in which the microphone was placed during acoustical measurement. In the affected frequency region, even a few mm can cause a major frequency shift due to small wavelengths.

[48] Digital_Appendix\6 crossValidation

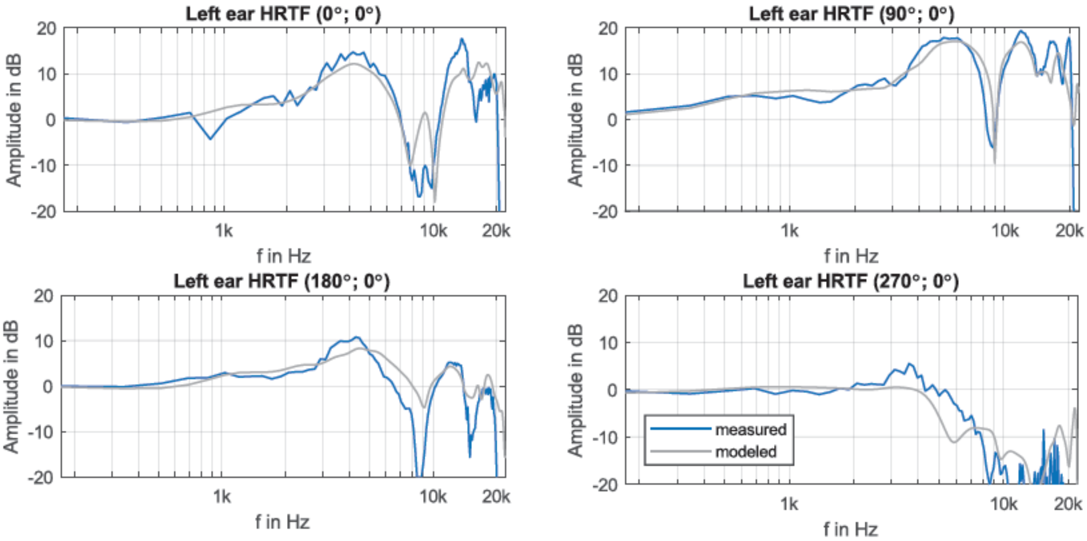


Figure 27: comparison of magnitude spectra for measured (blue) and modeled (grey) HRFT of subject 9's left ear for 4 different positions on the horizontal plane

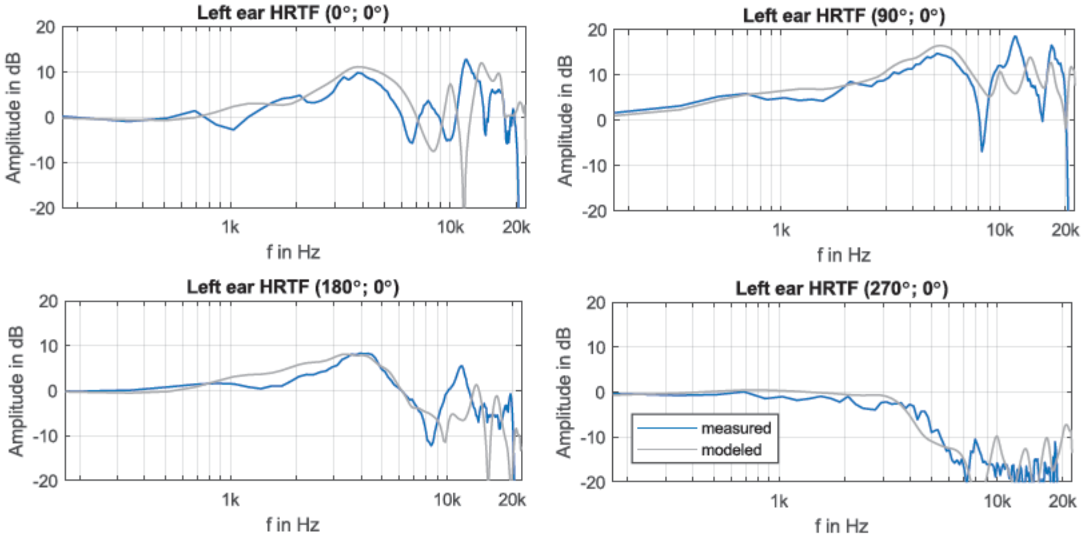


Figure 28: comparison of magnitude spectra for measured (blue) and modeled (grey) HRFT of subject 3's left ear for 4 different positions on the horizontal plane

As mentioned earlier, ERB errors are energetic errors between two HRIRs calculated in auditory filters of equivalent rectangular bandwidth. Due to their definition, ERB errors are psycho-acoustically relevant which is why they can give insight regarding the above-mentioned examples.

Figure 29 shows ERB errors for subject 9 and Figure 30 shows ERB errors for subject 3. When comparing the ERB errors in the top plots averaged across ears and positions for subject 3 and 9, the median (50th percentile, green line) looks similar with the highest values of around 5 dB between 9 and 10 kHz. The bottom plots of Figure 29 and Figure 30 show absolute median ERB errors averaged across ears and frequencies. The median

ERB error for subject 9 varies between 1 and 1.5 dB for most positions with some 2 dB values in elevation areas between 20° and 60°. For subject 3, the median ERB error averaged across ears and frequencies also varies between 1 and 1.5 dB for most positions but between azimuth angles of 90° to 130° and between 270° and 300° ERB errors can partly reach 4 dB. Interestingly these areas represent ear positions.

The dashed red lines representing the full range (percentiles 0 to 100) averaged across positions in the top plots also show that the ERB error for subject 3 is considerably higher in some positions, reaching above 20 dB at around 15 kHz.

Hence, the above-mentioned frequency offset of measured to modeled magnitude HRTFs for subject 3 also show in perceptually relevant ERB errors. However, for most positions ERB errors are low and are comparable to the ERB errors of subject 9.

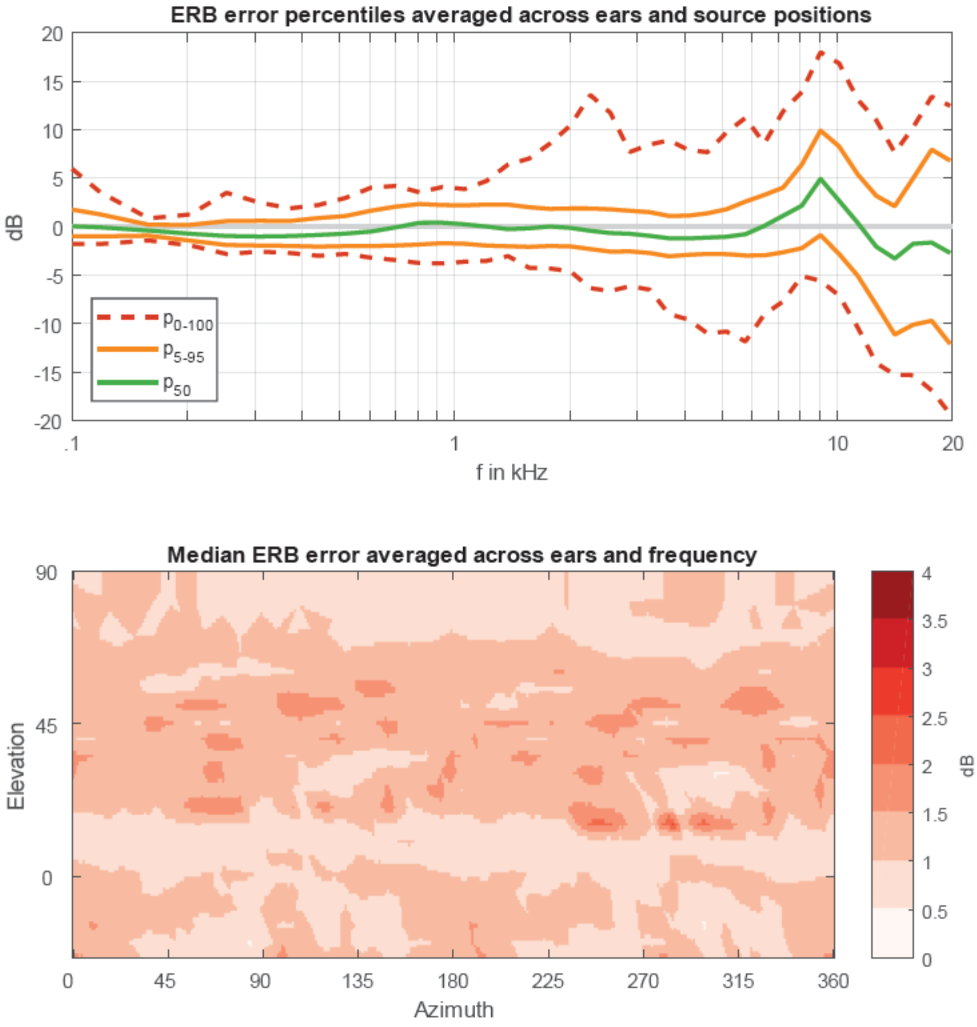


Figure 29: ERB errors between modeled and measured HRTFs of subject 9; ERB errors averaged across ears and positions are shown in the top and absolute ERB errors averaged across ears and frequency are shown at the bottom. The top plot also shows different percentiles.

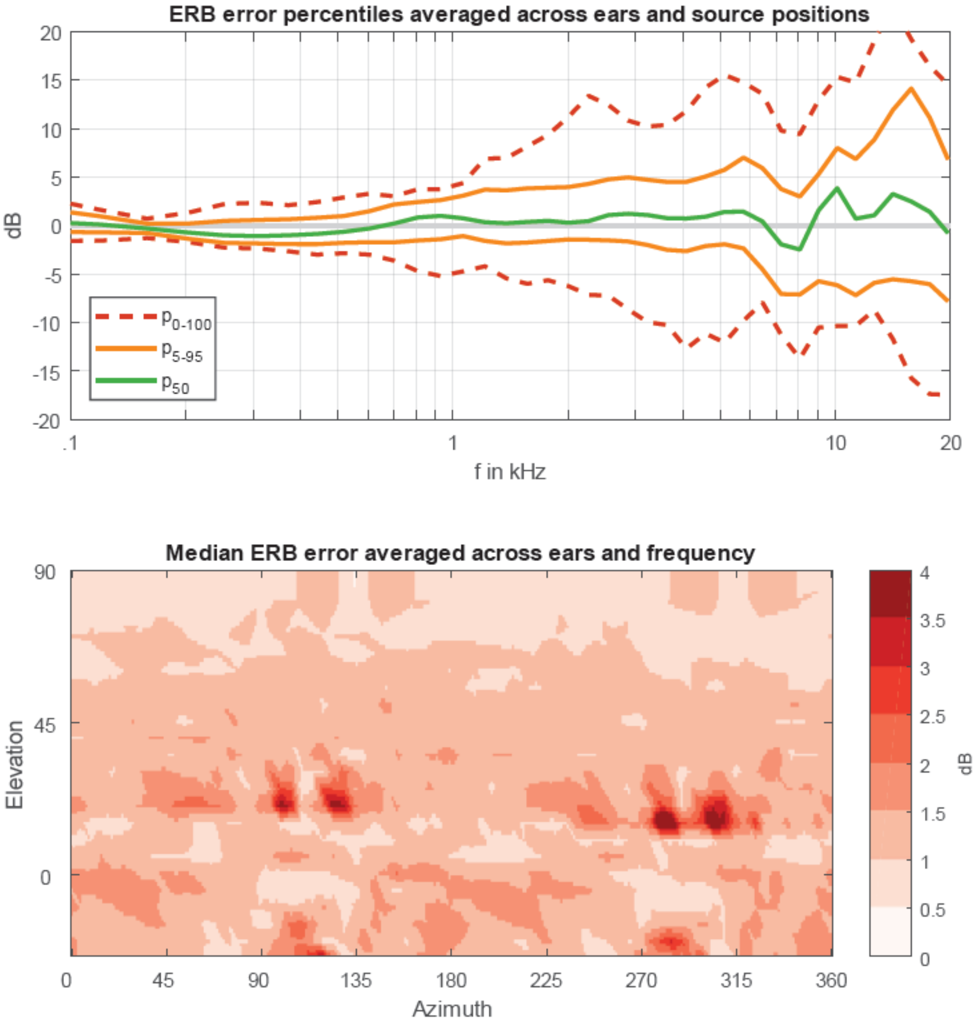


Figure 30: ERB errors between modeled and measured HRTFs of subject 3; ERB errors averaged across ears and positions are shown in the top and absolute ERB errors averaged across ears and frequency are shown at the bottom. The top plot also shows different percentiles.

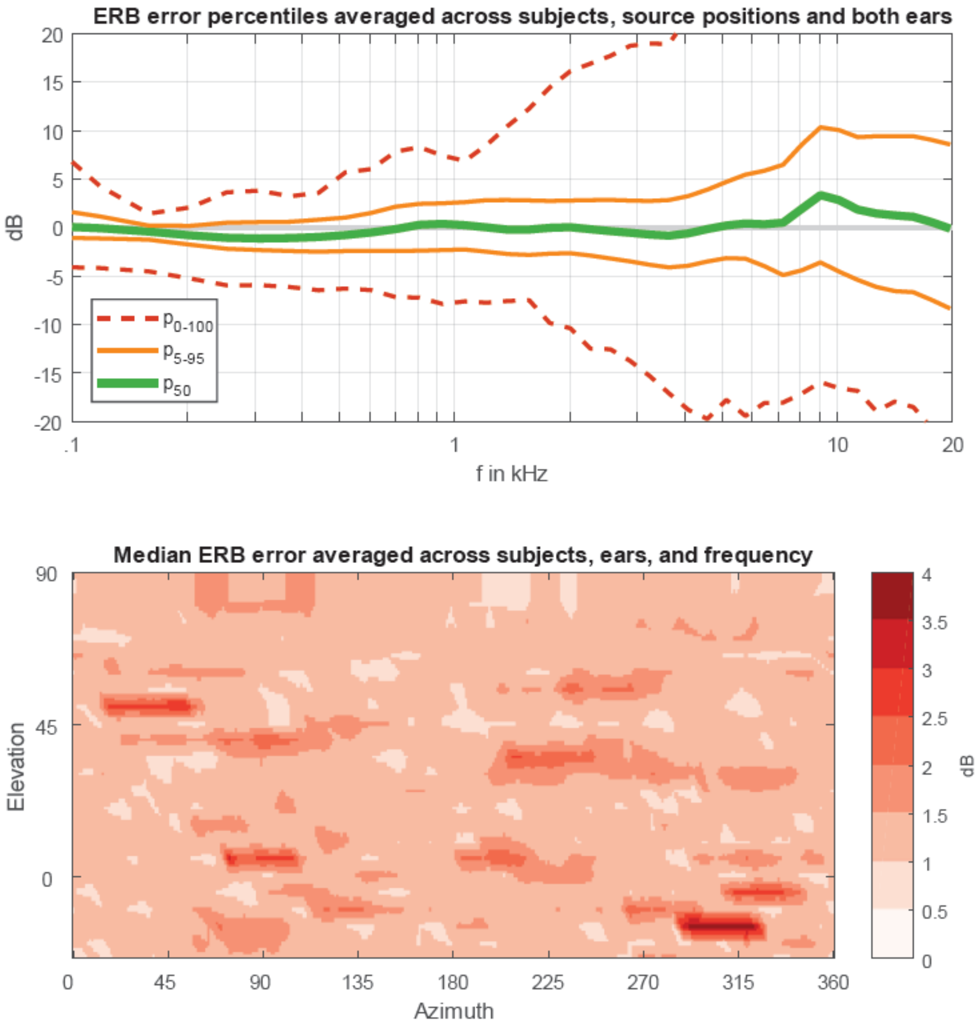


Figure 31: ERB errors between modeled and measured HRTFs averaged across all subjects and ears; ERB errors averaged across ears, subjects and positions are shown in the top and absolute ERB errors averaged across ears, subjects and frequency are shown at the bottom. The top plot also shows different percentiles.

Figure 31 shows the median ERB error averaged across all subjects, all ears and all frequencies. The plots indicate that ERB errors of measured and modeled HRTF data on average range between 1 and 2 dB for most positions. Only in some positions ERB errors of 2.5 to 3.5 dB can be reached. The median error in the top plot (green line) shows that at around 9 kHz the highest value of 4 dB is reached. Usually, the first pinna notch can be found in this frequency range, so it can be assumed that differences in the depth or the position of the first pinna notch between HRTF versions are responsible for ERB errors.

The bottom plot of Figure 31 shows that for at least 50% of data overall expectable results can be reached on average.

Results of the predicted localization performance using Baumgartner's model (Baumgartner et al., 2014) are shown in Figure 32 in form of a boxplot including the median, the interquartile range between upper and lower quartiles and the full range of the data indicated by the whiskers (vertical lines). The quadrant error rate (QE) and the polar root-mean-square (PE) are presented. In addition to the measured and modeled data, also cross-validated data was calculated by using the measured data as the template and the modeled as the target.

In the measured HRTF set the QE is 15% lower than in the modeled set. The PE median for the measured set is around 35° which is just slightly above the general inaccuracy of sensitivity of ± 30° for localization on the median plane (Baumgartner et al., 2014). The PE median for the modeled set is noticeably higher with 42°. Even 47° are reached for the PE median in the cross-validated case. Pretending the case that the measured set would represent the true HRTFs, a noticeable localization error would occur when using the modeled HRTF set. Given the sensitivity of ± 30°, the median error difference of around 12° seems moderate.

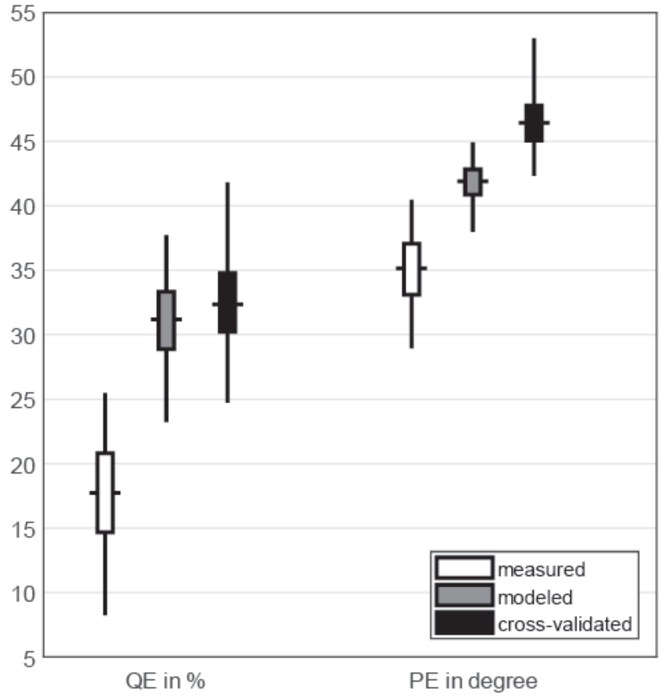


Figure 32: Boxplot of PE and QE for measured, modeled and cross-validated cases. QE and PE were calculated using Baumgartner's approach. The box represents the interquartile range (upper and lower quartiles) and the full range of the data via whiskers (vertical lines)

3.2 Parameter Tuning

Table 6 contains the results of the parameter tuning procedure to find the most useful relationship between the listener's AFs and the HRTF owner's AF. Here, only results of the measured dataset is being reported but the modeled dataset yielded similar values.

The table shows that the difference of AFs seems be the most effective transformation mean. However, R^2 values of .2 for *Difference* and *Coloration* still low but no interactions have been added up to this point.

Table 6: Parameter tuning for finding an optimal relationship of AFs

Relationship of AFs	***SAQI feature***	***R²***	
		Forced entry	stepwise backwards
$\frac{listener's\ AF}{owner's\ AF}$	Difference	.199	.197
	Coloration	.158	.157
	Externalization	.104	.104
	Source Position	.144	.143
$listener's\ AF - owner's\ AF$	Difference	.198	.198
	Coloration	.198	.198
	Externalization	.115	.113
	Source Position	.139	.139
$\frac{(listener's\ AF)^2}{(owner's\ AF)^2}$	Difference	.180	.179
	Coloration	.148	.146
	Externalization	.099	.098
	Source Position	.147	.146
$\frac{(listener's\ AF)^3}{(owner's\ AF)^3}$	Difference	.166	.164
	Coloration	.140	.139
	Externalization	.095	.094
	Source Position	.144	.141
$\frac{\sqrt{listener's\ AF}}{\sqrt{owner's\ AF}}$	Difference	.202	.201
	Coloration	.164	.162
	Externalization	.108	.107
	Source Position	.138	.137
$listener's\ AF^2 - owner's\ AF^2$	Difference	.197	.197
	Coloration	.167	.165
	Externalization	.116	.113
	Source Position	.146	.145

As previously mentioned, regression models were calculated using both an approach with forced entry of all predictors and a stepwise backwards method. When comparing both columns, almost similar values can be seen. This indicates that no or only small suppressor effects exist. Suppressor effects would occur if a predictor had a significant effect only when another variable is held constant (Field, 2009). This does not seem to be the case.

3.3 Comparison of rating means

Rating means for measured HRTFs are higher than ratings means for modeled HRTFs, as can be seen in Figure 33. This is interesting because as was described earlier, each subject rated the same set of HRTFs once in the measured and once in the modeled version. Comparison of rating means is therefore informative in terms of the question of how differences between a given HRTF and one's own HRTF stand out differently for the two data sets.

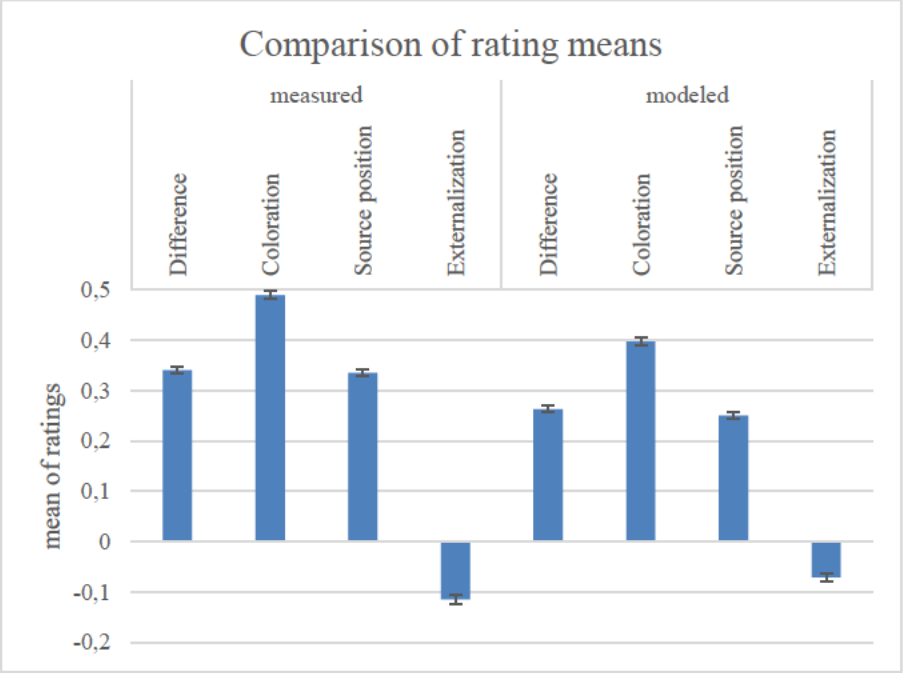


Figure 33: absolute rating means, including standard errors of the mean

Table 7 contains the results of paired-sample *t*-tests and calculated effect size *r*, as described earlier. On average, SAQI items *Difference*, *Coloration* and *Source position* were rated significantly higher with measured HRTFs than with modeled HRTFs. The effect size of these differences are almost medium strong (r=.3), according to Cohen's definition (Cohen, 1988). *Externalization* was rated significantly lower (meaning more internalized) with measured HRTFs than with modeled HRTFs, however with a rather small effect size.

These findings suggest, that differences between one's own and another person HRTF are presumably clearer when listening to measured HRTFs which results in higher ratings. This might be due to more individual details like hair and clothing which are covered in measured HRTFs but neglected in modeled versions. Also, measured HRTFs are more prone to uncertainties caused by noise or positioning errors of microphone and subject which might in consequence cause higher perceived differences in one's own and another person's HRTF.

Table 7: *t*-statistic and effect sizes for Comparison of rating means

Comparison pair	***Difference of means***	***SE***	***t***	***Effect size r***
Difference - measured and modeled	0.08	0.01	9.30**	.25
Coloration - measured and modeled	0.09	0.01	10.65**	.29
Source position - measured and modeled	0.08	0.01	10.61**	.29
Externalization - measured and modeled	-0.04	0.01	-4.03**	.11

**p < .01

3.4 Results of multiple regression with interaction

Regression models with interactions and block-wise entry were created. As mentioned above, only the main effects were force-entered into the model while interactions were added stepwise forward. This reduced 1276 predictors to 67 to 109, depending on the SAQI item. Tables with regression coefficients for each of the four SAQI items and both data sets can be found in the digital appendix[49]. Table 19 in the appendix contains results for *Difference* exemplarily with standardized and unstandardized coefficients, the standard error and p-values.

Table 8 contains R^2 and MSE values for all SAQI items. A comparison with regression results in Table 6 makes obvious that including squared predictors and interactions significantly enhances the models. This shows that the statistical relationship between HRTFs and morphological features is highly complicated.

Table 8: coefficients of determination and mean square errors for multiple regression with interactions for different SAQI items and data sets

SAQI item	***Data set***	***No. predictors***	***R^2***	***R^2 adj.***	***MSE***
Difference	measured	98	.47	.42	0.035
Coloration		86	.46	.42	0.041
Source position		84	.40	.35	0.044
Externalization		67	.29	.25	0.072
Difference	modeled	94	.47	.43	0.032
Coloration		95	.49	.45	0.039
Source position		109	.48	.43	0.029
Externalization		73	.26	.21	0.054

[49] Digital_Appendix\10 Statistical_Models\Regression_Models

All models are statistically significant ($p<.01$). Like the above shown results without interactions in Table 6, *Difference* and *Coloration* yield the highest coefficients of determination while *Externalization* yields the lowest. Interestingly, better results are achieved for the modeled HRTF set in general. The lowest MSE can be found for *Source position* in the modeled set.

3.5 Mixed model results

3.5.1 Model criteria and goodness-of-fit measures

Hierarchical models with random intercepts for participants were created and tested as explained under 2.13.5. Table 9 shows relevant model criteria for models without random intercept and for final models with random intercept which were further reduced in some cases.

Table 9: model criteria of multilevel mixed-effects models showing Deviance (-2logL) and AIC for models without intercept and final models with random intercept. The variance of random effects var(u_{0j}) and results from chi-square difference test are also included.

		No random intercept			***Final model with random intercept***			
SAQI item		***Deviance***	***AIC***	$\chi^2_{diff}(1)$	***Deviance***	***AIC***	***var(u_{0j})***	***No. pred.***
Difference	measured	-294.80	-96.80	372.2**	-632.15	-476.15	0.0059	75_r
Coloration		329.98	503.98	815.6**	-485.59	-307.59	0.0072	86
Source position		43.89	213.89	417.5**	-373.58	-199.58	0.0050	84
Externalization		312.56	448.56	46.6**	275.66	403.66	0.0015	61_r
Difference	modeled	-665.46	-475.46	94.5**	-759.92	-565.92	0.0003	94
Coloration		107.53	299.53	629.9**	-514.98	-330.98	0.0012	89_r
Source position		-676.70	-456.70	196.5**	-873.17	-649.17	0.0009	109
Externalization		-98.54	49.46	1.9	-95.69	50.31	0.0001	70_r

Note: ** $p< .01$. Significant results for the chi-square difference test indicate model enhancement through random intercept.

The χ^2_{Diff} column shows the deviance difference between models without and unreduced models with random intercept. The results of χ^2_{Diff} tests (chi-square distribution) indicate the significance of variance in intercepts across participants. This variance var(u_{0i}) is very small in most cases, however, the mostly significant results of χ^2_{Diff} tests and lower AIC values for models with random intercept show that adding random intercepts for participants enhances the models significantly. This is true for all cases except *Externalization* in the modeled data set. SAQI item *Coloration* in the measured data set shows the highest variance across participants, var(u_{0i}) $=0.0072$, $\chi^2_{Diff}(1) = 815.6$, $p<.01$.

As previously explained, some models were reduced further by deleting non-significant predictors after a random intercept was added. Again, χ^2_{Diff} tests and AIC comparison showed if reducing the models enhances them significantly. For 50% this was the case. Reduced models are marked with a lowered “r” in the number of predictors column and they also have fewer predictors than in Table 8.

Table 10 contains goodness-of-fit measures for the final hierarchical mixed models. As explained under 2.13.5., the conditional R^2 and MSE values account for full models including random effects, whereas marginal values describe variance explained by fixed effects only. Significant model enhancements by adding random intercepts and accounting for the hierarchical structure of the data was already expressed by the model criteria in Table 9 and can also be seen by MSE values given in Table 10. Comparison to MSE values in Table 8 shows that MSE values could be decreased.

Most variance is explained for *Difference* and *Coloration,* the highest $R^2_{\mathrm{conditional}}$ value being .48 for *Coloration* in the modeled data set. The lowest MSE of 0.028. can be found for *Source position* in the modeled data set. The least variance can be explained for *Externalization* with R^2=.28 for the measured and .26 for the modeled data set.

Table 10: marginal and conditional values for R^2 and the MSE for the final hierarchical regression models

SAQI item	Data set	$R^2_{marginal}$	$R^2_{conditional}$	$MSE_{marginal}$	$MSE_{conditional}$
Difference	measured	.37	.47	0.039	0.032
Coloration		.34	.45	0.045	0.036
Source position		.32	.39	0.046	0.040
Externalization		.27	.28	0.073	0.071
Difference	modeled	.46	.47	0.032	0.032
Coloration		.46	.48	0.039	0.037
Source position		.45	.47	0.030	0.028
Externalization		.25	.26	0.054	0.054

In general, acceptable values are achieved for most SAQI items predicted by AFs with interactions using multilevel mixed-effects models with random intercepts for participants. The models can account for approximately 50% of the variance, still leaving the rest unexplained. However, in Cohen’s definition of effect sizes for multiple regression the amount of explained variance is equivalent to a large effect (Cohen, 1988).

Developed mixed models with estimates, standard errors and confidence intervals can be found in the digital appendix[50] for each SAQI item. Table 20 in the appendix presents the estimates for *Difference* with measured HRTFs exemplarily.

[50] Digital_Appendix\10 Statistical_Models\Mixed_Models

3.5.2 Side note: hierarchical structure

On a short side note it should be mentioned that χ^2_{Diff} values in Table 9 are often much higher for the measured HRTF set. The differences between $R^2_{\text{conditional}}$ and R^2_{marginal} in Table 10 point in the same direction. In the modeled data set there is almost no difference between R^2 values with and without included random effects. This indicates that adding a hierarchical structure to the data was much more helpful for the measured HRTF set. In Figure 33 and Table 7 under 3.3 rating means were compared. On average, ratings for measured HRTFs were significantly higher than ratings for modeled HRTFs. It was hypothesized that differences between one's own and another person's HRTF are presumably clearer when listening to measured HRTFs. It therefore appears that individuality in rating plays a bigger role when higher differences are perceived than in cases with lower differences.

3.6 Evaluation of prediction models for HRTF individualization

In the previous section regression models were created to predict the given listening test ratings for perceived differences of one's own and another person's HRTF. The amount of explained variance of around 50% in the final multilevel mixed-effects models already suggests that the models could be used for HRTF individualization based on AFs. To express this finding more tangibly the models were used to predict the best fitting HRTF for each subject. This was done by identifying the HRTF out of the 15 HRTFs which each subjects evaluated for which the regression model predicted the lowest rating (which would then be the smallest predicted perceived difference). The actual rating of this HRTF was then compared to the minimal and maximal rating that the subjects gave for the 15 evaluated HRTFs.

Figure 34 shows the averaged results across all subjects for each SAQI item and both versions of HRTF (measured blue and modeled HRTFs light-blue). The blue bars are the averaged minimum and maximum ratings and represent the average range of HRTFs that the subjects evaluated. The white lines represents the averaged mean values of HRTF ratings. The black lines represent the averages of the actual HRTF ratings that were selected to be the best fits by the final multilevel models.

The figure shows that HRTF individualization based on AFs using the developed regression modeles seems to work well in most cases. Externalization again is the exception which is not surprising given the previous stated results. The selected HRTF can usually be found closer at the bottum of the blue bars representing the best case than at the white lines representing the mean ratings. However, also room for optimization of the regression models becomes obvious because it would be desirable that the predicted best fits would be even closer to to the real best fits.

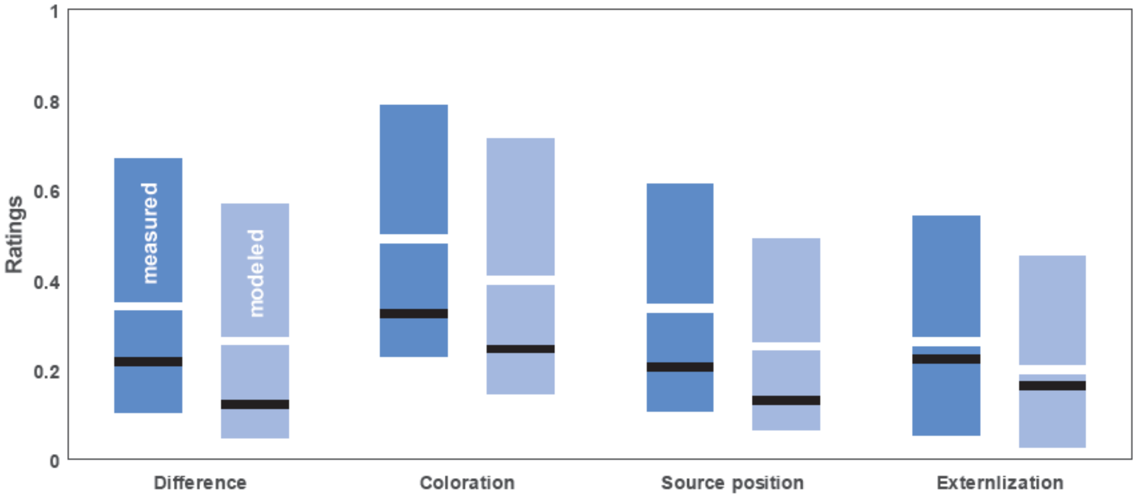


Figure 34: Evaluation of HRTF prediction models. The blue bars represent the average range of given ratings (blue bars for measured HRTFs and light-blue bars for modeled HRTFs). The white lines represent the averaged mean ratings. The black lines represent the actual ratings of the HRTF which was predicted to be the best fit using the developed multilevel regression models.

3.7 Anthropometric feature correlation

Pearson correlation coefficients were calculated for 21 AFs. Table 18 in the appendix contains correlation results. Results suggest several strong correlations between AFs. Some of these strong correlation results were selected and are presented in Table 11. They are sorted in such a way to show that some AFs correlate well with several other features.

Most AFs in Table 11 are torso, head or neck related. Shoulder circumference for example correlates strongly with shoulder width, which is not surprising. But it also correlates highly with neck width, r=.79 and even with pinna height, r=.58, (all ps<.001, one-tailed). Pinna height in turn correlates with several pinna related AFs.

Table 11: results for some selected strong anthropometric feature correlations

Anthropometric features		*Pearson correlation coefficient*
	Shoulder circumference and...	
x1, x17	... head width	.66**
x6, x17	...neck width	.79**
x9, x17	...torso top width	.74**
x12, x17	...shoulder width	.77**
x8, x17	...neck depth	.69**
d5, x17	...pinna height	.58**
	Pinna height and....	
d5, d4	...fossa height	.57**
d5, x8	...neck depth	.58**
d5, d6	...pinna width	.54**
d5, d1	...cavum concha height	.50**
d5, x6	...neck width	.57**
	Neck width and...	
x1, x6	...head width	.76**
x8, x6	...neck depth	.67**
x12, x6	...shoulder width	.70**
	Shoulder width and...	
x12, x8	...neck depth	.56**
x12, x9	...torso top width	.71**

**. Correlation is significant at the 0.01 level (1-tailed).

3.8 Effect sizes for Difference

For each evaluated SAQI feature anthropometric features that present themselves to be top predictors were identified. As mentioned before, this was done using two different data sets, one in which HRTFs where measured acoustically and one in which HRTFs were numerically modeled. To identify general anthropometric top predictors, it is therefore useful to work with two different data sets based on identical anthropometrics because a general identification of top predictors should show similar results. This is not always the case as will be shown in the following sections. Sometimes a certain AF appears to be in the top 3 of the modeled case but in the measured data set occupies the last place (for example x12 in Coloration). Self-contradicting results as these therefore were not considered in finding the best predictors for a certain SAQI feature. The first 10 predictors in the tables that where sorted by predictor importance where examined.

Table 12 contains effect sizes of anthropometric predictors for the perceived *Difference*.

Table 12: Effect sizes (R^2_{increase}) for *Difference*. Predictors which are among the top 10 in both HRTF sets are in bold letters.

Difference - measured		***Difference - modeled***	
Main Effect	***Predictor Importance ($R^2_{increase}$)***	***Main Effect***	***Predictor Importance ($R^2_{increase}$)***
d2	.053	**d1**	.061
d1	.047	**d2**	.057
d4	.036	θ2	.051
x7	.035	**θ1**	.042
d6	.033	**d3**	.033
x16	.032	x14	.033
d3	.030	x9	.030
θ1	.030	x5	.028
x12	.029	**d6**	.027
x8	.029	x2	.027
x1	.028	x1	.025
x4	.026	d4	.021
x17	.026	x8	.020
x14	.026	x7	.017
x2	.022	x3	.016
θ2	.021	x17	.014
x6	.016	x4	.014
x5	.015	x12	.014
x3	.013	x16	.012
d5	.010	x6	.011
x9	.008	d5	.011

Cavum concha height *d1* and cymba concha height *d2* clearly present themselves as the most important features for predicting *Difference*. That is, if cavum concha height and cymba concha height are deviating between listener and HRTF owner, the listener perceives a difference in respect to his own HRTF.

Pinna width *d6* and cavum concha width *d3* which are correlated, r=.41, p (one-tailed) < .01 are also among the first 10 predictors in both cases. The rotation angle θ1 also seems to play a big part in generating a perceivable difference when deviating between listener and HRTF owner.

Interestingly, for *Difference* no torso or head-related AFs can be identified among the first important predictors being present in both, the modeled and the measured HRTF set. Head circumference *x16* and height *x14* appear among the first predictors but not consistently among both sets.

3.9 Effect sizes for Coloration

Table 13 contains the sorted predictors for *Coloration*. The rotation angle of the pinna θ1 seems to play a major role in predicting a perceived difference in timbre when comparing one's own HRTF to another person's HRTF. Both, in modeled and in the measured test set θ1 is among the first three predictors. Although the rotation angle is significantly correlated to the flare angle θ2, r=-.27, p (one-tailed) < .01, the flare angle only appears in the measured data set. Cymba concha height *d2* is also among the top 5 predictors in both sets while being the second most important predictor for modeled data. Cavum concha width *d3* has the highest predictor importance in the measured set and also plays a role in the modeled part, taking position 6.

Again, it can be noted that no head or torso related features are present consistently in both data sets among the top 10 positions. However, shoulder width *x12* is the third strongest predictor in the modeled part. Shoulder width is significantly correlated to head height *x2,* r=.34, head depth *x3,* r=.31 and neck depth *x8,* r=.56 (all ps <.01, one-tailed) which are all among the first 8 predictors in the measured set. It can therefore be assumed that head and torso related features also play a role in predicting the perceived coloration of another person's HRTF in comparison to one's own.

Table 13: Effect sizes ($R^2_{increase}$) for *Coloration*. Predictors which are among the top 10 in both HRTF sets are in bold letters.

Coloration - measured		***Coloration - modeled***	
Main Effect	***Predictor Importance ($R^2_{increase}$)***	***Main Effect***	***Predictor Importance ($R^2_{increase}$)***
d3	.068	**θ1**	.124
θ2	.047	**d2**	.101
θ1	.041	x12	.074
x2	.035	d5	.071
d2	.032	d6	.067
x3	.029	**d3**	.064
x7	.028	**d4**	.057
x8	.027	x9	.048
d4	.027	x6	.044
x4	.027	x14	.041
x17	.026	x17	.039
x5	.025	d1	.036
d1	.025	θ2	.035
x6	.024	x4	.035
x14	.023	x5	.031
x16	.022	x8	.027
d6	.021	x7	.026
d5	.012	x3	.020
x1	.010	x1	.012
x9	.007	x2	.009
x12	-.002	x16	.008

3.10 Effect sizes for Source Position

Predictors for *Source position* stand out to the previous described items *Difference* and *Coloration.* Table 14 contains predictors, sorted by effect strength and this time head and torso related AFs can also be found among top predictors. The subject's height *x14* is the top predictor in the measured set and the sixth predictor in the modeled set. Height significantly correlates to various other features, the two strongest correlations being shoulder circumference *x17,* r=.51 and head depth *x3,* r=.49, (all ps <.01, one-tailed), both also playing a role as top predictors for *Source position.* Shoulder circumference *x17* takes the forth place in ratings for the measured data set but only place eleven for the modeled set. Head depth *x3* is the fifth most important predictor for measured HRTFs and the eight most important for modeled HRTFs, respectively. Pinna flare angle θ2 is among the top four predictors in both data sets. Rotation angle θ1 has an extraordinary high predictor importance in the modeled set but a much lower importance in the measured set, which is at least confusing but appears similarly for *Coloration* in Table 13. Other pinna-related features among the top ten most important predictors for *Source position* are cymba concha height *d2,* cavum concha width *d3,* fossa height *d4* and pinna height *d5.*

The fact that for *Source position* head and torso-related features play an important role is presumably since ITD plays a major role for source localization, as described earlier. ITD was shown to correlate strongly to head width, i.e. the inter-tragus distance (Algazi et al., 2001b).

Table 14: Effect sizes ($R^2_{increase}$) for *Source position*. Predictors which are among the top 10 in both HRTF sets are in bold letters.

Source Position - measured		***Source Position - modeled***	
Main Effect	***Predictor Importance ($R^2_{increase}$)***	***Main Effect***	***Predictor Importance ($R^2_{increase}$)***
x14	.048	**θ1**	.128
d2	.045	**d3**	.093
θ2	.043	**d4**	.079
x17	.039	**θ2**	.073
x3	.036	d1	.071
d3	.036	**x14**	.069
d5	.036	**d2**	.066
θ1	.033	**x3**	.047
d4	.029	x5	.046
d6	.027	**d5**	.045
x9	.026	x17	.042
x1	.025	x1	.039
x12	.021	x7	.035
d1	.021	x8	.033
x16	.021	x12	.031
x8	.021	x2	.029
x2	.019	x6	.028
x7	.018	d6	.022
x4	.017	x9	.019
x6	.015	x4	.014
x5	.013	x16	.013

3.11 Effect sizes for Externalization

Cavum concha width *d3* and pinna width *d6* can be identified as the most relevant features for *Externalization.* No real consensus can be found for cavum concha height *d1* which is the most important predictor for *Externalization* for modeled HRTFs and only the ninth most important one for measured HRTFs. Pinna offset down *x4* did not play a huge role for previously discussed SAQI item but for *Externalization* it can be found at the sixth and seventh rank. Pinna offset down describes the pinna relative to the head. It is therefore interesting that another feature sharing this specific characteristic, namely the pinna flare angle θ2 can also be found among the most important predictors for *Externalization* among both datasets.

Head circumference *x16* is the tenth most important predictor on both cases and adds a head-related feature to the model.

Table 15: Effect sizes ($R^2_{increase}$) for *Externalization*. Predictors which are among the top 10 in both HRTF sets are in bold letters.

Externalization - measured		***Externalization - modeled***	
Main Effect	***Predictor Importance ($R^2_{increase}$)***	***Main Effect***	***Predictor Importance ($R^2_{increase}$)***
d3	.049	**d1**	.036
x3	.042	**d3**	.022
d6	.033	**d6**	.019
x14	.028	**θ2**	.019
x5	.026	d5	.019
x2	.025	**x4**	.018
x4	.024	θ1	.018
θ2	.024	x7	.017
d1	.022	x8	.016
x16	.021	**x16**	.015
d2	.020	d2	.014
d5	.020	x3	.013
θ1	.016	x12	.011
d4	.015	x1	.010
x1	.012	x14	.009
x7	.010	d4	.007
x9	.008	x17	.007
x17	.007	x6	.007
x6	.005	x9	.006
x12	.003	x5	.005
x8	.002	x2	.000

3.12 Anthropometric key features

The listening test was conducted with four SAQI features representing different perceivable differences in binaural hearing. To identify general anthropometric key features for HRTF individualization, the above results for all SAQI qualities and data sets were combined calculating the average predictor importance (the amount of R^2_{increase}). The results are presented in Table 16. AFs are sorted by the average predictor importance strength.

Table 16: anthropometric features with average predictor importance sorted by strength

Identifier	*Anthropometric feature*	*Predictor Importance ($R^2_{increase}$)*
θ1	pinna rotation angle	.054
d3	cavum concha width	.049
d2	cymba concha height	.048
d1	cavum concha height	.040
θ2	pinna flare angle	.039
x14	height	.034
d4	fossa height	.034
d6	pinna width	.031
d5	pinna height	.028
x3	head depth	.027
x17	shoulder circumference	.025
x5	pinna offset back	.024
x7	neck height	.023
x12	shoulder width	.023
x4	pinna offset down	.022
x8	neck depth	.022
x2	head height	.021
x1	head width	.020
x9	torso top width	.019
x6	neck width	.019
x16	head circumference	.018

Eight among the top nine features are pinna related, the strongest being pinna rotation. The pinna flare angle is the fifth strongest predictor on the list. The findings are based on perceptual motivated analysis but are in agreement with another work in which the morphological influence on HRTFs was studied (Xu et al., 2007) and pinna rotation and flare angle were isolated as key features. Liu and Zhong (2016) partly link the enhancement of their anthropometry-based matching methods in comparison to prior methods to the inclusion of pinna rotation and flare angle because these measures add information about pinna position relative to the head surface.

The second most important feature is the cavum concha width and the forth predictor in the list is the cavum concha height. This finding does not surprise and underlines the results of previous other findings (Algazi et al., 2001b; Fels and Vorländer, 2009; Ghorbal and Auclair, 2017).

Cymba concha height takes the third place, right behind cavum concha. This finding was in some way expected because of the importance of the cymba concha height for *Difference.* However, so far only in Bomhardt's work the cymba concha was also identified as a highly influential feature (Bomhardt, 2017). Besides that, no other studies are known to validate this finding and so far, cymba concha height has always been found to be among minor parameters to influence HRTFs (Ghorbal and Auclair, 2017).

Among the top ten key features only two non-pinna related AFs can be found. The subject's height is the sixth most important predictor. For *Source position* the height was the top predictor which was not surprising due to its correlation with shoulder, head and torso related features.

The results in Table 16 also help identify minor parameters. Most neck, shoulder and head related features seem to contribute only little to the explained variance. The already mentioned height of the subject seems to contribute more than head depth and head width, the latter being important measures for ITD (Algazi et al., 2001b). It can be noted that results suggest that key features for HRTF individualization mainly consist of pinna features, namely cavum concha, cymba concha and pinna angles.

3.13 SVM regression results

A support vector machine regression was used to create a recommender to predict unknown data by splitting the listening test results into training and testing set. More details about the procedure can be found under 2.13.7.

In Table 17 the relevant kernel hyperparameters are presented for a gaussian kernel. Also, R^2 measures were calculated using a standard regression approach, which is why these values must be viewed with caution. Using machine learning techniques, overfitting of data can easily happen which is why 10-fold cross-validation was used. However, it can still be seen, that R^2 values for the training set (when testing the SVM-R on the same data it was trained with) are noticeably higher than in the previous mentioned multilevel modeling. Feeding new and unseen data to the SVM-Rs however, leads to much lower amount of explained variance.

More interesting are MSE values for test and training set because they can directly be compared to other models. Here again, for the training set comparably low MSE values are reached which are much higher for the test set.

Table 17: SVM-R kernel hyperparameters and goodness-of-fit values for all SAQI items

SAQI item		***Kernel scale***	***Epsilon***	***Box constraint***	***R$^2_{train}$***	***R$^2_{test}$***	***MSE$_{train}$***	***MSE$_{test}$***
Difference	measured	4.0281	0.1421	0.5761	.67	.26	0.021	0.050
Coloration		6.2411	0.1743	1.1480	.58	.27	0.032	0.056
Source position		6.3282	0.1709	0.5736	.46	.13	0.042	0.054
Externalization		5.6309	0.2381	0.2956	.36	.04	0.065	0.101
Difference	modeled	6.4510	0.1146	0.4990	.54	.25	0.026	0.053
Coloration		5.0772	0.1607	0.5552	.62	.24	0.029	0.055
Source position		5.0609	0.1228	0.4011	.54	.29	0.025	0.043
Externalization		0.0158	0.2756	0.1428	.30	.00	0.048	0.085

The prediction of SAQI ratings using SVM-R works differently well for different SAQI qualities. While *Externalization* ratings seem to be difficult to predict (which might have to do with the very low reported difference during this part of the test) other items, such as *Difference* and *Coloration* work better. The lowest MSE$_{test}$ could be achieved for *Source position* with the modeled data set, which is in accordance with multilevel model results (compare to Table 10). However, the amount of variance explained and the achieved MSE values for predicting new data still leave room for further investigation. Also, the huge gap between error for training and error for test set despite cross-validation suggests an overfitting of the training set data.

Finally, a multilevel mixed-effects model was created for the test set and tested on the training set for the items *Difference* (measured) and *Coloration* (modeled). As explained above, a previous block-wise regression with all parameters, including interactions and squares was carried out. This was done to compare multilevel modeling to SVM-R in respect to a recommender application predicting unknown results. An $\mathrm{MSE}_{\mathrm{test}}$ of 0.053 was achieved for *Difference* and 0.062 was achieved for *Coloration.* The achieved values using SVM-R, which are presented in Table 17, indicate that SVM-R fits unknown data slightly better and that for recommender applications SVM-R should be preferred over multilevel mixed-effects modeling.

3.14 Similarity test results

The A-B comparison between the subject's own modeled against the own measured HRTF was carried out for several SAQI qualities. The mean and SE were calculated with combined rating results. As it was noted earlier, all ratings were normalized between 0 to 1 and -1 to 1 for unipolar scales, respectively. Results are shown in Figure 35 to Figure 37.

More general SAQI items were gathered in Figure 35 showing that subjects perceived a distinct *Difference* between their measured and modeled HRTF with a rating of around 0.3. Considering the results in Figure 36 in which coloration-related SAQI items are presented, it can be assumed that the perceived *Difference* by a great proportion stems from measured HRTFs which are brighter on average and seem to have more energy in the higher frequency region. This can clearly be seen at items *High-frequency tone color*, *Tone color bright-dark* and *Sharpness*, and the negative rating of *Low-frequency tone color*. Participant's comments after the listening test also reflect and undermine these findings. SH order equalization, as described in 2.12.2 was applied to adjust the frequency roll-off of measured HRTFs to modeled HRTFs, however it could be that a stronger equalization is needed to lift higher frequency regions of measured HRTFs to the same level as modeled HRTFs.

More amplitude in higher frequency regions might also be partially reflected in the items *Crispness*, *Loudness* and even *Speech intelligibility* which are rated higher for modeled HRTFs.

The items *Naturalness* and *Degree-of-liking* are relatively equal for measured and modeled HRTFs, which is noteworthy because of the mentioned coloration effects.

When it comes to *Localizability* modeled HRTFs tend to be slightly more ambiguous compared to measured HRTFs which is suggested by the rating of approximately -0.05. Other source-related items such as *Width* and *Height* suggest that using modeled HRTFs the source felt bigger, which, however does not reflect in *Source expansion*. Additionally, modeled HRTFs are perceived as being closer and more internalized which is reflected in *Distance* and *Externalization* ratings of approximately -0.1.

When comparing measured and modeled HRTFs some *Front-back position* confusion with a rating of about 0.2 can be seen. Also, a vertical offset exists with a mean rating of around 0.7 which represents an offset of around 11°.

It should be noted, that in offsets and especially in *Front-back confusions* it is not clear which of the HRTF versions cause the effect. This is due to a lack of external reference, as mentioned earlier.

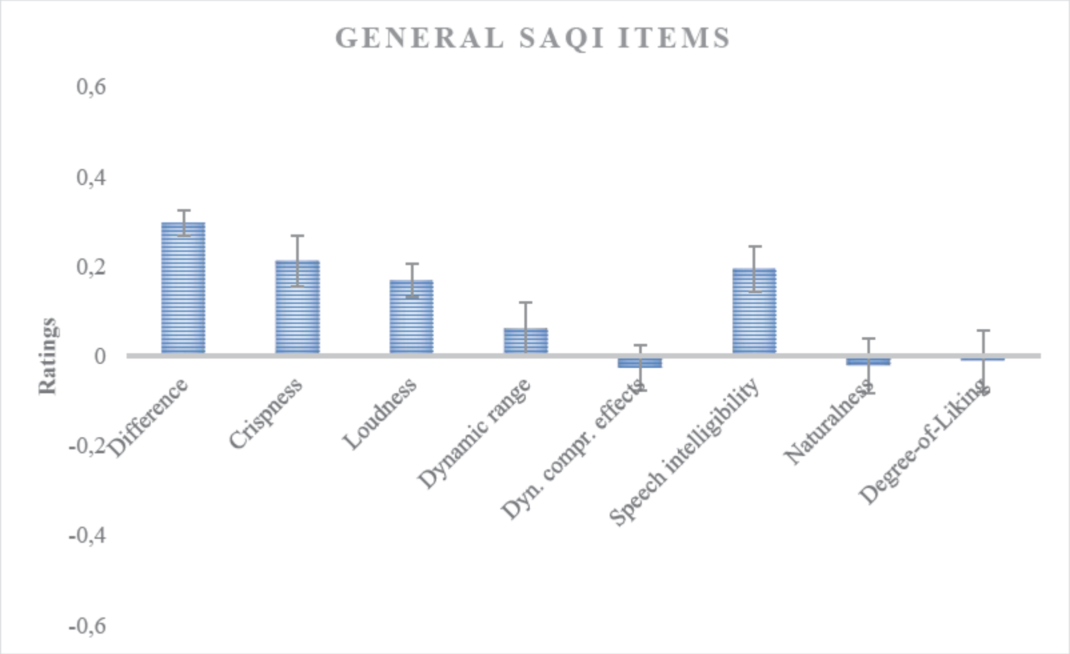


Figure 35: Similarity test for measured and modeled HRTFs – general SAQI items with error indicators

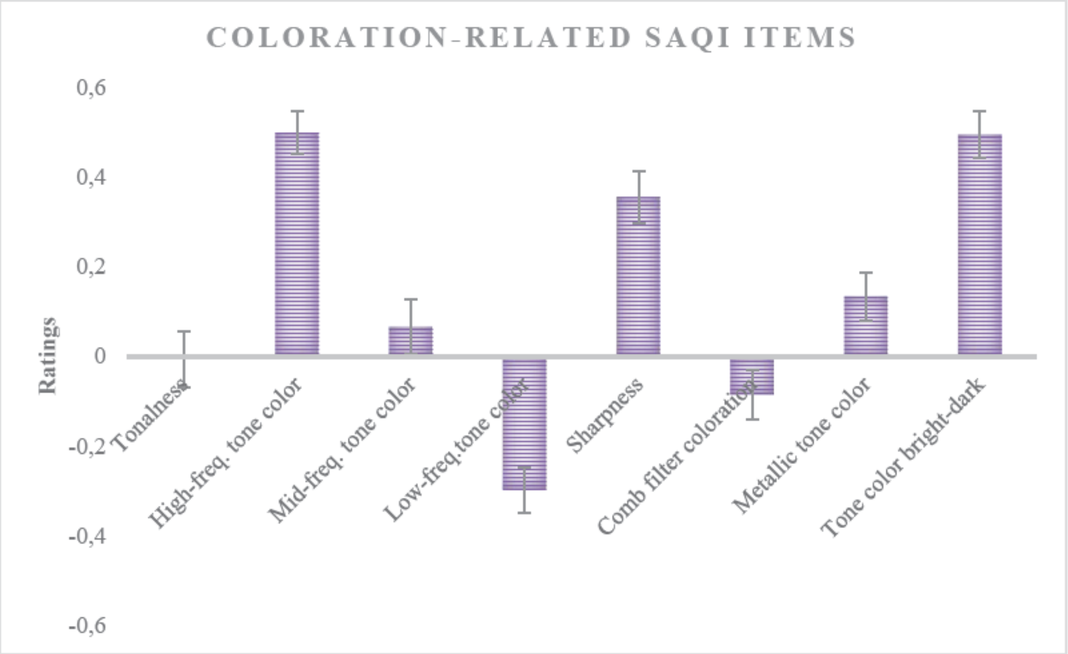


Figure 36: Similarity test for measured and modeled HRTFs – coloration-related SAQI items with error indicators

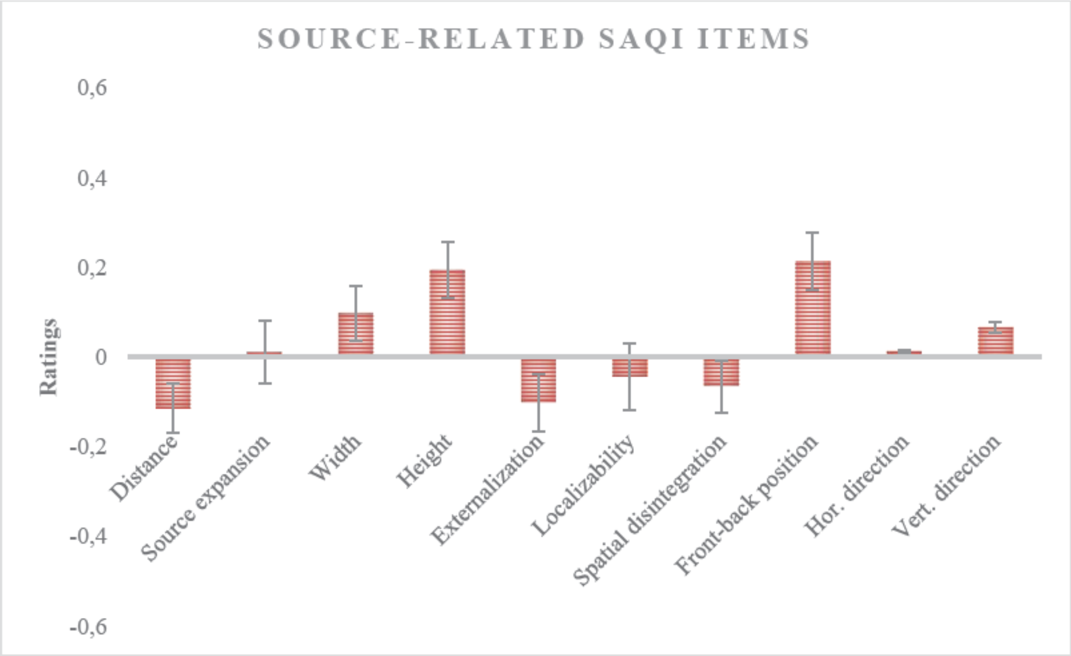


Figure 37: Similarity test for measured and modeled HRTFs – source-related SAQI items with error indicators

4 Conclusion and Discussion

4.1 BEM calculated HRTFs

A comprehensive data set containing measured and modeled HRTFs for 93 subjects was created in this work. HRTFs were acoustically measured and BEM calculated based on 3D meshes. One advantage of calculating HRTFs, especially for subjects is that the 3D scanning procedure might be less uncomfortable than acoustical measurements in which subjects must sit still for a longer period of time with microphones inserted into both ear canals. On the other hand, there are quite a few challenging steps following 3D scanning that must be carried out before the subject's HRTF is attained. For example, 3D mesh preparation can be quite comprehensive including mesh cleaning, mesh combination, mesh-alignment and remeshing. Semi-automatic working steps using Python scripts for mesh alignment and material assignment were developed throughout this work and helped to generate a quicker and more reproducible work flow. Faster options of BEM calculations, namely SL-FMM and ML-FMM could not be used due to a bug. Hence, BEM calculation of HRTFs for the left and right pinna took approximately 40 days for all 93 subjects.

Cross-validation of measured and modeled HRTFs showed differences between these two versions for some subjects, but an overall comparison showed moderate median ERB errors. Localization performance evaluation using the model of Baumgartner et. al (2014) revealed higher PE and QE values for modeled HRTFs. The median PE of the measured data set was 35° and therefore just slightly above the general sensitivity of ± 30°. The PE for the modeled data set was noticeably higher with 42°. This suggests that the measured HRTFs represent true HRTF better than modeled HRTFs.

Another aspect that points in a similar direction was the comparison of rating means with paired-sample *t*-tests. It was shown that differences between one's own HRTF and another person's HRTF are being rated higher when listening to the measured HRTF set. The effect strength of these findings was medium for *Difference, Coloration* and *Source position* but weak for *Externalization.* These findings suggest that differences between one's own and another person's HRTF were perceived more distinct using HRTFs from the measured set. This again could result from measured HRTFs representing the true HRTFs of participants more accurately.

However, the similarity test in the second part of the listening test which inquired this topic by letting all subjects compare their own modeled and their own measured HRTF contradicts the above conclusion. The test showed no difference in rated *Degree-of-liking* and *Naturalness* for measured and modeled HRTFs. A clear difference between the two versions can be perceived, mainly in tone-color-related SAQI items. It appears that in modeled HRTFs high frequencies are more emphasized and that they are also perceived sharper than their measured correspondences. It was speculated that despite an SH order equalization of measured HRTFs, modeled HRTFs show more energy in higher frequency areas resulting from the measurement methods. This might also be the cause for the finding that modeled HRTFs are perceived with higher *Loudness* and more *Crispness* and that also *Speech intelligibility* is slightly better. A clear overall *Difference* between modeled and measured HRTFs can be seen, while *Source position* varies slightly, and the source is perceived wider and higher with a modeled HRTF.

4.2 HRTF individualization using anthropometric key features

4.2.1 Explained variance

This work provides a list of anthropometric key features that were derived from tables of effect sizes for 4 different SAQI items. These SAQI items were selected owing to their good representation of relevant qualities in terms of binaural hearing. To search for the influence of anthropometrics and the perceived fit of a certain HRTF, the listener's AFs HRTF owner's AFs were set into relationship. The mere difference between corresponding AFs showed to be the transformation yielding the best results in terms of explained variance.

Explainable variance differs between different SAQI items and among data sets. For SAQI items *Difference* and *Coloration* $R^2_{\text{conditional}}$ values of .45 to .48 could be achieved. That most variance could be explained for these qualities, is probably due to their good perceivability. Especially differences in *Coloration* are very easy to rate using pink noise, which was reported by many participants. For *Source position* also, a large proportion of variance could be explained but values differ between the measured ($R^2_{\text{conditional}} = .39$) and modeled data set ($R^2_{\text{conditional}} = .47$). For *Source position* the overall lowest MSE could be achieved with modeled HRTFs. Exceptional low values of explained variance were achieved for *Externalization* with 26 to 28%. An explanation is again delivered by the participants feedback who reported that ratings for *Externalization* were very hard to give because a difference was hardly perceivable. Most participants reported that while one source felt very internalized and close to the head the other source was perceived almost the same. It can be assumed that the difficult perceivability in *Externalization* differences results from the absence of room reflections in HRIRs.

According to Cohen's definition of effect sizes the amount of explained variance for all SAQI items equates to a large effect (Cohen, 1988).

4.2.2 Recommender

An evaluation of HRTF individualization using the developed multilevel mixed-effects regression models was done to show how close a predicted best fitting HRTF is to the real best fit among a set of 15 HRTFs. Figure 34 showed that HRTF individualization based on

AFs using the developed models yields results that are quite promising. The selected HRTF was closer to the best case in the set than to the average rating. However, also room for optimization was shown because it is desirable that a recommender selects the HRTF which provides the real best fit.

For unknown data, i.e. a test set, it was shown that SVM-R delivers slightly better prediction results than multilevel models. However, the large gap between training and test set errors (measured $\text{MSE}_{\text{train}}$=0.021, MSE_{test}=0.050 for *Difference,* measured) despite 10-fold cross-validation suggests an overfitting of the training data. This indicates that the rating of HRTFs of a given data set is highly individual and generalization should be done carefully. However, a recommending system that uses SVM-R to predict perceptual differences for different HRTFs of a data set based on the relationship of the listener's AFs to AFs underlying to each HRTF in the data set could still lead to satisfactory results. SVM-R could predict ratings for *Difference, Coloration* and *Source position* and the HRTF with the best (i.e. the lowest) rating could be chosen from the data set.

4.2.3 Hierarchical structure

Rating of SAQI items is an extremely individual process, which can be seen at the significant enhancement of models by adding random intercepts. Interestingly, this effect appears much stronger in the measured data set. The deviance difference χ^2_{diff} between models with and without intercept is much bigger for the measured data set which can be seen in Table 9. Besides, the difference between R^2_{marginal} (describing the model without random effects) and $R^2_{\text{conditional}}$ (describing the full mixed model) is much higher for the measured HRTF data set. Figure 33 and Table 7 showed that on average, ratings for measured HRTFs are significantly higher than for modeled HRTFs. It therefore appears that individuality in rating plays a higher role when higher differences are perceived than in cases with lower differences. In other words, subjects seem to agree on how to assess lower differences, but higher differences are more likely to be rated individualistically.

4.2.4 Effect sizes and top predictors

Calculation of effect sizes for AFs in multilevel mixed-effects regression models for different SAQI items was an important step to identify key predictors. For this purpose, it was useful that subjects rated the same HRTFs in two different versions. Consistencies could be identified among top 10 predictors for each SAQI item.

For perceived overall *Difference*, cavum concha height and cymba concha height are of great importance. For *Coloration*, cavum concha width, cymba concha height and the pinna rotation angle play a major role. *Coloration* between different HRTFs was reported to be very noticeable, which is why it can be assumed that for the *Difference* ratings, often

Coloration played a role. Reflection and resonances caused by the cavum concha are extensively described by varies publications (Hebrank and Wright, 1974; Lopez-Poveda and Meddis, 1996). It is therefore not surprising to find said feature among the list of top predictors.

For *Source position*, the pinna rotation angle, the pinna flare angle and two non-pinna related features are important, namely the subject's height and the head depth. That head and torso-related measures are vital for *Source position* is not surprising because, as discussed earlier, for horizontal localization ITD plays a major role. ITD is highly determined by the inter-tragus distance (Algazi et al., 2001b) which strongly correlates to head width. Interestingly, head width only takes the 12th position among sorted effect sizes in Table 14 but both, head depth and the subject's height significantly correlate with head width (see Table 18). Besides that, the subject's height correlates significantly with 14 other AFs.

For *Externalization*, rating differences was reported to be very ambivalent, as mentioned earlier. The top predictors are cavum concha width, pinna width and two features describing the pinna relative to the head, namely rotation angle and the pinna's offset down.

Table 16 finally presented general effect sizes of AFs for HRTF individualization which were attained by combining the results of all SAQI items and data sets. The above described characteristics can be found here as well, meaning that mainly pinna-related features are among the top 10 predictors. Sorted by effect size, the following top predictors could be identified: pinna rotation angle, cavum concha width, cymba concha height, cavum concha height, pinna flare angle, height, fossa height, pinna width, pinna height and head depth. These findings are in basic agreement with other studies that were mentioned under 1.2.2. Under 3.12 it was already discussed that the important roles of the cavum concha and the pinna angles were already shown in previous studies. Cymba concha height had not been identified as a top predictor in any of the studies that were reviewed by Ghorbal and Auclair (2017). This is somewhat surprising, because cymba concha height is the third most important feature for HRTF individualization of this work.

This work can contribute a ranked list of key features to already existing findings. Results suggest that key predictors for HRTF individualization must most importantly include features describing the pinnae relative to the head, such as the pinna angles. Secondly, features describing the fine structures of the pinnae, especially the cavum concha must be included. Thirdly features describing the entire auricle, such as the pinna width should be included. Lastly, the subject's height and the head depth seem to deliver extensive information for non-pinna related effects. The results also show that most torso and neck-related features play a minor role in the individualization of HRTFs.

It should however be noted that top predictors and effect sizes must be viewed with caution because results seem to be very dataset dependent. In *Coloration*, the second most important predictor only appears as the second last predictor within the other HRTF set. Also, it was shown that adding feature interactions and squared features significantly enhanced regression models. This shows that anthropometry has complicated dependencies and nonlinearities, and that attempting oversimplification could lead to unrealistic results. Also, the frequent occurrence of strong intercorrelation between AFs show that an identification of single AFs as key features can only be considered as a simplification. Statistical analysis of the relationship between perceived HRTF differences and various AFs suggests that pinna-related spectral behavior is complicated and individualized.

4.3 Possible sources of error

Previous studies showed that results can differ for different HRTF measurements even for the same subject (Majdak et al., 2013). Reasons for this can be a slightly different position of the microphone in the ear canal or a microphone which changes its position during measurement. It is therefore unavoidable that differences between the amplitude spectra of measured and modeled HRTFs occur.

In this work, the biggest challenge during mesh preparation was the decision of selecting ear canal entrances which then determined where microphone elements for the reciprocal calculation were being placed. Although a careful examination of each mesh was carried out, it cannot be assured that the exact position has always been hit. This was because ear canals sometimes were hard to identify in the 3D meshes. Earlier research concerning this topic has shown that the exact element selected as microphone element does not matter as long as the element is in the ear canal (Ziegelwanger et al., 2015b). However, this cannot always be guaranteed. Consequently, there are several causes for differences between measured and modeled HRTFs or the difference of any version of these HRTFs to the true HRTF of a person.

Another aspect which must be considered as a possible source of error is the removal of the torso for reasons which were stated earlier (see 2.4). The torso shapes the HRTF mainly in frequency regions below 3.5kHz (Katz, 2001a) through elevation dependent torso reflections (Algazi et al., 2001a, Algazi et al., 2002) and therefore should not be neglected. At must be assumed that this aspect also adds to a deviation between measured and modeled HRTFs.

No reliability test was included during the listening test to assess if participants' ratings would be given similarly another time. It can be assumed that some of the unexplained variance stems from ratings in which participants were not sure how exactly they should rate. Especially for small differences between HRTFs this could be the case, as well as for the SAQI item *Externalization* in general, for which it has been repeatedly reported that ratings were difficult to give. Also, order effects which might play a role were not considered due to the listening test design, in which each participant rated different HRTFs.

For SAQI items *Difference, Coloration* and *Source position* almost 50% of variance in ratings could be explained through AFs as predictors. This also means that the models could not account for at least half of the variance. This means that general statements about AFs for HRTF individualization must be given cautiously.

Adding AF squares and simple interaction terms enhanced the models but also limitations due to lacking degrees of freedom were reached. The data sets contain 1260 ratings and after adding AF squares and interactions already 1276 predictors were included. Hence, the available degrees of freedom, which are N-1 are already outnumbered. Stepwise selection of predictors into regression models quickly reduced the number of parameters. But it is obvious that much more ratings would be needed to analyze the complicated relationship of AFs to HRTFs on a perceptual basis in order to attain findings that can be generalized.

4.4 Future work

Modeled HRTFs were BEM calculated with simplified meshes, not containing torsos, and therefore, elevation dependent torso reflections were neglected. A parametric model to include torso influence could improve BEM modeled HRTFs, especially in lower frequency regions.

Statistical analysis showed that the relationship between AFs and the perceived degree of similarity between one's own and another person's HRTF is not straight forward but rather interdependent and complicated. Regression models were enhanced by adding squared terms and by including simple interaction between each term. It can be assumed that when triple interaction and a higher polynomial order of transformation is added to AFs, even more variance could be explained. However, to achieve this, much more data in form of listening test ratings would be needed to retrieve a higher number of degrees of freedom.

Further analysis with the available data of this work could be possible using other statistical approaches. As it was mentioned before, some research used PCA to create factors within AFs to reduce dimensions. While this work tried SVM-R, other machine learning techniques could also be applied to the underlying data. Cluster analysis or the implementation of a bagged tree algorithm could be among promising approaches.

Finally, a subsequent listening test could be carried out in which the best fit of the 93 HRTFs of this work could be selected for participants based on the results of this work. A localization test could then provide useful insights for the effectiveness of HRTF individualization based on AFs.

Bibliography

AES Standards Comittee (2015): „AES69-2015: AES standard for file exchange - Spatial acoustic data file format.“ In: (Audio Engineering Society).

Algazi, V. Ralph; Carlos Avendano and Richard O. Duda (2001a): „Elevation localization and head-related transfer function analysis at low frequencies.“ In: *The Journal of the Acoustical Society of America*, 109(3), pp. 1110–1122.

Algazi, V. Ralph; Richard O. Duda; Ramani Duraiswami; et al. (2002): „Approximating the head-related transfer function using simple geometric models of the head and torso.“ In: *The Journal of the Acoustical Society of America*, 112(5), pp. 2053–2064.

Algazi, V.R.; R.O. Duda; D.M. Thompson; et al. (2001b): „The CIPIC HRTF database.“ IEEE, pp. 99–102.

Batteau, D. W. (1967): „The role of the pinna in human localization.“ In: *Proc. R. Soc. Lond. B*, 168(1011), pp. 158–180.

Baumgartner, Robert; Piotr Majdak and Bernhard Laback (2014): „Modeling sound-source localization in sagittal planes for human listeners.“ In: *The Journal of the Acoustical Society of America*, 136(2), pp. 791–802.

Ben-Hur, Zamir; Fabian Brinkmann; Jonathan Sheaffer; et al. (2017): „Spectral equalization in binaural signals represented by order-truncated spherical harmonics.“ In: *The Journal of the Acoustical Society of America*, 141(6), pp. 4087–4096.

Bernschütz, Benjamin (2016): „Microphone arrays and sound field decomposition for dynamic binaural recording.“ Technische Universität Berlin.

Bernschütz, S., B.; Pörschmann, C.; Spors, S.; Weinzierl (2011): „SOFiA Sound Field Analysis Toolbox.“ *Proc. of the International Conference on Spatial Audio*, Vortrag auf der Detmold.

Blauert, J. (1997): *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press.

Bomhardt, R. (2017): *Anthropometric Individualization of Head-Related Transfer Functions Analysis and Modeling*. Logos Verlag Berlin, (Aachener Beiträge zur Akustik).

Bomhardt, Ramona; Matias de la Fuente Klein and Janina Fels (2016): „A high-resolution head-related transfer function and three-dimensional ear model database.“ pp. 050002.

Brinkmann, Fabian; Alexander Lindau; Stefan Weinzerl; et al. (2017): „A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations.“ In: *Journal of the Audio Engineering Society*, 65(10), pp. 841–848.

Brinkmann, Fabian and Stefan Weinzierl (2017): „AKtools—An Open Software Toolbox for Signal Acquisition, Processing, and Inspection in Acoustics.“ *Audio Engineering Society Convention 142*,.

Bronkhorst, Adelbert W. (1995): „Localization of real and virtual sound sources.“ In: *The Journal of the Acoustical Society of America*, 98(5), pp. 2542–2553.

Ciba, Simon; André Wlodarski and Hans-Joachim Maempel (2009): „WhisPER – A New Tool for Performing Listening Tests.“ *Audio Engineering Society Convention 126*,.

Cohen, Jacob (1988): *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, N.J: L. Erlbaum Associates.

Dinakaran, Manoj; Fabian Brinkmann; Stine Harder; et al. (2018): „Perceptually motivated analysis of numerically simulated head-related transfer functions generated by various 3D surface scanning systems.“ Vortrag auf der ICASSP.

Dinakaran, Manoj; Peter Grosche; Fabian Brinkmann; et al. (2016): „Extraction of Anthropometric Measures from 3D-Meshes for the Individualization of Head-Related Transfer Functions.“ *Audio Engineering Society Convention 140*,.

Fahrmeir, Ludwig; Christian Heumann; Rita Künstler; et al. (2016): *Statistik: Der Weg zur Datenanalyse*. 8. Aufl. Springer Spektrum, (Springer-Lehrbuch).

Fels, Janina and Michael Vorländer (2009): „Anthropometric parameters influencing head-related transfer functions.“ In: *Acta Acustica united with Acustica*, 95(2), pp. 331–342.

Field, Andy P. (2009): *Discovering statistics using SPSS: and sex, drugs and rock „n“ roll*. 3rd ed. Los Angeles: SAGE Publications.

Fuß, Alexander; Fabian Brinkmann; Thomas Jürgensohn; et al. (2015): „Ein vollsphärisches Multikanalmesssystem zur schnellen Erfassung räumlich hochaufgelöster, individueller kopfbezogener Übertragungsfunktionen.“ *Fortschritte der Akustik – DAGA 2015*,.

Geier, Matthias; Jens Ahrens and Sascha Spors (2008): „The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods.“ *Audio Engineering Society Convention 124*,.

Ghorbal, Slim and Théo Auclair (2017): „Pinna Morphological Parameters Influencing HRTF Sets.“ In: pp. 7.

Hall, Judith G (2007): *Handbook of physical measurements*. Oxford; New York: Oxford University Press.

Hartmann, William M. (1999): „How We Localize Sound.“ In: *Physics Today*, 52(11), pp. 24–29.

Hastie, Trevor; Robert Tibshirani and Jerome Friedman (2009): *The Elements of Statistical Learning*. New York, NY: Springer New York, (Springer Series in Statistics).

Hebrank, Jack and D. Wright (1974): „Spectral cues used in the localization of sound sources on the median plane.“ In: *The Journal of the Acoustical Society of America*, 56(6), pp. 1829–1834.

Hox, J.J.; M. Moerbeek and R. van de Schoot (2010): *Multilevel Analysis: Techniques and Applications, Second Edition*. Taylor & Francis, (Quantitative Methodology Series).

Katz, Brian F. G. (2001a): „Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation.“ In: *The Journal of the Acoustical Society of America*, 110(5), pp. 2440–2448.

Katz, Brian F. G. (2001b): „Boundary element method calculation of individual head-related transfer function. II. Impedance effects and comparisons to real measurements.“ In: *The Journal of the Acoustical Society of America*, 110(5), pp. 2449–2455.

Kreuzer, Wolfgang; Piotr Majdak and Zhengsheng Chen (2009): „Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range.“ In: *The Journal of the Acoustical Society of America*, 126(3), pp. 1280–1290.

LaHuis, David M.; Michael J. Hartman; Shotaro Hakoyama; et al. (2014): „Explained Variance Measures for Multilevel Models.“ In: *Organizational Research Methods*, 17(4), pp. 433–451.

Lebedev, V. I. (1977): „Spherical quadrature formulas exact to orders 25–29.“ In: *Siberian Mathematical Journal*, 18(1), pp. 99–107.

Lindau, Alexander and Fabian Brinkmann (2012): „Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings.“ In: *J. Audio Eng. Soc*, 60(1/2), pp. 54–62.

Lindau, Alexander; Vera Erbes; Steffen Lepa; et al. (2014): „A Spatial Audio Quality Inventory (SAQI).“ In: *Acta Acustica united with Acustica*, 100(5), pp. 984–994.

Lindau, Alexander and Stefan Weinzierl (2006): „FABIAN - An Instrument for Software-Based Measurement of Binaural Room Impulse Responses in Multiple Degrees of Freedom.“ *Proceedings of the VDT International Convention*,.

Liu, Xuejie and Xiaoli Zhong (2016): „An improved anthropometry-based customization method of individual head-related transfer functions.“ IEEE, pp. 336–339.

Lopez-Poveda, E. A. and R. Meddis (1996): „A physical model of sound diffraction and reflections in the human concha.“ In: *The Journal of the Acoustical Society of America*, 100(5), pp. 3248–3259.

Majdak, Piotr; Matthew J. Goupell and Bernhard Laback (2010): „3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training.“ In: *Attention, Perception, & Psychophysics*, 72(2), pp. 454–469.

Majdak, Piotr; Bruno Masiero and Janina Fels (2013): „Sound localization in individualized and non-individualized crosstalk cancellation systems.“ In: *The Journal of the Acoustical Society of America*, 133(4), pp. 2055–2068.

Masiero, Bruno; Dietrich Pascal; Martin Pollow; et al. (2012): „Design of a Fast Individual HRTF Measurement System.“ Vortrag auf der DAGA, Darmstadt.

Middlebrooks, John C. (1999a): „Individual differences in external-ear transfer functions reduced by scaling in frequency.“ In: *The Journal of the Acoustical Society of America*, 106(3), pp. 1480–1492.

Middlebrooks, John C. (1999b): „Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency.“ In: *The Journal of the Acoustical Society of America*, 106(3), pp. 1493–1510.

Møller, Henrik (1992): „Fundamentals of binaural technology.“ In: *Applied Acoustics*, 36(3–4), pp. 171–218.

Møller, Henrik; Michael Friis Sørensen; Clemen Boje Jensen; et al. (1996): „Binaural Technique: Do We Need Individual Recordings?“ In: *J. Audio Eng. Soc*, 44(6), pp. 451–469.

Moorrees, Coenraad F.A. (1994): „Natural head position—a revival.“ In: *American Journal of Orthodontics and Dentofacial Orthopedics*, 105(5), pp. 512–513.

Nakagawa, Shinichi and Holger Schielzeth; Robert B. O’Hara (Hrsg.) (2013): „A general and simple method for obtaining R^2 from generalized linear mixed-effects models.“ In: *Methods in Ecology and Evolution*, 4(2), pp. 133–142.

Nicol, R. (2010): *Binaural Technology*. Audio Engineering Society.

Peugh, James L. (2010): „A practical guide to multilevel modeling.“ In: *Journal of school psychology*, 48(1), pp. 85–112.

Rafaely, Boaz (2015): *Fundamentals of Spherical Array Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, (Springer Topics in Signal Processing).

Ramírez, Dayhanara Martínez; Joaquín Canseco Jiménez; Enrique González Ramírez; et al. (2013): „Discrepancies in cephalometric measurements in relation to natural head position.“ In: *Revista Mexicana de Ortodoncia*, 1(1), pp. e27–e32.

Takemoto, Hironori; Parham Mokhtari; Hiroaki Kato; et al. (2012): „Mechanism for generating peaks and notches of head-related transfer functions in the median plane.“ In: *The Journal of the Acoustical Society of America*, 132(6), pp. 3832–3841.

Thurlow, Willard R.; John W. Mangels and Philip S. Runge (1967): „Head Movements During Sound Localization.“ In: *The Journal of the Acoustical Society of America*, 42(2), pp. 489–493.

Turku, Julia; Asta Kärkkäinen; Leo Kärkkäinen; et al. (2008): „Perceptual Evaluation of Numerically Simulated Head-Related Transfer Functions.“ *Audio Engineering Society Convention 124*,.

Williams, Earl G. (1999): *Fourier acoustics: sound radiation and nearfield acoustical holography*. San Diego, Calif.: Academic Press.

Xie, Bosun (2013): *Head-related transfer function and virtual auditory display*. Second edition. Plantation, FL: J. Ross Publishing.

Xu, S.; Z. Z. Li; L. Zeng; et al. (2007): „A study of morphological influence on head-related transfer functions.“ IEEE, pp. 472–476.

Yang, Lin; Longyu Zhang; Haiwei Dong; et al. (2015): „Evaluating and Improving the Depth Accuracy of Kinect for Windows v2.“ In: *IEEE Sensors Journal*, 15(8), pp. 4275–4285.

Ziegelwanger, Harald; Wolfgang Kreuzer and Piotr Majdak (2015a): „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 (ICSV)*, Vortrag auf der Florence, Italy ICSV 2015, pp. 1–8.

Ziegelwanger, Harald; Wolfgang Kreuzer and Piotr Majdak (2016): „A priori mesh grading for the numerical calculation of the head-related transfer functions.“ In: *Applied Acoustics*, 114, pp. 99–110.

Ziegelwanger, Harald; Piotr Majdak and Wolfgang Kreuzer (2015b): „Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization.“ In: *The Journal of the Acoustical Society of America*, 138(1), pp. 208–222.

Ziegelwanger, Harald; Andreas Reichinger and Piotr Majdak (2013): „Calculation of listener-specific head-related transfer functions: Effect of mesh quality.“ pp. 050017–050017.

Zotkin, D.N.; R. Duraiswami; L.S. Davis; et al. (2002): „Virtual audio system customization using visual matching of ear parameters.“ IEEE Comput. Soc, pp. 1003–1006.

Zotkin, D.Y.N.; J. Hwang; R. Duraiswaini; et al. (2003): „HRTF personalization using anthropometric measurements.“ IEEE, pp. 157–160.

List of Figures

List of Tables

Appendix

Pearson correlation coefficients of anthropometric features with each other N = 95

Table 18: Pearson correlation results of anthropometric features.

Note: N=95.

**. Correlation is significant at the 0.01 level (1-tailed).

*. Correlation is significant at the 0.05 level (1-tailed).

AF	x1	x2	x3	x4	x5	x6	x7	x8	x9	x12	x14
x1	1	.468**	.372**	-.063	.205*	.764**	-.151	.573**	.493**	.551**	.246**
x2	.468**	1	.502**	.042	.138	.358**	-.160	.354**	.337**	.337**	.405**
x3	.372**	.502**	1	-.083	.174*	.411**	.095	.333**	.209*	.308**	.490**
x4	-.063	.042	-.083	1	-.072	-.048	.092	-.045	-.067	-.098	-.089
x5	.205*	.138	.174*	-.072	1	.186*	-.020	.120	.201*	.183*	.140
x6	.764**	.358**	.411**	-.048	.186*	1	-.068	.671**	.543**	.694**	.324**
x7	-.151	-.160	.095	.092	-.020	-.068	1	-.292**	-.071	-.062	.275**
x8	.573**	.354**	.333**	-.045	.120	.671**	-.292**	1	.381**	.558**	.213*
x9	.493**	.337**	.209*	-.067	.201*	.543**	-.071	.381**	1	.706**	.420**
x12	.551**	.337**	.308**	-.098	.183*	.694**	-.062	.558**	.706**	1	.468**
x14	.246**	.405**	.490**	-.089	.140	.324**	.275**	.213*	.420**	.468**	1
x16	-.018	.077	.103	.119	-.004	-.023	-.011	-.139	-.039	.000	.111
x17	.661**	.444**	.374**	-.071	.222*	.790**	-.067	.689**	.735**	.767**	.510**
d1	.214*	.310**	.347**	-.004	.094	.326**	-.152	.339**	.205*	.313**	.236*
d2	.191*	.186*	.175*	-.042	.106	.171*	-.093	.129	.095	.237*	.182*
d3	.102	.016	.126	.137	.089	.256**	-.044	.284**	.204*	.319**	.096
d4	.201*	.124	.139	.003	.098	.366**	.064	.341**	.155	.262**	.204*
d5	.425**	.231*	.386**	.002	.133	.568**	-.002	.577**	.405**	.488**	.251**
d6	.200*	.224*	.356**	.073	.188*	.350**	.103	.312**	.239**	.220*	.218*
θ1	.182*	-.063	.120	.035	-.193*	.322**	.213*	.171*	.154	.169	.127
θ2	-.250**	-.126	-.052	.085	-.189*	-.194*	.089	-.235*	-.263**	-.168	-.095

Table 18 continuation: Pearson correlation of AFs

	x16	x17	d1	d2	d3	d4	d5	d6	θ1	θ2
x1	-.018	.661**	.214*	.191*	.102	.201*	.425**	.200*	.182*	-.250**
x2	.077	.444**	.310**	.186*	.016	.124	.231*	.224*	-.063	-.126
x3	.103	.374**	.347**	.175*	.126	.139	.386**	.356**	.120	-.052
x4	.119	-.071	-.004	-.042	.137	.003	.002	.073	.035	.085
x5	-.004	.222*	.094	.106	.089	.098	.133	.188*	-.193*	-.189*
x6	-.023	.790**	.326**	.171*	.256**	.366**	.568**	.350**	.322**	-.194*
x7	-.011	-.067	-.152	-.093	-.044	.064	-.002	.103	.213*	.089
x8	-.139	.689**	.339**	.129	.284**	.341**	.577**	.312**	.171*	-.235*
x9	-.039	.735**	.205*	.095	.204*	.155	.405**	.239**	.154	-.263**
x12	.000	.767**	.313**	.237*	.319**	.262**	.488**	.220*	.169	-.168
x14	.111	.510**	.236*	.182*	.096	.204*	.251**	.218*	.127	-.095
x16	1	-.089	.143	.142	-.135	-.178*	-.040	-.061	-.094	.127
x17	-.089	1	.396**	.186*	.222*	.317**	.578**	.377**	.207*	-.343**
d1	.143	.396**	1	.088	.031	.336**	.496**	.229*	-.137	.231*
d2	.142	.186*	.088	1	.094	-.041	.249**	.138	-.084	-.074
d3	-.135	.222*	.031	.094	1	.388**	.327**	.405**	.198*	-.125
d4	-.178*	.317**	.336**	-.041	.388**	1	.572**	.367**	.069	.206*
d5	-.040	.578**	.496**	.249**	.327**	.572**	1	.539**	.257**	-.078
d6	-.061	.377**	.229*	.138	.405**	.367**	.539**	1	.016	-.135
θ1	-.094	.207*	-.137	-.084	.198*	.069	.257**	.016	1	-.271**
θ2	.127	-.343**	.231*	-.074	-.125	.206*	-.078	-.135	-.271**	1

Regression Coefficients for SAQI item *Difference*

Table 19: regression coefficients for *Difference* with interactions from the measured data set
Note: R^2 =.47, R^2 adj. = .42, *p < .05, **p < .01

	Unstandardized Coefficients		*Standardized Coefficients*
Predictor	**B**	**Std. Error**	**β**
(Constant)	0.29	0.01	
x1_diff_centered	-0.01	0.01	-.04
x2_diff_centered	0.04	0.01	.21**
x3_diff_centered	-0.04	0.01	-.17**
x4_diff_centered	-0.06	0.03	-.07
x5_diff_centered	-0.03	0.02	-.05
x6_diff_centered	0.00	0.01	-.01
x7_diff_centered	-0.01	0.01	-.05
x8_diff_centered	-0.03	0.01	-.22**
x9_diff_centered	-0.02	0.00	-.29**
x12_diff_centered	0.01	0.00	.23**
x14_diff_centered	0.00	0.00	.13**
x16_diff_centered	0.00	0.00	.18**
x17_diff_centered	0.00	0.00	.01
L_d1_diff_centered	0.16	0.07	.15*
L_d2_diff_centered	-0.38	0.07	-.26**
L_d3_diff_centered	0.05	0.05	.05
L_d4_diff_centered	0.11	0.05	.16*
L_d5_diff_centered	0.02	0.05	.05
L_d6_diff_centered	0.01	0.04	.02
L_θ1_diff_centered	0.01	0.00	.19**
L_θ2_diff_centered	0.00	0.00	.09
R_d1_diff_centered	-0.04	0.09	-.03
R_d2_diff_centered	-0.04	0.08	-.02
R_d3_diff_centered	0.06	0.05	.07
R_d4_diff_centered	-0.26	0.06	-.37**
R_d5_diff_centered	0.09	0.05	.19
R_d6_diff_centered	-0.02	0.05	-.02
R_θ1_diff_centered	-0.01	0.00	-.22**
R_θ2_diff_centered	0.00	0.00	-.14**
x14_diff_c_square	0.00	0.00	.10**
L_d2_diff_c_square	0.86	0.21	.15**
x7_diff_c_square	0.00	0.00	-.02
L_d5_diff_c_square	-0.04	0.02	-.07
R_d4_diff_c_square	0.35	0.08	.27**
L_d4_diff_c_square	-0.06	0.07	-.05
x16_diff_c_square	0.00	0.00	-.04
L_d1_diff_c_square	0.11	0.15	.03
R_d1_diff_c_square	0.28	0.23	.08

x3_diff_c_square	0.00	0.00	-.02
x1_diff_c_square	0.01	0.01	.04
x8_diff_c_square	0.01	0.00	.17*
x7_R_d6_inter	0.07	0.01	.18**
x7_L_d5_inter	0.00	0.01	-.01
L_d2_R_d1_inter	-0.18	0.24	-.03
x3_R_d2_inter	-0.09	0.05	-.06
x8_L_d4_inter	-0.01	0.01	-.02
x9_L_d3_inter	0.02	0.01	.08**
R_d1_R_d4_inter	-0.58	0.15	-.22**
x3_x17_inter	0.00	0.00	.14**
R_d2_R_θ1_inter	0.00	0.01	.01
x5_L_d5_inter	-0.06	0.03	-.06*
R_d2_R_d4_inter	0.58	0.15	.13**
x3_x12_inter	0.00	0.00	-.08
x4_R_d3_inter	-0.19	0.09	-.06*
x1_x9_inter	0.01	0.00	.28**
x14_R_d5_inter	-0.01	0.00	-.32**
x14_R_d1_inter	0.02	0.00	.20**
L_d1_R_θ2_inter	0.01	0.00	.11**
x2_R_d6_inter	0.06	0.02	.10**
L_d2_R_d3_inter	-0.97	0.24	-.18**
L_d2_L_d3_inter	0.54	0.24	.10*
x6_x9_inter	-0.01	0.00	-.21**
x17_R_θ1_inter	0.00	0.00	.17**
x8_R_d3_inter	-0.07	0.02	-.15**
x1_x17_inter	0.00	0.00	-.17*
x7_x12_inter	0.00	0.00	-.08**
x8_R_d2_inter	0.12	0.03	.13**
x12_R_d6_inter	0.03	0.01	.23**
x7_R_d3_inter	0.05	0.02	.09**
x7_R_θ1_inter	0.00	0.00	-.08**
x1_L_θ1_inter	0.01	0.00	.19**
x1_x6_inter	-0.05	0.01	-.29**
x1_x14_inter	0.00	0.00	.09*
x1_R_θ1_inter	0.00	0.00	-.12*
x12_L_d6_inter	-0.02	0.01	-.18**
x2_L_d2_inter	-0.11	0.04	-.10**
x9_L_d2_inter	0.02	0.01	.08*
x6_x8_inter	0.01	0.01	.16
x8_square_R_d1_square_inter	-0.06	0.02	-.19**
L_d6_square_R_d2_square_inter	-2.67	1.14	-.07*
R_d3_square_R_d6_square_inter	0.53	0.24	.06*
L_θ1_square_L_θ2_square_inter	0.00	0.00	-.08*
x4_square_x8_square_inter	0.03	0.01	.09**
x4_square_L_d6_square_inter	-0.58	0.31	-.06
x6_square_x8_square_inter	0.00	0.00	.10

x3_square_L_θ2_square_inter	0.00	0.00	.07*
x9_diff_L_d2_square_inter	0.00	0.03	.01
x7_diff_x17_square_inter	0.00	0.00	.15**
x7_diff_x12_square_inter	0.00	0.00	-.10*
x4_diff_R_d2_square_inter	2.01	0.61	.12**
x7_diff_R_d2_square_inter	0.35	0.10	.11**
x6_diff_R_d2_square_inter	0.31	0.19	.07
x14_diff_R_d2_square_inter	-0.06	0.02	-.13**
x12_diff_R_d2_square_inter	0.12	0.05	.13*
x7_diff_R_d3_square_inter	-0.09	0.04	-.07*
x2_diff_R_d1_square_inter	0.27	0.08	.15**

Multilevel mixed-effects model estimates for SAQI item *Difference*

Table 20: multilevel mixed-effects model for Difference (measured data set)

Note: $R^2_{conditional}$ =.47. $R^2_{marginal}$ = .37, *p < .05, **p < .01

			95% CI	
Name	b	SE b	Lower	Lower
(Intercept)	.29**	0.02	0.26	0.32
x1_diff	-.01	0.01	-0.04	0.01
x2_diff	.04**	0.01	0.02	0.05
x3_diff	-.03**	0.01	-0.05	-0.01
x4_diff	-.06	0.04	-0.13	0.01
x5_diff	-.03	0.02	-0.07	0.00
x6_diff	.02	0.01	-0.01	0.05
x7_diff	-.00	0.01	-0.02	0.01
x8_diff	-.02**	0.01	-0.04	-0.01
x9_diff	-.01**	0.00	-0.02	0.00
x12_diff	.00	0.00	0.00	0.01
x14_diff	.00	0.00	0.00	0.00
x16_diff	.00**	0.00	0.00	0.00
x17_diff	-.00	0.00	0.00	0.00
L_d1_diff	.09	0.08	-0.07	0.26
L_d2_diff	-.30**	0.09	-0.47	-0.13
L_d3_diff	.05	0.05	-0.06	0.15
L_d4_diff	.11*	0.06	0.00	0.22
L_d5_diff	.04	0.06	-0.07	0.15
L_d6_diff	-.07	0.05	-0.16	0.03
L_θ1_diff	.00*	0.00	0.00	0.01
L_θ2_diff	.00	0.00	0.00	0.00
R_d1_diff	-.09	0.10	-0.29	0.11
R_d2_diff	-.02	0.09	-0.21	0.16
R_d3_diff	.02	0.06	-0.10	0.14
R_d4_diff	-.24**	0.06	-0.36	-0.12
R_d5_diff	.05	0.06	-0.06	0.16
R_d6_diff	.06	0.05	-0.05	0.16
R_θ1_diff	-.01**	0.00	-0.01	0.00
R_θ2_diff	-.00	0.00	0.00	0.00
x1_diff_square	.01	0.01	-0.01	0.03
x8_diff_square	.01**	0.00	0.01	0.02
x14_diff_square	.00	0.00	0.00	0.00
L_d2_diff_square	.70**	0.18	0.35	1.05
R_d4_diff_square	.23**	0.05	0.12	0.33
x1_x6_inter	-.05**	0.01	-0.07	-0.03
x1_x9_inter	.01**	0.00	0.00	0.01
x1_x14_inter	.00**	0.00	0.00	0.00
x1_L_θ1_inter	.01**	0.00	0.00	0.01
x1_R_θ1_inter	-.00	0.00	-0.01	0.00
x2_L_d2_inter	-.12**	0.03	-0.18	-0.05

x2_R_d6_inter	.05**	0.02	0.02	0.08
x3_x17_inter	.00	0.00	0.00	0.00
x4_R_d3_inter	-.16	0.08	-0.32	0.01
x5_L_d5_inter	-.08**	0.03	-0.13	-0.02
x6_x9_inter	-.00	0.00	-0.01	0.00
x7_x12_inter	-.00**	0.00	0.00	0.00
x7_R_d3_inter	.04*	0.02	0.01	0.07
x7_R_d6_inter	.06**	0.01	0.04	0.08
x7_R_θ1_inter	-.00**	0.00	0.00	0.00
x8_R_d2_inter	.10**	0.03	0.04	0.16
x8_R_d3_inter	-.05**	0.02	-0.08	-0.01
x9_L_d3_inter	.01*	0.01	0.00	0.02
x12_L_d6_inter	-.02**	0.01	-0.03	-0.01
x12_R_d6_inter	.02**	0.01	0.01	0.04
x14_R_d1_inter	.02**	0.00	0.01	0.02
x14_R_d5_inter	-.01**	0.00	-0.01	0.00
x17_R_θ1_inter	.00**	0.00	0.00	0.00
L_d1_R_θ2_inter	.01**	0.00	0.01	0.02
L_d2_R_d3_inter	-.48**	0.15	-0.77	-0.19
R_d1_R_d4_inter	-.31**	0.11	-0.54	-0.09
R_d2_R_d4_inter	.44**	0.13	0.18	0.69
x3_square_L_θ2_square_inter	.00*	0.00	0.00	0.00
x4_square_x8_square_inter	.02*	0.01	0.00	0.04
x4_square_L_d6_square_inter	-.52	0.28	-1.07	0.04
x8_square_R_d1_square_inter	-.03**	0.01	-0.06	-0.01
L_d6_square_R_d2_square_inter	-1.55	1.04	-3.59	0.49
L_θ1_square_L_θ2_square_inter	-.00	0.00	0.00	0.00
x2_diff_R_d1_square_inter	.24**	0.07	0.10	0.39
x4_diff_R_d2_square_inter	1.33*	0.56	0.22	2.43
x7_diff_x12_square_inter	-.00**	0.00	0.00	0.00
x7_diff_x17_square_inter	.00**	0.00	0.00	0.00
x7_diff_R_d2_square_inter	.28**	0.09	0.10	0.47
x12_diff_R_d2_square_inter	.14**	0.03	0.08	0.21
x14_diff_R_d2_square_inter	-.06**	0.02	-0.09	-0.03
x17_diff_L_d1_square_inter	-.02*	0.01	-0.03	0.00