Essentials on HRTF measurement and storage format standardization

Bachelor Thesis

Wolfgang
Hrauda

Supervisor: Dr. Sontacchi, Alois

Graz, June 14, 2013

institut für elektronische musik und akustik

Abstract

This bachelor thesis presents a detailed overview of the acquisition and storage of head-related transfer functions (HRTFs) and similar directional audio data, providing both theoretical basics and an examination of publicly available datasets and databases. The need for more and larger high-quality HRTF databases is mainly addressed in two ways: Firstly, knowledge is established, which allows to assess and enhance the quality of existing or future databases in terms of spatial resolution, optimization of the measurement process, choice of the impulse response measurement method and post-processing. Secondly, particular focus is laid on the possible creation of a standardized directional audio storage format. For sake of more universal applications of such a format, the scope is extended from HRTFs to general head-related directional audio data.

Contents

1 Introduction

When Baron Rayleigh claimed in 1876 that "the possibility of distinguishing a voice in front from a voice behind [...] depend[s] on the compound character of the sound in a way that is not easy to understand",[1] he certainly was not aware that more than 130 years later - although research efforts have made us understand a lot about the phenomenons of the human nature - spatial hearing is still a research field with mysteries yet to be discovered. While the basic principles related to the localization of a sound source seem to be rather clear as of 2013, modelling these principles and their perception and neuronal processing in our hearing sense remain a challenge to mankind's inquiring mind.

Since 1876, it has been understood that distinguishing sound sources in front from sound sources in the back can mainly be attributed to spectral differences depending on the source direction, which are the result of filtering with what is now referred to as "Head-related transfer functions" (HRTFs). In fact, HRTFs are a key component of spatial hearing and when creating so-called virtual auditory environments. All over the world, research institutes are measuring HRTFs or derivatives thereof and are utilizing them for research purposes such as modelling HRTFs based on the anthropometry of a person's head, torso and pinna. One obstacle in this process is that data exchange between the institutions is difficult as they all use different ways of measuring and storing their data. More uniform measurement setups with results stored in a standardized file format would thus be of great advantage. The purpose of this thesis is to examine current trends and possibilities regarding the acquisition and preservation of HRTFs and similar data under the perspective of the creation of a common storage format. Furthermore, some guidelines for high-quality measurements shall be given.

To this end, the role of HRTFs in spatial hearing and some basics regarding their mathematical and geometrical representation are shown in *Chapter 2* of this thesis. The actual measurement process is discussed, as well as the basic principles, advantages and disadvantages of the most important available impulse response measurement methods which are used to obtain the head-related impulse responses. In *Chapter 3*, an examination of existing databases, focusing on their geometrical setups, measurement and post-processing procedures and data storage, is presented. This includes the largest publicly available HRTF databases, but also binaural room impulse response (BRIR) measurements and some special cases which serve to show the diversity of existing setups. Finally, a critical summary of these insights is given in *Chapter 3* which also involves raising various issues to be considered when thinking about the quality of HRTF and related measurements and the standardization of the data storage format. As such a format entitled "Spatially Oriented Format for Acoustics" (SOFA) is being developed at the time of writing, its potential is showcased by the exemplary representation of a dataset in SOFA.

[1] From "Our perception of the direction of a source of sound" in Rayleigh, Nature, XIV. pp. 32-33, 1876.

Acknowledgements

Special thanks go out to my supervisor Alois Sontacchi who supported me with his comprehensive knowledge and experience in a way that fruitfully combined challenging and understanding. I would also like to thank Piotr Majdak who supervised my internship at the Acoustic Research Institute (ARI) where I was first introduced to this topic. Eventually, wholehearted thanks go out to every single person who supported and encouraged me during the process of writing this thesis.

2 Head-related transfer functions

2.1 Basics about HRTFs

Their role in spatial hearing Head-related transfer functions (HRTFs) deal with the sound transmission from a sound source point[2] to a point in the ear canal of a human being. They describe the filtering of incident sound waves through reflections and diffractions at the head, torso and pinna, depending on the angle of sound incidence. HRTFs play an important role in the localization of sound sources - often referred to as "spatial hearing" - which can be understood in conjunction with the so-called duplex theory first introduced by Rayleigh in 1907. This theory states that for the localization in the front half of the horizontal plane, the human brain makes use of interaural-time- and level-differences (ITDs and ILDs):

- ITDs occur when a sound wave reaches one ear earlier than the other, because the sound source is located on the left or the right side of the human being (mathematically spoken the azimuth angle $\varphi \neq 0$ and elevation angle $\theta \neq \pm\frac{\pi}{2}$, see *Section 2.2*). The brain detects the phase difference in the signals of the left and the right ear which allows to estimate the azimuth angle from which the sound was emitted. Above approximately 1.6 kHz, the wavelengths are shorter than the distance between the two ears which leads to ambiguity in the detection of the phase information. Therefore, in the "classic" duplex theory the ITDs are only significant up to a frequency of roughly 0.8 - 1.6 kHz.
- ILDs occur when the head causes an "acoustic shadow" and, thus, decreased sound pressure level at the contralateral ear.[3] However, this effect is only relevant at frequencies above approximately 1 kHz, because at lower frequencies, the sound waves bend around the head, resulting in identical levels at both ears.

Thus, in the classic duplex theory, ITDs determine the localization up to about 800 Hz and ILDs determine the localization above 1.6 kHz with a crossover range between 800 Hz and 1.6 kHz. However, real-world signals are usually broadband and in this case, the situation is more complex, because at high frequencies the ITDs of the entire signal's envelope (in contrast to the phase information of a single sine wave) can be detected and spectral differences between left and right ear are also considered.[4]

The duplex theory forms the basis of spatial hearing when only the front half of the horizontal plane is considered. However, as shown in *Figure 1* (left), a "cone" of points exists, at which the ITDs and ILDs are constant. Therefore, sound sources located on this so-called "cone of confusion" cannot be distinguished via ITDs and ILDs. Research

[2] As explained in Zaar [Zaa10], the distance between source and head is only relevant below 1.3 meters and can thus be neglected for sake of simplicity if the sound source point is located in the far-field. However, for sound sources in the near-field, like in telecommunications, it is important to also consider the radius.

[3] When a sound source is located to the left of a subject, the left ear is referred to as the "ipislateral" ear and the right ear is referred to as the "contralateral" ear and vice versa.

[4] See Macpherson, E.A. and Middlebrooks, J.C. (2002): Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. Journal of the Acoustical Society of America., 111(5), p. 2219-2236.

has shown that there are two ways in which human beings dissolve this ambiguity:

- Instinctively, small head movements lead to small changes in ITDs and ILDs which - in relation to the direction and velocity of the movement - help to discern the correct sound source position.
- The frequency- and direction-dependent filtering of the incident sound waves due to reflections and diffractions at head, torso and pinna (referred to as HRTFs, as described above) is analysed by the human brain and also improves the localization, even without head movement.

Of course, the actual localization process of human beings is far more complex because it is also influenced by other types of cues (for example, visual cues) and the cognitive abilities of our brain. However, this is not relevant for the purpose of this thesis, and is thus not considered in greater depth herein.

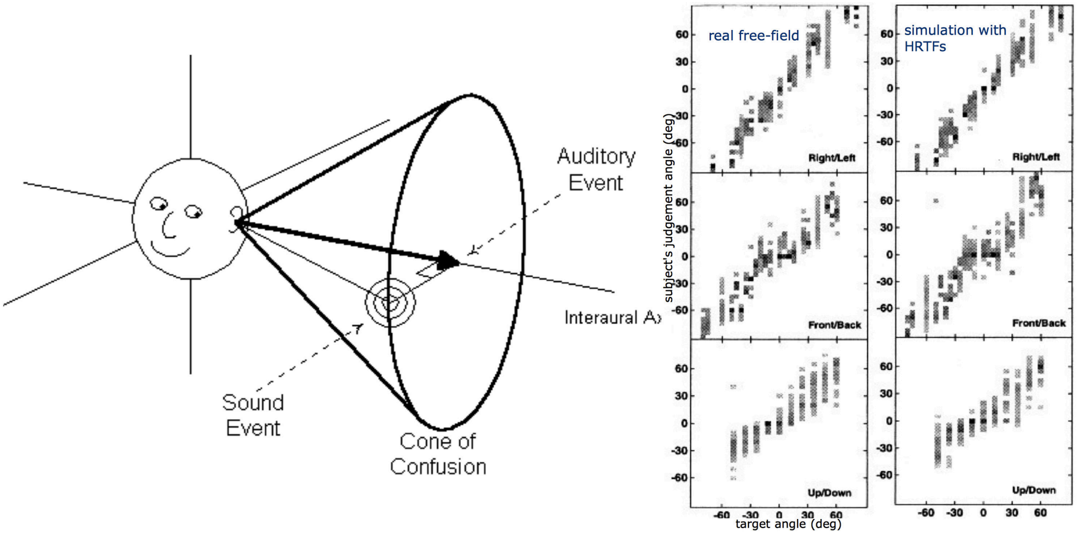


Figure 1: Cone of confusion on which ITD and ILD are constant (left), comparison of localization performance between real free-field situation and a corresponding simulation using HRTFs (right), (from: [Zaa10] and [Maj10])

Mathematical representation Head-related transfer functions basically comprise of two parts as shown by Møller et al. [MSHJ95]:

– Direction-dependent part: $\frac{P_2}{P_1}(\varphi, \theta) = \frac{\text{sound pressure at entrance to blocked ear canal}}{\text{sound pressure at head center without subject}}$
φ ... azimuth angle, θ ... elevation angle (see *Figure 2*)
This part is measured when placing a microphone at the entrance of a blocked ear canal (called "blocked-meatus" technique).

– Direction-independent part: The radiation impedance at the entrance of the ear canal, the transfer function of the ear canal itself and the impedance of the eardrum are also taken into account.
This part is measured when the ear canal is not blocked and the microphone is either situated at the entrance of the ear canal or directly at the ear drum.

The insertion of microphones deeper into the ear canal is difficult and dangerous to human subjects. Furthermore, the ear canal introduces additional, undesired variation between different people when only the filtering effects of torso, head and pinna are of interest. Therefore, the blocked-meatus technique is the most common choice for HRTF measurements.

Thus, head-related transfer functions can be expressed as

$$\underline{H}(\varphi,\theta,r,f) = \frac{\mathsf{FT}\{P_2(\varphi,\theta,r,t)\}}{\mathsf{FT}\{P_1(t)\}} \tag{1}$$

with f being the frequency, t being the time and FT being short for Fourier transformation.

Usually, HRTFs are assumed to be linear-time variant (LTI) systems (see *HRTF measurement*) and as such they can be split into a minimum-phase portion $\underline{H}_{min}(f)$ and an all-pass portion $\underline{H}_{ap}$:

$$\underline{H}(f) = \underline{H}_{ap}(f) \cdot \underline{H}_{min}(f) \tag{2}$$

However, most of the energy is contained in the minimum-phase part. And as explained by Hölzl [Höl12], pp. 9, various listening tests have confirmed that in most cases, the all-pass term can be neglected. The advantage of this approach is that the phase of minimum-phase systems can be derived from its magnitude spectrum via the Hilbert transform. In addition, the initial propagation delay can be removed and replaced by an additional ITD parameter, if desired. Thus, the magnitude responses for the left and right ears plus the corresponding ITDs is a sufficient representation of the spectral, time- and level-based cues for localization.

This also allows for more flexible post-processing: For example, spectral features which are common to all HRTFs of one subject can be assumed not to contain relevant spectral cues. Thus, the so-called Directional Transfer Functions (DTFs), containing only the direction-dependent cues, can be derived. First, the Common Transfer Function (CTF) - which contains the direction-independent parts - is calculated as the logarithmic mean value of the HRTF amplitude spectra of all source positions from one subject (i indicates the different source positions and N is the total number of source positions):

$$C(f) = \frac{20}{N} = \sum_{i=1}^{N} \log|\underline{H}_i(f)| \tag{3}$$

Then, the DTFs are obtained for each source position by subtracting the common parts of the HRTFs:

$$D_i(f) = 20\log|\underline{H}_i(f)| - C(f) \tag{4}$$

As only the amplitude spectrum of $\underline{H}(f)$ is considered in *Equations 3* and *4*, the phase information is discarded. This is valid, because the most significant parts of the system are minimum-phase (as explained above) and it is thus possible to reconstruct the phase later by using the Hilbert transformation.

Generally, it should be pointed out that the term "head-related transfer functions" (HRTFs) actually refers to the frequency domain representation of the corresponding "head-related impulses responses" (HRIRs) in the time domain. As one can be derived from the other by a simple Fourier transform, they can be regarded as equal, but they are not identical.

Applications and HRTF modelling To sum it up, amplitude, time-of-arrival and direction-dependent filtering of sound waves are different for the two ears and determine the perceived direction of the source. Basically, all of these effects are captured in HRTFs which explains their importance for the understanding of spatial hearing and psychoacoustics and, thus, the importance of this research area. In addition, HRTFs play a key role in the rendering of virtual environments where the impression of being situated in a virtual soundscape is generated via headphones. For this puprose, sounds have to be processed so that the user actually has the impression that its source is situated in a certain direction and at a certain distance. When only using ITDs and ILDs in conjunction with headphones, the sounds move in the desired direction but only within the horizontal plane. Furthermore, they are still localized within the head which is referred to as in-head localization or lateralization. The impression of real sound sources elevated from the horizontal plan and outside the head - referred to as externalisation - is only possible by filtering the source signal with the HRTFs corresponding to the desired source direction. *Figure 1* (right) shows that sound localization with HRTFs is comparable to the localization in a "real" free-field situation.
However, it is absolutely crucial that a test subject uses their own HRTFs, because even small changes in the anthropometry of the pinna often result in significant changes of the spectral cues. In fact, as explained by Majdak [Maj10], correct localization with someone else's HRTFs or even with the own, but altered, HRTFs is only possible after extensive training sessions. Thus, the acquisition of individual HRTFs for a human subject based on her/his anthropometric data is an important research area. Many different attempts have been made to find connections and to relate certain spectral features of HRTFs to certain anthropometric details of the pinna and other parts of a subject's body. However, as it is well beyond the scope of this thesis to go into more detail on this issue, please refer to Hölzl [Höl12] and Xu et al.[5] for an overview of different approaches. Still, it is save to say that enormous amounts of work will be required for obtaining satisfactory results. Large and more high-quality HRTF archives would be an important boost to these efforts.
In addition, the concept of HRTFs can also be expanded in several ways: For example, the influence of a certain room can be deliberately included in the measurements which results in so-called Binaural Room Impulse Responses (BRIRs). For moving sources in a virtual environment setup, the problem of switching or interpolating between HRTFs for different source directions can be avoided by usage of time-varying system identification (see *Time-variant system identification* in *Section 2.4*). Apart from that, measurements with several microphones placed at typical positions in hearing-aid device cases or with 16 microphones located on a robot dummy's head for research purposes have been performed (see *Section 3.6*). It is not the goal of this thesis to present all applications and variations of HRTFs. However, when considering a new common storage format, it is important to keep the format flexible enough to handle all these specific cases.

[5] Individualization of head-related transfer function for three-dimensional virtual auditory display: A review, S. Xu, Z. Li and G. Salvendy, in R. Shumaker: Virtual Reality, 2007

2.2 Geometry of HRTF measurement setups

HRTF measurement setups consist of one or multiple sources and one subject, which usually is a human being or a head and torso simulator with two ears where two sound receivers (microphones) are located.[6] The measurement room and its acoustic features must also be taken into account as they influence the measurement.

Coordinate systems *Figure 2, left* shows the description of a head and source configuration in spherical coordinates with the subject's head located at the origin of the coordinate system. It consists of:

- azimuth angle φ (between head center and source in horizontal plane)
- elevation angle θ (between head center and source in vertical plane)
- radius (distance between head center and source)

Alternatively, lateral and polar angels can be used, as shown in *Figure 2, right*. This representation is connected to the binaural (ITD and ILD) and spectral localization cues: The perceived lateral angle α is defined by the binaural cues while the perceived polar angle β is dominated by the spectral cues due to head, pinna and torso filtering. It is a logical step to also describe the measurement setup and thus the sound source directions with lateral and polar angles.[7] In most cases, the use of spherical or polar coordinates naturally emerges from the radial symmetric measurement setup. However, different definitions for the range of azimuth and elevation angles can be used to cover all possible directions. Evaluation of the databases presented in *Chapter 3* reveals three

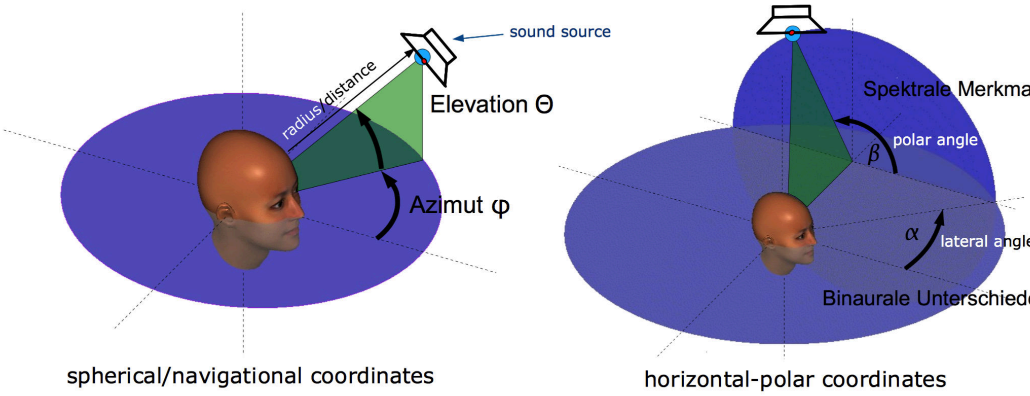


Figure 2: Two different ways of describing an HRTF measurement setup using two angles and a radius to determine the source position; the subject's head is located at the center of the system (adapted from: [Maj10])

[6] In some cases there might be more than two receivers. For example, when the measurements are related to hearing-aid devices with several microphones.

[7] In certain cases, for example the CIPIC database (see *Section 3.1*), the way the loudspeakers are moved for achieving different source directions might also inherently suggest to use lateral and polar angles.

approaches shown in *Table 1*.[8] In some cases - for example when acquiring BRIRs (as in *Section 3.3*) - the setup is not radial symmetric and it is thus more convenient to use standard Cartesian coordinates. For a common HRTF archive, all of these definitions might be allowed as long as the used coordinate system is clearly defined. Once this requirement is met, conversion from one system to the other is no problem.

Name	azimuth/ lateral	elevation/ polar	front	back	left	right	top
			azimuth/elevation *or* lateral/polar				
Geographic	$0^\circ \dots 360^\circ$	$-90^\circ \dots 90^\circ$	$0^\circ/0^\circ$	$180^\circ/0^\circ$	$270^\circ/0^\circ$	$90^\circ/0^\circ$	$0^\circ/90^\circ$
Navigational	$-180^\circ \dots 180^\circ$	$-90^\circ \dots 90^\circ$	$0^\circ/0^\circ$	$180^\circ/0^\circ$	$90^\circ/0^\circ$	$-90^\circ/0^\circ$	$0^\circ/90^\circ$
Horizontal-polar	$-90^\circ \dots 90^\circ$	$-90^\circ \dots 270^\circ$	$0^\circ/0^\circ$	$180^\circ/0^\circ$	$270^\circ/0^\circ$	$-90^\circ/0^\circ$	$0^\circ/90^\circ$

Table 1: Different versions of spherical and lateral coordinates with varying azimuth and elevation angle ranges and the corresponding azimuth/elevation value pairs for important directions

Realisation of different source directions Normally, in one measurement cycle, the transfer functions for different sound incidence directions on the head are obtained while the distance between head and source is kept constant. In practice, the variation of azimuth and elevation angles can be realised in different ways:

- Loudspeaker(s) located in horizontal plane:
  - Only one loudspeaker is used and the azimuth angle is varied by rotating the subject. When only source positions in the horizontal plane are of interest, this method works well. However, elevation angles other than 0° are difficult to realise with human subjects as the loudspeaker has to be moved. One possibility is rising the loudspeaker with a crane (as in *Section 3.2*).
  - Multiple loudspeakers at different azimuthal positions in the horizontal plane are used. This allows to measure HRTFs for several azimuth source positions without rotating the subject (though it might be rotated in addition to increase the azimuthal resolution) which is faster than with a single loudspeaker. Regarding realisation of the elevation angles, similar difficulties occur; but moving multiple loudspeakers is even more difficult than moving a single loudspeaker (although, it was accomplished at CIPIC, see *Section 3.1*).
- Loudspeakers located in vertical plane:
  Several loudspeakers are located at multiple elevations and the subject is rotated to achieve different azimuth values (for example, see *Section 3.4*). Thus, realisation of elevation angles is no longer a problem. On the other hand, the achievable elevation resolution strictly depends on the number of loudspeakers.

Apart from practical considerations, it must also be taken into account, that the question of rotating/moving the subject or the sound sources also changes the way the room influences the measurement via reflections, modes etc. When the loudspeaker positions are constant and the subjects are rotated, the room influence does not change by varying

[8] Note that in other contexts, other names might be used for the same specification of a coordinate system. For example, in the CIPIC database, "navigational" is referred to as "interaural-polar".

the azimuth and elevation angle during the measurement cycle. However, when loudspeaker positions are changed, the way they interact with the room is affected and thus also the room influence on the measurement during the course of the measurement. This must be taken into account when trying to minimize the room influence by removing its spectral characteristics from the measurement results in the post-processing stage. It is also another argument why it is advisable to record HRIRs with as little reflections from the room as possible by adequate acoustic preparation of measurement room and setup and by sensible time-windowing of the obtained impulse responses (see *Section 2.3*). These considerations must also reflect in the definition of a common storage format. Furthermore, more sophisticated and special setups such as a varying torso angle and BRIRs with arbitrary source positions (as used at TU Berlin, see *Section 3.3*) must be describable in such a new format. More thoughts and propositions on these issues can be found in *Section 4.1* of this work.

Spatial resolution requirements and interpolation The spatial resolution is an important indicator for the quality and usability of an HRTF database. It is defined by the number of measurement points in azimuthal and elevational directions. Increasing the number of measurement points usually results in either a more complex measuring setup, a more time-consuming measurement process or both, so a trade-off between effort and resolution has to be found. Of course, the actual required resolution also depends on the purpose of the acquired data and a single sensible resolution value can certainly not be given. However, as suggested in [MBL07], a resolution of 5° in the horizontal plane (azimuth angle) and 10° in the vertical plane (elevation angle) seem to be a reasonable reference. Databases with roughly 1000 or beyond measurement points fulfil these requirements.[9] However, when using moving sources in virtual environments or when performing very critical listening tests, the required resolution might even be around 2° as shown by Minnaar et. al in [MPC05].

One possibility to overcome limited spatial resolution is interpolation between measrurement points. For example, Martin and McAnally [MM07] showed that interpolating between HRTFs with a resolution of 20° achieved the same performance in a localization test as the same dataset with a higher actually measured resolution. Although this might be possible and applicable for certain purposes, it must be kept in mind that interpolation is always just an estimation. To give an example, in the research field of HRTF customization, interpolated HRTFs might be too inaccurate for the evaluation of HRTF models. It is well beyond the scope of this thesis to present and examine the performance the various interpolation approaches used with HRTFs. However, with regard to a common storage format or archive, it is important to argue that such a database should not use interpolation as a means of "tailoring" different data sources to fit into one pre-defined measurement point scheme. If such steps are thought to be necessary for certain applications, interpolation can still be performed on actually measured data

[9] In many cases, the elevation angle range is not covered in its entity (particularly for negative angles corresponding to source directions "from the floor"), which results in less measurement points for a given resolution. Also, at high elevations, the azimuthal space between points naturally gets smaller and, thus, the azimuthal angular resolution is often reduced while retaining a comparable resolution regarding the "absolute" distance between measurement points.

from such a database, as necessary.

Room description Apart from describing the measurement setup with the sound sources and the subject, it is also crucial to document the measurement room and the placement of the setup within the room, because both influence the measurement and can be considered for evaluation of its quality. Appropriate treatment of the room in order to avoid disturbing reflections and distortions of the frequency response is of utter importance.[10] The most relevant parameters, which should be specified in any case, are:
– Dimensions: length x width x height, if not applicable: at least room volume
– Room-type: Anechoic, semi-anechoic etc., possibly verified with absorption values
– Reverberation time T_{60}
– Reflectogram: graphical representation of the early reflections
Furthermore, sketches of the room and the measurement setup are relatively simple and appropriate means of documentation. While this can easily be included in an accompanying paper, it is already a challenge how to store the basic measurement room geometry in an HRTF archive. Single parameters are easy to store, but provide generalized and sometimes simplified information which might be too little for an accurate description. Thus, more sophisticated approaches such as allowing the possibility to integrate or link entire room models to a database can be considered.

2.3 HRTF measurement

Generally, the target of head-related transfer functions is to describe the behaviour of a system comprising head, torso and pinna of a subject's body. The identification of this system is usually realized by measuring the impulse response. This implies the assumption that the measurement system and path are linear and time-invariant systems (so-called LTI systems). However, in reality, non-linearities are introduced in the measurement equipment and human subjects might move their heads during their measurements resulting in slightly time-variant and non-linear systems. This has to be taken into account when designing the measurement setup and chosing the measurement method (for a deeper evaluation of impulse response measurement methods refer to *Section 2.4*).

The measurement process In *Figure 3*, a general example of such a measurement setup is given: Like most contemporary systems, it is based on a measurement software which generates the excitation signal $x(t)$ for the impulse response measurement and records the system response $y(t)$ while also taking care of additional tasks such as controlling the rotation of subject and/or loudspeakers and monitoring the subject movement (head tracking, for example via infrared). The excitation signal is converted to the analog domain, amplified and fed to the loudspeaker. Propagation in the measurement room results in filtering with the room impulse response $R(f)$. When the

[10] A detailed discussion of acoustic measurement room treatment is beyond the scope of this thesis.

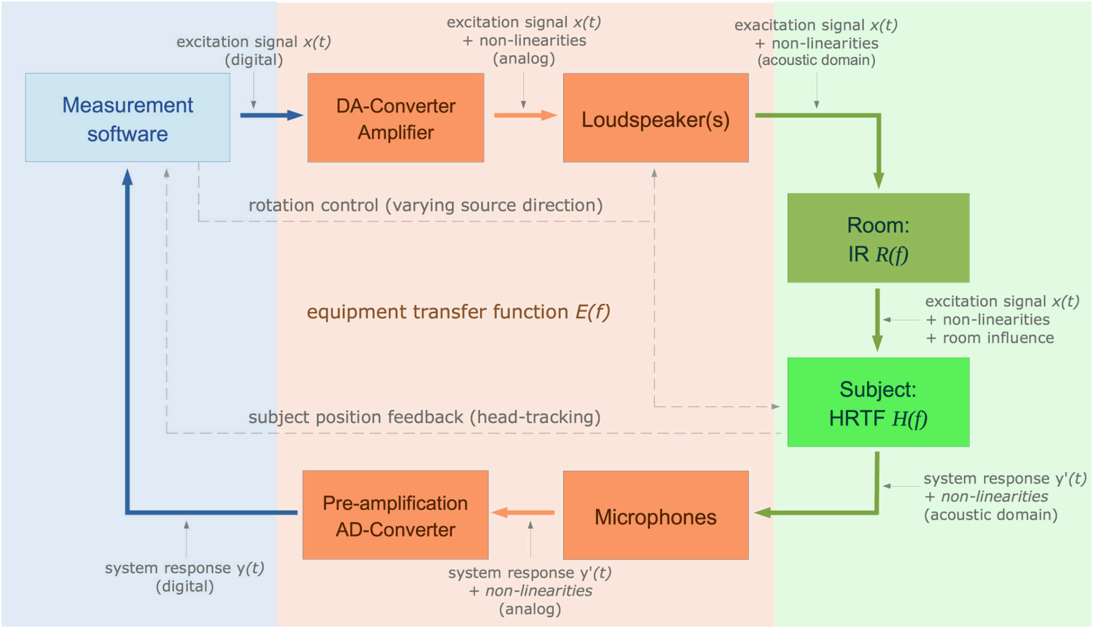


Figure 3: Schematic overview of a typical software-based HRTF measurement system (such as used for the acquisition of the Listen HRTF and ARI databases)

sound waves reach the subject they are filtered again by head, pinna and torso which corresponds to the actual head-realted transfer function that to be obtained. Microphones measuring the sound pressure of the system response are located at or in the ear. Their output signal is sent to microphone pre-amplifiers, converted back to the digital domain and then recorded by the measurement software. Depending on the method of system identification, the impulse response is usually obtained from the system response via auto-correlation, spectral division or deconvolution.[11]

During one measurement cycle the sound source direction (azimuth and elevation angle) is varied to cover the desired spatial range. As described in *Section 2.2* this can be realised by either rotating/moving the subject or the loudspeaker(s). Often, the measurement software also controls the rotation which allows for optimal timing of rotation and measurement periods. In case that head-tracking is used to limit the errors introduced by movements of the subject, this signal is practically also fed to the measurement software.

Many laboratories also document anthropometric data of their subjects so as to be able to develop and investigate anthropometry-based HRTF individualization and generation models. A typical set of parameters is shown in *Section 3.1* about the *EarLab CIPIC HRTF database search*. Approaches for the acquisition of the data range from measuring them by hand with tape to retrieving the desired data from 3D head scans.

Post-processing As indicated in *Figure 3*, distortion is introduced by the measurement equipment (corresponding transfer function $E(f)$) and the frequency response is

[11] See *Section 2.4* for a more detailed discussion of these methods and the associated steps to obtain the impulse response from the system response.

also altered by reflections and modes of the measurement room ($R(f)$). Instead of the pure head-related transfer function $H(f)$, the distorted system frequency response $Y(f)$ is obtained after the auto-correlation/deconvolution step:

$$Y(f) = R(f) \cdot E(f) \cdot H(f) \tag{5}$$

In order to remove the distortion from $Y(f)$ and obtain $H(f)$, the minimum-phase portion of the room and equipment transfer functions can be discarded from $H(f)$ by "spectral division" (corresponds to filtering with a filter's inverse).

$$H(f)R_{max}(f)E_{max}(f) = \frac{S(f)}{R_{min}(f) \cdot E_{min}(f)} \tag{6}$$

$$\Rightarrow H'(f) = |H(f)| \cdot e^{arg\{H(f)+R_{max}(f)+E_{max}(f)\}} \tag{7}$$

This way, the magnitude spectrum of $H(f)$ is compensated, but the phase of the all-pass components remains. It is not possible to also remove them, because inverting these filters would result in unstable systems. However, it was shown in *Section 2.1* that the phase of HRTFs is usually set to minimum-phase, anyway. Thus, the uncompensated phase of $H(f)$ is practically irrelevant. The transfer functions $R(f)$ and $E(f)$ can either be obtained by separately measuring the impulses responses of all components (room, amplifier, loudspeaker etc.) or by summarizing the effects of all components by performing an IR measurement at the laboratory setup without the subject. The latter approach is more commonly used as it is simpler and also more accurate than the first one, as all distortion is captured directly at the very measurement setup.

A second important post-processing step is truncating the obtained head-related impulses responses to sensible lengths. This is necessary because the actual HRIRs usually are only a few milliseconds long. The measurement inevitably contains undesired reflections from the room floor, walls and ceiling, from the measurement equipment (even if wrapped in absorptive acoustic foam) and possibly also from knees and legs of the subject. They can be removed by applying a window of adequate length to the impulse responses in the time-domain (usually about 200-300 samples depending on the purpose).

In addition, custom post-processing is performed as necessary: For example, the DTFs might be calculated or errors like inconsistent ITD and ILD values due to loudspeaker misplacement might be corrected. When the HRTFs are used to present stimuli to a subject via headphones in the context of a psychoacoustic experiment or a virtual environment application, it is important that the magnitude response of the used headphones is also taken into account. Usually the headphone transfer function is measured and compensated for in the frequency domain, similar as with the room and equipment transfer functions.

Reciprocal method An important alternative to the measurement approach discussed in this section relies on Helmholtz' principle of reciprocity and basically consists of exchanging the positions of loudspeakers and microphones: Miniature loudspeakers are inserted into the ears and play the excitation signal for system identification while a microphone array is placed around the subject to capture the system response. It is an interesting idea allowing for very fast HRTF measurements with the spatial resolution being only limited by the number of microphones available.[12] At the moment of writing, the main problem lies in the achievable SNR due to the limited performance of the miniature speakers and the limitation in absolute sound pressure level when measuring HRTFs of human subjects so as not to harm their hearing sense. It is beyond the scope of this thesis to discuss the reciprocal method in depth,[13] however, a reciprocal setup, as documented in [MH10], is shown and discussed as an example in *Section 3.6*. It certainly is important that a standardized storage format is flexible enough to handle the description of reciprocal HRTF measurement setups. In this case, the common assumption in many databases that the subject's head and thus the sound receivers are always located at the origin of the coordinate system can not be held any longer.

2.4 Methods for impulse response measurement

As already explained in the previous section, non-linearities, time-variance and additional noise are introduced when acquiring the impulse response of real systems such as HRTFs:

- non-linearities: measurement equipment (mostly amplifieres, loudspeakers, microphones)
- time-variance:
  - movements of the subject
  - changes in room conditions (for example temperature, humidity)
- noise/outside error influences:
  - background noise of the room
  - transient disturbances
  - measurement equipment

Therefore, in order to reduce the effects of these influences, the choice of a good method for transfer function measurement is crucial. Some important features are:

- measurement duration
- signal-to-noise ratio (SNR) (vs. measurement duration)
- robustness against error noise and transients
- separation of linear and non-linear parts of the impulse response

The SNR can always be increased by extending/repeating the measurement. Therefore, a trade-off between these two parameters has to be found.

In the following, an overview and discussion of the most common impulse response measurement methods used in HRTF databases is given.

[12] Also, it is necessary that the reflections from the microphones do not disturb the measurement result when a large number thereof is used.

[13] For more information refer to Fast head-related transfer function measurement via reciprocity by Zotkin et al., Journal of the Acoustical Society of America, Vol. 120, 2006 and to Zaar [Zaa10].

Golay-codes [14]

Golay-codes are pseudo-random binary codes derived from two simple recursive relations:

$$a_1 = [1, 1] \text{ and } b_1 = [1, -1] \tag{8}$$

$$a_{n+1} = [a_n, b_n] \text{ and } b_{n+1} = [a_n, -b_n] \tag{9}$$

The sum of the auto-correlations r_{aa} and r_{bb} of these codes cancel each other, except for a Dirac impulse with an amplitude of $2L$ (L is the length of the sequence):

$$r_{aa}[n] + r_{bb}[n] = 2L\delta[n] \tag{10}$$

Both codes a and b are used as excitation signals during the measurement and the corresponding responses y_a and y_b of the system are recorded. Because of equation 10, the impulse response $h[n]$ can be obtained as follows:

$$h[n] = \frac{1}{2L}(r_{ay_a} + r_{by_b}) \tag{11}$$

The energy contained in the two Golay-codes, each of length L, is $2L$ times greater than the energy in a single impulse. Thus, Golay-codes increase the SNR of an IR measurement by $10\log(2L)$ compared to excitation with a single impulse. However, the measurement must be performed with both excitation codes a and b, which doubles the measurement duration and deteriorates the robustness against time-variance during one measurement cycle. Zahorik [Zah99] has shown that head movement during an HRTF measurement might lead to significant artefacts when using Golay-codes. By contrast, the artefacts were not present in the results when using the MLS technique (described below). It should be taken into account that the effect of this flaw increases when longer measurement durations are required because of a long system response or high SNR demands. For example, this might be the case when the response of the room shall also be captured, such as with binaural room impulse responses (BRIR).

Maximum Length Sequences (MLS)

Similar to Golay-codes, maximum length sequences (MLS) are pseudo-random binary codes, that have a flat magnitude spectrum. The sequences are generated by a one-bit shift register with a feedback loop that contains an "exclusive OR" bit operator. For usage in impulse response measurement, the bits "0" and "1" bits are mapped to "-1" and "+1". The length L of a sequence is determined by its order m:

$$L = 2^m - 1 \tag{12}$$

As circular cross-correlation is used to obtain the impulse response $h[n]$ from the system response $y[n]$, the excitation sequence must be at least as long as the expected impulse response of the measured system, in order to avoid time-aliasing:

$$T_{MLS} = \frac{L}{f_s} = \frac{2^m - 1}{f_s} \geq \tilde{T}_{IR} \tag{13}$$

[14] See [BS09] and [Maj10].

Given the minimal length, the required order of the MLS sequence therefore is:

$$m \geq \frac{log(\tilde{T}_{IR} \cdot f_s + 1)}{\log(2)} \tag{14}$$

Due to the non-periodic nature of the signal, most of the energy in the auto-correlation function of an MLS sequence x is centered around 0. The uneven number of samples in the code prevents a perfect balance between "+1" and "-1" and thus leads to a slight offset of $\frac{1}{L}$ (which is increasingly dampened with larger lengths L):

$$r_{xx}[n] = \delta[n] - \frac{1}{L} \tag{15}$$

As shown in [Wes11], the cross-correlation between the excitation sequence x and the measured system response y equals the impulse response of the system (except the DC offset dampened by $\frac{1}{L}$):

$$r_{xy}[n] = \ldots = r_{xx}[n] * h[n] \approx (\delta[n] - \frac{1}{L}) * h[n] = h[n] - \frac{1}{L}\sum_{k=0}^{L-1} h[k]$$

$$\rightarrow r_{xy}[n] \approx h[n] \tag{16}$$

For efficient calculation of the cross-correlation, the so-called Fast Hadamard Transformation can be used.[15]

The phase spectrum of an MLS sequence is erratic with uniform density of probability between $-\pi$ and π. Therefore, all portions of the measured system response y, that do not correlate with the excitation signal x, are phase-randomized during the cross-correlation. Practically, these portions are equally distributed along the time axis of the resulting impulse response $h[n]$. This makes the MLS measurement very robust against transient error signals, because they do not correlate with the MLS sequence.

As only the maximum amplitudes "$+1$" and "-1" occur in the sequence, the sequences have a very low crest factor of 1, which is another advantage. Therefore, the theoretical SNR of an IR measurement with MLS is $10\log(L)$ dB higher than a measurement with a single impulse excitation. However, it is impossible to create the infinitely steep slopes of a pulse signal with real equipment due to non-ideal anti-aliasing filters in D/A-converters, electronic circuits and electro-acoustic conversion in the loudspeaker. This leads to transient overshoot causing distortion which increases the noise floor and, thus, decrease the SNR. Therefore, the amplitude of the signal that is fed to the measurement system must be reduced which means that the advantage of a low crest factor is lost. On the other hand, lowering the amplitude too much will also decreases the SNR at a certain point. Thus, the optimal setting for the amplitude usually has to be found in a time-consuming "trial-and-adaption" process.

[15] For more information see The Fast Fourier-Hadamard Transform and its use in Signal Representation and Classification, J. E. Whelchel Jr. and D. F. Guinn, 1968

Inverse Repeated Sequence (IRS)

IRS is an impulse response measurement method based on MLS sequences which suppresses even-order distortions as shown in [DH93]. This is achieved by using an MLS sequence $m[n]$ with the length L and its inverse $-x[n]$, then interleaving them on the time axis to generate an IRS $x[n]$ of length $2L$ (see *Figure 4*, left)

$$x[n] = \begin{cases} m[n], & n \text{ even}, 0 \leq n < 2L \\ -m[n], & n \text{ uneven}, 0 \leq n < 2L \end{cases} \tag{17}$$

The deconvolution process, which is necessary to derive the actual impulse response from the measured system response, is exactly the same as with the MLS technique. However, like the excitation sequence, the obtained response will also have a length of $2L$. Due to the anti-symmetry of the excitation sequence (see figure 4, right), the impulse response $h(t)$ is located in the first half of the response (samples 0 to $L-1$) while an inverted version $-h(t)$ of the impulse response will occur in the second half (samples L to $2L$). The second half holds no additional information and can thus be discarded. This means that the benefit of suppressing even-order harmonics comes at the expense of doubling the measurement time while non-even order harmonics are not suppressed at all.

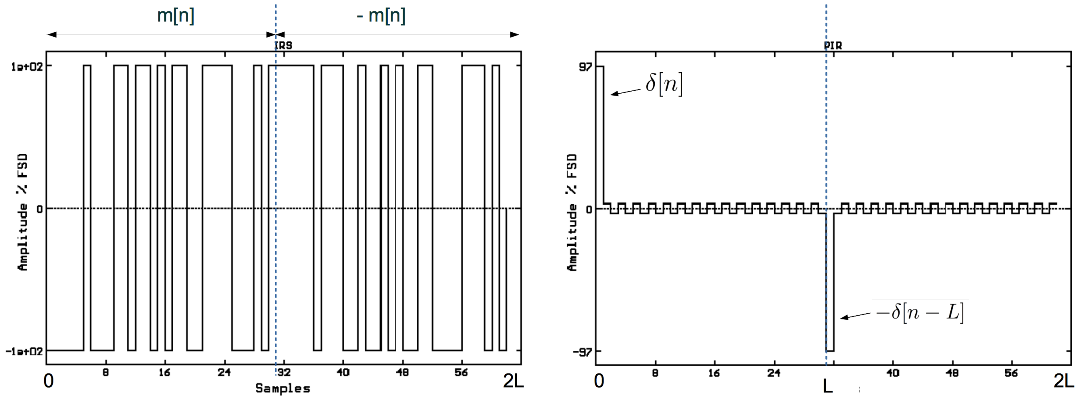


Figure 4: IRS $x[n]$ consisting of an MLS sequence $m[n]$ and its inverse $-m[n]$ (left); auto-correlation $r_{xx}[n]$ of the IRS with a positive and a negative Dirac delta impulse (right; both adapted from: [DH93])

Modified Inverse Repeated Sequence (MIRS)

Ewert and Kayser use a modification of the IRS in their measurements for a database of multichannel in-ear and behind-the-ear HRTFs and BRIRs at Universität Oldenburg. A modified inverse repeated sequence (MIRS) consists of successive IRSs of different orders. In the post-processing stage, the median of all the IRs with different orders is calculated and the result is windowed to the length of the lowest order IR. This grants suppression of uneven-order non-linear distortions and thus makes the MIRS immune against distortion in general.

In [HEA$^+$09], Ewert and Kayser claim that to have achieved an SNR of between 86 and 105 dB (depending on the measurement location) with their MIRS-based measurement. This is comparable to the performance of the sine sweep method proposed by Farina [Far00] in anechoic conditions. However, a much longer measurement time (about 30 s for one sequence) has to be taken into account. It might be argued, that when repeating a short sine sweep for 30 s, an even higher SNR can be obtained as it increases by +3 dB with every repetition. Still, for the measurements in public locations, which were necessary for this database, the MIRS has two main advantages over sweeps (see below):

- high immunity against transient noise (see MLS above)
- noise is less disturbing than sine sweeps (→ higher sound pressure levels acceptable to people in the public → higher SNR can be achieved)

Sine sweeps

As the name suggests, sine sweeps are pure sine signals usually starting at a certain low frequency, which is then continually increased up to a high frequency. When used as excitation signals for impulse response measurements, the frequency range of the sweep should match the desired frequency range of the measurement plus some offset to allow for smooth fade ins and outs. Mathematically, a sweep $x(t)$ can be represented as a sine function with a time-dependent phase $\varphi(t)$:

$$x(t) = \sin(\varphi(t)) \tag{18}$$

The spectrum of the sweep is determined by the phase term $\varphi(t)$, because it influences the way the frequency rises. When the frequency rises in a linear fashion, the resulting magnitude spectrum is "white" (equal energy in each frequency band); when the frequency rises in an exponential fashion, the resulting magnitude spectrum is "pink" (equal energy in each octave band), which is the more common one. However, arbitrary spectra can easily be constructed when synthesizing the phase term $\varphi(t)$ of a sweep in the frequency domain. [16]

The system's response $y(t)$ is measured and then used to obtain the impulse response $h(t)$: An inverse filter $x_i(t)$ is constructed from the excitation signal $x(t)$ such that

$$x(t) * x_i(t) = \delta(t - \tau) \tag{19}$$

with τ being the length of the impulse response of $x_i(t)$. Basically, the inverse filter $x_i(t)$ is a time-reversed sweep (with a modulated amplitude spectrum in case of non-linear frequency rise behaviour). In order to get the linear impulse response $h(t)$, the system response $y(t)$ is convolved with the inverse filter $x_i(t)$ to obtain $h'(t)$ in a first step:

$$h'(t) = y(t) * x_i(t) \tag{20}$$

In the frequency domain, the convolution with the inverse filter $x_i(t)$ corresponds to a multiplication with the inverse of the spectrum of the excitation signal $x(t)$:

$$H'(j\omega) = Y(j\omega) \cdot X^{-1}(j\omega) \tag{21}$$

[16] For more information on the construction of sweeps in the frequency domain see Mueller [MM01], p. 33 ff.

Before the convolution, the linear system responses for each frequency are spread along the time axis (as shown on the left in *Figure 5*). After the convolution, they are shifted back into a "compact form" with the responses for all frequencies starting at the same time (*Figure 5*, right): This is the linear impulse response $h(t)$.

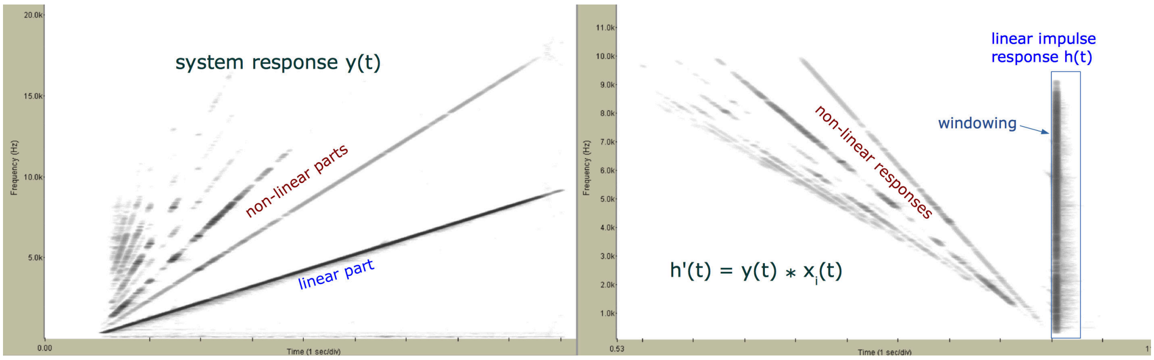


Figure 5: Time-frequency plot of the system response $y(t)$ (left) and the deconvoluted impulse response $h'(t)$ (right) containing harmonic distortion and the linear IR $h(t)$ (adapted from: [Far00])

It is also evident in *Figure 5*, that the non-linear responses of the system are moved to the left of the linear response as a result of the deconvolution process. Thus, linear and non-linear parts no longer overlap on the time axis and the non-linear parts can easily be separated from the linear parts by using time windowing. This allows for high immunity against distortion, increased SNR and the possibility to simultaneously measure the impulse response and the THD (total harmonic distortion) of a system.
Tests by Stan et al. [SEA02] show a difference of about 20 dB in the practically achievable SNRs of sweep-based and MLS-based measurements. It can be argued that, at the time of writing, these features make sweep-based impulse response measurements the best solution for acquiring HRTF data except for special applications.

Multiple Exponential Sine Sweep Method (MESM)

The MESM takes the concept of measuring impulse responses via exponential sweeps one step further and allows to significantly reduce the measurement time while retaining the advantages of this method for HRTF measurements. The number of sweeps in a given time is increased by placing them more effectively in a time-frequency representation. To that end, the concepts of "interleaving" and "overlapping" are employed:

– Interleaving: Each sweep is timed so that its linear response is "between" the linear response and the second-order harmonic response of the previous sweep in a time-frequency chart of the measurement process. This way, several sweeps and their harmonics are interleaved as shown in *Figure 6* (left).
– Overlapping: A new sweep is started as soon as its highest order harmonic does not interfere with the previous sweep. This concept is particularly effective with weakly non-linear systems (such as in *Figure 6*, middle) where high-order harmonics - which would disturb the next sweep - are not present.

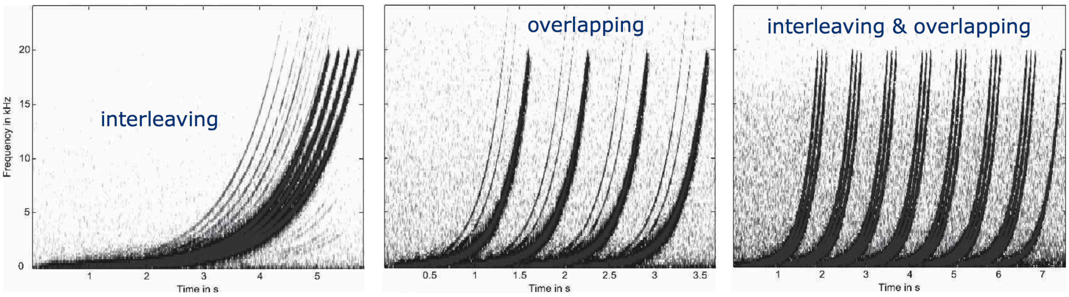


Figure 6: Time-frequency plot of a weakly non-linear system's responses to four interleaved sweeps (left), four overlapped sweeps (middle) and to an MESM excitation signal with overlapping groups of three interleaved sweeps (right), (adapted from: [MBL07])

In the MESM, both concepts are combined as shown in *Figure 6*, right: In this example, groups of three interleaving sweeps are formed. Then, seven of these groups (plus a single sweep at the end) are overlapped to obtain the excitation signal used for the measurement. An efficient way of grouping the sweeps has to be found and the exact time delay between the sweeps and the groups have to be optimised for each measurement setup and purpose.[17] This procedure may take some time as it involves a reference measurement with a conventional exponential sweep to obtain the necessary parameters for the MESM. However, once the system is calibrated, it usually does not need to be changed and drastically reduces measurement time. This might be less appropriate for mobile measurements, but for steady HRTF measurement setups it is optimal.

The rest of the process is basically the same as with conventional sweeps. After the deconvolution, the linear impulse responses of the different systems can be separated by time windowing, as they do not overlap in time. However, the harmonic impulse responses of the different systems do partially overlap in time. Still, the immunity against distortion and the ability to simultaneously measure the impulse response and the total-harmonic distortion of a system are retained, even though the harmonic impulse responses of the different systems cannot be accessed separately due to their overlapping.

Practical tests in [MBL07] proved that the measurement time for an entire HRTF set could be reduced to $25\ \%$ compared to conventional exponential sweeps. Apart from allowing laboratory staff to use their working time more efficiently, this makes the measurement process less fatiguing for the test subjects and, thus, probably alleviates retrieving candidates for such measurements. The immunity against time-variance is also improved due to the shorter measurement times. All in all, MESM currently seems to be the best suited impulse response measurement method for the acquisition of larger HRTF databases.

Time-variant system identification

Another possibility for retrieving HRIRs is to use time-variant system identification methods while performing a continuous 360°-azimuthal rotation of the subject at a constant

[17] For details of this process including a mathematical description of the required parameters refer to [MBL07].

source elevation. An adaptive filter $w[n]$ is used to estimate and track the time-variant acoustic transfer function $h[n]$. Based on the assumption that the changes in the HRIRs due to the azimuthal movement are slow compared to the convergence speed of the adaptive filter, the HRIRs can be captured for a continuous range of azimuth angles with sufficient quality.

Figure 7 (left) shows a discrete-time model for time-variant system identification using the normalized least mean-square (NLMS) algorithm. The excitation signal $x[n]$ drives the system-under-test $h[n]$ which produces the measured system response $\tilde{y}[n]$, comprised of the pure system response $y[n]$ plus the equipment and room background noise $r[n]$:

$$\tilde{y}[n] = x[n]h^T[n] + r[n] \tag{22}$$

The excitation signal $x[n]$ is also fed to the linear adaptive filter $w[n]$ which estimates and tracks the behaviour of the real system $h[n]$. The output of the adaptive filter $\hat{y}[n]$ is subtracted from the measured system response $\tilde{y}[n]$ to obtain the error signal $e[n]$:

$$\tilde{y}[n] = \tilde{y}[n] - x[n]w^T[n] \tag{23}$$

The error signal $e[n]$ is fed back to the NLMS algorithm which adjusts the parameters of the filter $w[n]$ in order to match the current state of the real system $h[n]$:

$$h(k+1) = w[n] + \mu_0 \frac{e[n]x[n]}{||x[n]||^2} \tag{24}$$

The speed with which the adaptive filter catches up with any changes in the real system is referred to as "convergence time" or "convergence speed" and is an important indicator for the quality of adaptive systems. It is influenced by several factors which also interact with the choice of an optimal excitation signal:

- It rises with the factor μ_0 in euqation 24, which is the step-size of the adaptation process. However, higher values of μ_0 comprise the robustness against noise.
- A setting for the lengths N of the adaptive filter $w[n]$ has to be found: A short filter allows for faster convergence, but when the filter is too short, significant parts of the IR's tail might be truncated resulting in an additional error noise $r_{tail}[n]$.
- Antweiler and Antweiler[18] show that an important aspect for achieving a high convergence speed is to choose a so-called "perfect signal" as an excitation signal, which means that it has an impulse-like auto-correlation function:

$$r_{xx}[n] = E_p \cdot \delta(k \mod N),\ E_p \ldots \text{energy in one period} \tag{25}$$

Thus, white noise is a possible choice as an excitation signal, but as shown in [AE09], the deterministic nature of "perfect sequences" (similar to MLS) make them less sensitive to $n_{tail}[n]$. However, perfect sequences lack the advantages of sweep-based impulse response measurements described above in *Sine sweeps*. Therefore, Antweiler et al. [ATVE12] show how to construct a perfect-sweep excitation signal that has the impulse-like auto-correlation feature requested for perfect signals in equation 24:

$$X(\nu) = \begin{cases} e^{\frac{-j4m\pi\nu^2}{N^2}} & \text{for } 0 \leq \nu \leq \frac{N}{2} \\ X^*(N-\nu) & \text{for } \frac{N}{2} < \nu < M \end{cases} \tag{26}$$

[18] C. Antweiler and M. Antweiler, "System Identification with Perfect Sequences based on the NLMS Algorithm", *International Journal of Electronics and Communications*, Vol. 3, pp. 129-134 1995

with ν being the frequency index, and a time-stretch-factor m specifying which portion of one period N the sweep covers ($m = M/2$ means that the sweep covers the entire period).

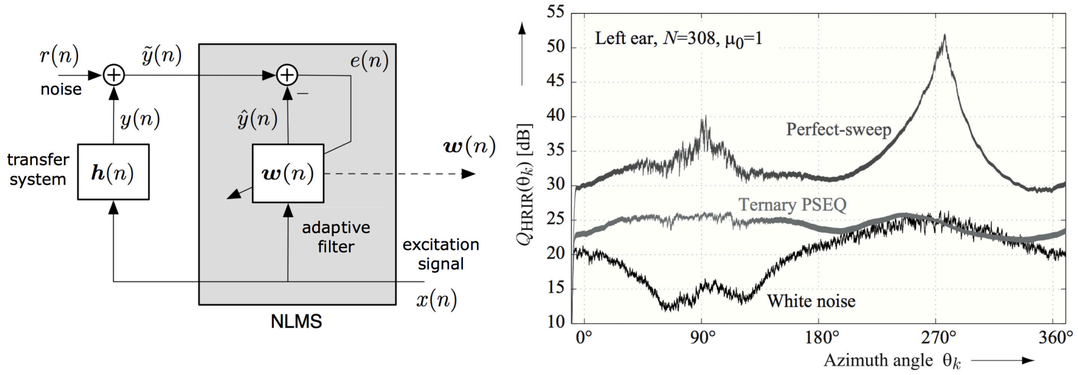


Figure 7: Model of the NLMS-based time-variant identification of the system $h[n]$ (left), quality of HRIR measurements with white noise, perfect-sequences and perfect sweeps covering full azimuthal space from $0°$-$360°$ (right) (from: [TAV10] and [ATVE12])

Figure 7 (right) compares the acquistion of quasi-continuous HRIRs using white noise, perfect-sequences and perfect-sweeps: During the measurement, the subject is continuously rotated to obtain a continuous change in azimuth angle while utilizing the NLMS to identify the time-varying system. In this case, one full rotation takes 20 s, which is very little considering the fact that during this time, HRIRs of the entire azimuthal space are acquired with almost infinite resolution. The plot in *Figure 7* (right) shows the quality of the measurement for the left ear microphone. The perfect signals both yield a better performance than the simple white noise, however, the perfect-sweep clearly outperforms the perfect-sequences on the entire azimuth range. The obvious change in performance at the azimuth angles $\theta_k = 90°$ and $\theta_k = 270°$ occur, because at these angles the minimum respectively maximum amount of direct sound from the loudspeaker is present at the left ear. This influences the significance of the truncated tail error $r_{tail}[n]$ and, thus, also the performance of the measurement. In [ATVE12], stationary measurements using these three excitation signals were compared to the time-variant measurements at both angles $\theta_k = 90°$ and $\theta_k = 270°$, showing no significant difference in the quality parameters except for the inferior white noise excitation.

Antweiler and Vary [19] have developed a computionally effective method for direct and sequential access to the single HRIRs from the continuous measurement, which is currently limited to perfect-sequences, though. Apart from that, the sweep-based measurement again is the optimal solution combining a rapid tracking ability with robustness against harmonic distortion. Generally, time-variant acquistion of HRIRs is a very attractive alternative to the conventional stationary measurement methods as they offer both reduced reduced measurement time and a quasi-infinite azimuthal resolution.

[19] Sequential and direct access of head-related transfer functions (HRTFs) for quasi-continuous angular positions, Christiane Antweiler and Peter Vary, 2011

Summary It has been shown that the set of pseudo-random binary sequences are a valid method of impulse response measurement, although, their features and SNR achievements are not as sophisticated as with sweep-based measurements. Practical studies in [SEA02] even showed a difference of about 20 dB in signal-to-noise ratio between MLS/IRS and sweeps which is mostly accounted to the sweep's immunity against harmonic distortion. The *MIRS* seems to be an interesting development as it discards the major disadvantage of the bianry sequences' sensitivity to harmonic distortion and, according to Ewert and Keyser [HEA$^+$09], also matches the sweep's SNR. Thus, in some cases, like when the sound of noise is preferred to the sound of sweeps due to the measurement environment, the MIRS could be more advantageous than sweeps. Still, in particular for the requirements of "usual" HRTF acquistion, the *MESM* offers unchallenged time efficiency while retaining all advantages of sweep measurements. Time-variant HRTF measurement seems to be a promising alternative, in particular for virtual environment purposes, where high spatial resolution is required for moving sound sources.

3 Evaluation of existing databases

In this chapter, publicly available databases and datasets are examined so as to get an overview of similarities and differences regarding measurement methods and setups, post-processing practices and data storage methods. As the classic HRTF measurement setup is often modified or expanded for special applications, the creation of a standardized file format also requires to handle these cases. Thus, some examples of such general head-related directional audio data measurements are included, as well.

3.1 EarLab CIPIC HRTF Search

The "EarLab CIPIC HRTF Search" database consists of head-related transfer functions (HRTFs) and relevant anthropometric data for 45 subjects, including a KEMAR manikin,[20] originally published by the "Center for Image Processing and Integrated Computing" (CIPIC) at University of California in 1998 (see [CIP12]). In 2001, EarLab - a laboratory located in the Hearing Research Center at Boston University, USA - created an online repository from this database and made it available on their website [Ear08].

Geometry The measurement setup basically consists of five loudspeakers which were mounted on a hoop with a radius of 1 m around the subject. The subject's head was located at the center of this hoop, which is also the origin of the coordinate system. Rotating the hoop about the interaural axis allows to vary the elevation angle. Different source directions are specified by azimuth and elevation angles in interaural-polar coordinates (equivalent to "HorizontalPolar" in *Table 1*, *Section 2.2*).[21] Angles of 0°/0° (azimuth/elevation) correspond to a source on the horizontal plane, in front of the subject's head. Azimuth angles range from $-90°$ to $+90°$ and elevation angles range from $-90°$ to $+270°$ with the following resolutions:

- For the measurement, the azimuth angles of the source positions are increased in steps of $5°$ for azimuth angles between $0°$ and $\pm 45°$. The resolution is decreased to $10 - 15°$ for azimuth angles between $\pm 45°$ and $\pm 90°$. This means that measurement locations are less densely spaced at the sides of the head - which can clearly be seen in *Figure 13* (*Section 4.1*).
- The elevation angles of the source positions are distributed uniformly and are derived by dividing the full circle in 64 parts. This results in a resolution of $360°/64 = 5.625°$. Of those 64 positions, 50 were used for the measurement (ranging from $-45°$ to $230.625°$).

All in all, impulse responses were captured at 1250 measurement position for each subject which ensures sufficient spatial resolution as explained in [MBL07].

[20] KEMAR Manikin is a head and torso simulator based on average anthropometric dimensions of female and male human beings. It has become an industry standard for anthropomorphic testing. For more information see: `http://www.gras.dk/00012/00330/`

[21] See the corresponding paper [ADTA01] and [CIP98] (an informal document found on the EarLab website).

Measurement & Post-processing For the measurement of the HRTFs, Golay-Codes were used as an excitation signal at a resolution of 16 bit and 44.1 kHz. As explained in *Chapter 2.4*, this is a valid choice for system identification, although, more recent developments show a clear tendency towards sweep-based measurement methods due to practically achievable SNR values and possible separation of distorted signal parts. Undesired room reflections were mostly removed by shortening the impulse responses with a modified Hanning windows. The resulting duration of one impulse response is 200 samples, which corresponds to a length of about 4.5 ms @44.1 kHz. However, at low elevations, reflections from the subjects' knees and from the floor, and at high elevations, reflections from the ceiling, are still present in the IRs. The spectral magnitude of the free-field response without subject was removed via spectral division (see *Post-processing* in *Section 2.3*) from the raw HRIRs for compensation of the equipment and room transfer functions.

In addition to the impulse responses, anthropometric data of each subject was measured as shown in Figure 8:

- length parameters x_1 - x_{17}: head, pinna, neck, torso and shoulder sizes and positions
- length parameters d_1 - d_8: pinna details
- angles θ_1 and θ_2: angular position of the pinna

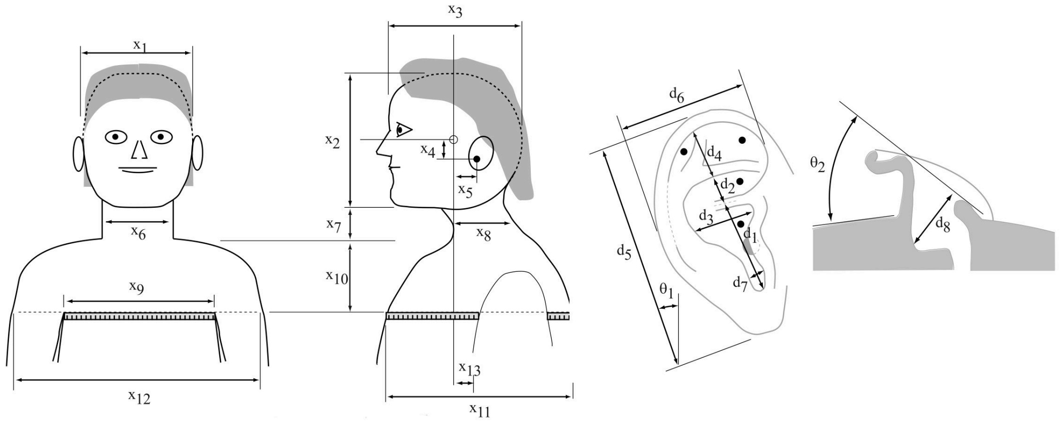


Figure 8: Overview of anthropometric data in the CIPIC database (from: [ADTA01])

The choice of parameters followed the proposition of Genuit[22] and aimed at investigating possible links between the anthropometry and HRTF features. Correlations in-between the anthropometric measurements of one subject were also examined but were found to be rather weak, in general. Although, a few parameters, for example d_1 (cavum concha height) and d_5 (pinna height), show fairly good correlation. However, it can be said that estimation of certain anthropometric parameters from others, such as estimating pinna dimensions from head and torso measurements, is not possible.

[22] Refer to: K. Genuit, Ein Modell zur Beschreibung von Außenohrübertragungseigenschaften, Rheinisch-Westfälische Technische Hochschule Aachen, Germany, 1984.

Data Storage The CIPIC originally used MAT files[23] to store the impulse responses. For each ear, one matrix of the dimension 25 x 50 x 200 (azimuth angles x elevation angles x number of samples) was used.
The EarLab database is website-based and uses a graphical interface to a search engine to access the data. The required search parameters are subject identification number and two ranges for azimuth and elevation angles. For a given position and subject, the impulse responses of both ears are available for download as either TXT files or MAT files. Additionally, a graphical plot of the IRs can be shown. The anthropometric data of a subject can be viewed as a table. Generally, the graphical online access is very quick and intuitive, particularly for users who are not familiar with MAT files or similar ways of data storage. However, it can be argued that this form of distributing HRTF data is not very suitable for scientific purposes, as it is rather inconvenient when standard scientific tools such as Matlab and Octave are used.

3.2 Listen HRTF database

The Listen project[24] aims at augmenting the reality through virtual soundscapes. HRTFs play a key role when placing a human being in such a virtual soundscape using headphones. For this reason, measurements of HRTFs and associated anthropometric data were performed at "Institut de Recherche et Coordination Acoustique/Musique" (IRCAM, Paris) and at AKG Acoustics in 2003. The results are available and documented on the IRCAM website [IRC03].

Geometry The measurements took place in an anechoic room with a volume of 324 m^2. It is interesting to note, that contrary to most of the other HRTF measurement setups with more than one elevation angle, a single loudspeaker was used for all measurement positions. This was made possible through a crane, driven by a stepper motor, which lifted the speaker to different heights for varying the elevation angle. The azimuth angle could be changed by a turntable on which the chair with subject was mounted. In order to minimize spatial errors, a head-tracking system, which monitors the head position, was used. It was connected to the measurement software and only allowed a measurement to be started when the head actually is at the desired position.
The elevation angles of the source used for the measurements range from $-45°$ to $90°$ with a constant resolution of $15°$. $0°$ elevation angle corresponds to a source position on the horizontal plane. The azimuth angles are all located on a $15°$ grid, but for higher elevations ($\geq °60$) only every second, forth or twenty-forth point is taken into account, as shown in *Table 2*. All in all, this results in 187 measurement points, which is rather limited compared to other databases such as CIPIC (see 3.1) and ARI (see 3.4), which both contain well beyond 1000 points.

[23] MAT is the internal format of Matlab®. Data that is available in Matlab® can be exported to MAT files. For more information see: `http://www.mathworks.com/help/pdf_doc/matlab/matfile_format.pdf`

[24] The former project website is no longer available. Some information can be found at `http://cordis.europa.eu/ist/ka3/iaf/projects/listen.htm`.

elevation angles	azimuth increment	measurement points
$-45°...45°$	$15°$	24
$60°$	$30°$	12
$-75°$	$60°$	6
$90°$	$360°$	1

Table 2: Number of measurement points at different elevation angles (adapted from: [IRC03])

Measurement & Post-processing The impulse responses were measured at a resolution of 24 bit and 44.1 kHz using a logarithimic sweep with a length of 8192 samples (corresponds to about 486 ms). A software-library for Max/MSP[25] controlled the entire measurement system including excitation signal playback, system response recording, the loudspeaker crane, turntable and head-tracker (see paragraph about *Geometry* above). The HRIR data are available in "raw" and "compensated" versions:
The raw data is the direct result of the deconvolution of the system response and, thus, has the full length of 8192 samples. Although the measurement room is theoretically anechoic, it is impossible to completely prevent reflections from the subject's lower body parts, the various mounting and measurement devices and - to some extent - even from the absorptive walls. Under these conditions, impulses responses with a length of almost 0.5 s are not optimal unless the influence of the room is desired - which is rather unlikely in case of an unechoic room, though. However, for sake of completeness and possible custom post-processing, it is still reasonable to also publish these "raw" data.
The compensated versions are based on the raw data. They are windowed to a length of 512 samples in the time domain. Then they are equalized using a reference soundfield: This reference is measured for all source directions with the microphones at the typical ear positions but without the subject being present. In the frequency domain, the minimum-phase parts of the measured HRIRs are equalized with the inverse of the reference soundfield magnitude of the corresponding source position. The compensated HRIRs are acquired by transforming the resulting filters back to the time domain. The transfer functions of the amplification system and the microphones were measured seperately for equalization purposes. Unfortunately, the online documentation does not give accurate information whether the effects of these devices are also already removed in the "compensated" data. Generally, the choice of exponential sweeps as an excitation signal, the measurement room conditions and setup as well as the post-processing indicate a good overall quality of the measurements except for the limited spatial resolution.

Data storage The impulse responses are available as WAV files[26] and as parts of MAT files. In case of the WAV files, the measurement location of a file is encoded in

[25] Max/MSP - the commercial counterpart to Pure Data (PD) - is a visual programming language mostly used for audio and multimedia applications. For more information see `http://cycling74.com/`.

[26] Waveform Audio File Format (WAVE or WAV) is a file standard for storing audio data on computers. It usually contains uncompressed data and is compatible with most common operating systems. For more information see `http://en.wikipedia.org/wiki/WAV`.

the file name. Though this might be sufficient for some purposes, it is obvious that this is a very limited form of describing measurement data.
In the MAT files, the actual impulse responses and corresponding metadata are separately stored for both ears in structure fields as shown in *Table 3*. Entries in the data matrix are primarily sorted by elevation angles (ascending) and secondarily by azimuth angles (ascending). In addition to the data provided in the MAT files, more detailed information about each subject and the measurement is available both online and as a downloadable XML [27] file. [28]

field name	**description**
content_m	2-dimensional data matrix (one line per position)
elev_v	a column vector of elevation angles in degrees
azim_v	a column vector of azimuth angles
type_s	a keyword which indicates the content type (e.g. FIR)
sampling_hz	sampling rate (in Hertz)

Table 3: Structure fields of a MAT file from the "Listen HRTF database"

3.3 T-Labs (TU Berlin)

The Spatial Audio Group of the Quality and Usability Lab at TU Berlin has performed two measurement sessions in 2010 and 2011 using a KEMAR manikin. The results are available in three different file formats (WAV, MAT and Open Directional Audio File Format (OpenDAFF), the latter one is not yet released, see *Data storage* below) on their website [WHGS12], along with documentation about the database and its acquisition. The first of the two sessions and its results are also documented and analysed in [WGRS11].

Geometry The first session was dedicated to acquiring the head-related impulse responses (HRIRs) of the manikin and took place in the anechoic chamber shown in *Figure 9*, right. The HRIRs of the full azimuthal space ($360°$) were captured with a resolution of $1°$. The manikin was mounted on a rotary turntable as shown in *Figure 9*, right. The loudspeaker was positioned in the horizontal plane, at ear height of the manikin. The measurement cycle was repeated four times with distances from source to head of 0.5 m, 1 m, 2 m and 3 m. In the second session, the 80 loudspeakers of the WFS studio are placed in a slightly odd-shaped rectangular form [29] with the manikin positioned in the center of the room. The procedure of acquiring the 80 BRIRs for each loudspeaker was repeated for head rotations from $-80°$ to $80°$ in steps of $2°$.
The geometrical description of the measurement setups in the MAT files is based on a

[27] XML is a flexible format that allows data and documents to be described and encoded in files (for more information see `http://www.w3.org/XML/` and `http://en.wikipedia.org/wiki/XML`)

[28] Online access: `http://recherche.ircam.fr/equipes/salles/listen/info.html`.
XML-file: `ftp://ftp.ircam.fr/pub/IRCAM/equipes/salles/listen/archive/INFO/INFO.zip`

[29] The exact shape is shown in *Figure 13*, *Section 4.1*.

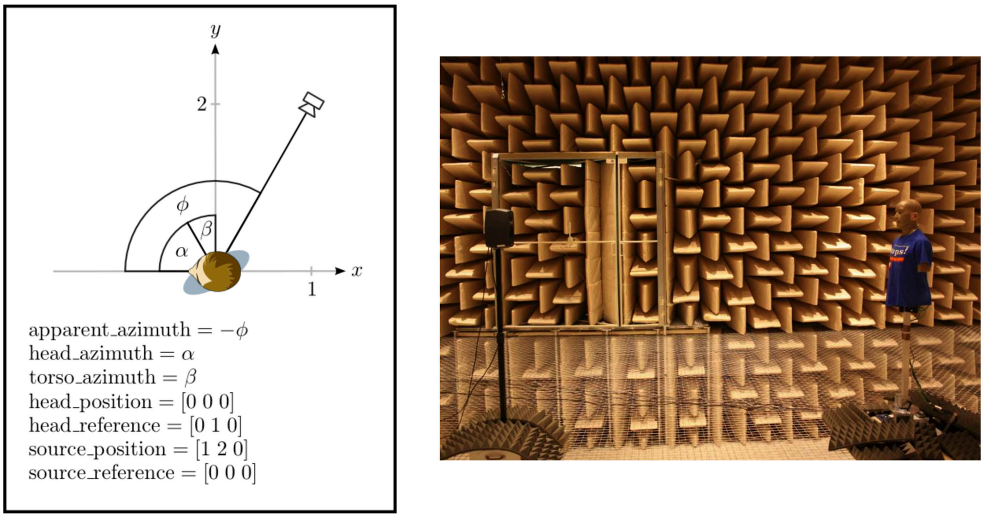


Figure 9: Geometrical description of measurement configurations (left, from: [WHGS12]); measurement setup with KEMAR manikin in the TU Berlin anechoic chamber (right, from: [WGRS11])

Cartesian coordinate system in which the position of the sound source and the head are specified (see *Figure 9*, left). "Reference directions" of both source and head are defined by points in the coordinate system which the head/source "looks at". This describes the basic measurement configuration without any movement.
In addition, head and torso can be rotated. Thus, azimuth and elevation angles are specified for torso and head. The head rotation (α in *Figure 9*) is relative to the torso rotation (β). The parameters "apparent azimuth/elevation" give the resulting angles between head and source (ϕ).
The geometrical description used in the OpenDAFF files seems not to be publicly documented, as these files are not release yet. However, it can be assumed that it will probably be similar to the description used in the MAT files.

Measurement & Post-processing In the first session, the excitation signal was a linear sine sweep with a length of 5.3 s. For the recording, a sampling rate of 44.1 kHz and an unusually high bit resolution of 32-bit floating point were used. In the post-processing stage, the impulse responses were truncated to a length of 2048 samples (corresponding to 46.4 ms). As already mentoined in *Section 3.2*, shorter IRs of about 200-500 samples are more beneficial because of possible undesired reflections in long IRs. An inaccurate loudspeaker positioning was detected and then compensated, based on the assumption that ITD and ILD have to be zero at an azimuth of $0°$. The loudspeaker magnitude response was removed from the HRIRs by applying an 128-point finite-impulse-response (FIR) filter with the inverse magnitude response. All in all, a mean SNR value of about 80 dB was achieved for the HRIRs. Headpohne equalization filters for a few different models were also designed and made public.

The second measurement session was located in a Wave Field Synthesis (WFS)[30] studio in Detmold, Germany. Binaural room impulse responses (BRIRs) were recorded for each of the 80 loudspeakers that are part of the WFS array in the studio. Although a few demo samples are already published online, the entire results are not yet publicly available.

Data storage In the WAV format, all impulse responses from one measurement cycle are combined into a multi-channel WAV file. For example, the IRs for a distance of 3 metres between source and head are all contained in a single WAV file in the following order:[31]

Channel 1: virtual source position $0°$ azimuth angle, left ear
Channel 2: virtual source position $0°$ azimuth angle, right ear
Channel 3: virtual source position $1°$ azimuth angle, left ear
Channel 4: virtual source position $1°$ azimuth angle, right ear
Channel 5: etc.

This format is very efficient for some purposes such as rendering virtual audio environments. However, it is not sufficient for sophisticated storage of measurement results.

The MAT files contain structure field entries which describe the measurement (usually referred to as "metadata" in this thesis) and the actual impulse responses (usually referred to as "data"). The metadata include information about the measurement room, the loudspeaker, the head, the sample rate and the geometry of the measurement setup as described above. The impulse responses are split into two fields for left and right ears respectively. A set of m-files[32] - usable in both Octave and Matlab® - is available for easier access to data and metadata.

The OpenDAFF (Open Directional Audio File Format) versions of the datasets are not officially available at the moment of writing, because - as stated on the website [WHGS12] - this format is still under development. Generally, OpenDAFF is a file format for storing directional audio data developed by the Institute of Technical Acoustics at RWTH University in Aachen, Germany. The intention is to facilitate the exchange and representation of directional audio data by providing the DAFF file format and a related open-source software package consisting of C++ and Matlab® libraries and tools. OpenDAFF uses its own numerical container and writes custom binary files.

[30] Wave Field Snythesis (WFS) is a holophonic technique which aims at reproducing the original soundfield of a recorded sound source. It relies on the idea of the Huygens-Fresnel principle which states that a wave front can be yielded by superposition of an infinite number of secondary waves, which are positioned on the envelope of the wave front at a given time. With WFS, loudspeakers are used as source points of these secondary waves. Therefore, WFS systems require a large number of loudspeakers in order to approximate the infinite number of secondary waves that would be necessary for a perfect representation of the wave front. For more information see: `http://www.holophony.net/Wavefieldsynthesis.htm`

[31] The internal strucutre of the multi-channel WAV files is described in the manual "Introduction to the SoundScapeRenderer (SSR)", 2012, by J. Ahrens, M. Geier and S. Spors, p. 18.

[32] M-files are script or function files containing Matlab® code.

3.4 ARI HRTF database

The ARI HRTF database has been acquired at the Acoustics Research Institute (ARI) which is part of the Austrian Academy of Sciences[33] located in Vienna, Austria. It is continually being expanded and at the time of writing, it consisted of data from more than 90 subjects. Most of the measurements were performed using the blocked-ear-canal technique with in-ear microphones (see *Figure 10*, middle). However, behind-the-ear HRTF measurements were also performed on some subjects (partially carriers of cochlea implants[34]) with the microphones mounted in a casing of a hearing-aid device (see *Figure 10*, left). For storage of the database, the proprietary "ARI HRTF format v2" is currently being used (see *Data storage* below).

Geometry The measurements are performed in a semi-anechoic chamber (6.2 m x 5.5 m x 2.9 m) with a reverberation time of about 55 ms. As shown in *Figure 10*, right, 22 loudspeakers are mounted on a circle in a vertical plane with elevation angles ranging from $-30°$ to $+80°$. The subject is placed on a computer-controlled rotating chair with the head located exactly in the center of the circle formed by the loudspeakers. The source positions during one measurement cycle cover the full azimuthal space ($0°$ to $360°$) with a resolution of $2.5°$ within a range of $\pm 45°$ and a resolution of $5°$ outside this range. The elevation angles of the sources increase in steps of $5°$ from $-30°$ to $70°$ plus one source position at an elevation of $80°$. For elevations other than $0°$, the number of azimuth positions is accordingly reduced with higher values to keep the spherical angles between the measurements points constant. For example, at an elevation of $80°$, only 18 HRTFs are measured. All in all, 1550 positions are measured for each subject.
The ARI HRTF format simultaneously uses all three different specifications of coordinate systems shown in *Table 1* in *Section 2.2* to describe the geometry of the measurement setup. This increases flexibility, although, they can also be derived from each other by simple arithmetic operations. In addition, the channel of the loudspeaker which emitted the excitation signal is given for each measurement, allowing to determine the elevation angle which corresponds to that loudspeaker.

Measurement & Post-processing Exponential sweeps with a duration of about 1.7 s and frequency range of 50 Hz to 20 kHz are used for the measurements. However, the single sweeps are interleaved and overlapped according to the Multiple Exponential Sweep Method (MESM)[35], which was also developed at ARI. This method allows to simultaneously use all 22 loudspeakers and, thus, to measure the HRTFs of all 22 elevations for one azimuth angle at one time. Playback and recording of excitation signals and system responses is done with the software AMTatARI which is part of ARI's ExpSuite

[33] See `http://www.oeaw.ac.at/english/`.

[34] Cochlear implants are a special kind of hearing aid devices for people with impaired inner ears. They are tiny electronic devices, which are surgically implanted in a person's ear and emit electronic neuron pulses, similar to those emitted by the hair cells of a normal hearing person. This can provide a sense of hearing to people who are often deaf or nearly deaf.

[35] For more information see the corresponding paragraph in section 2.4 and [MBL07].

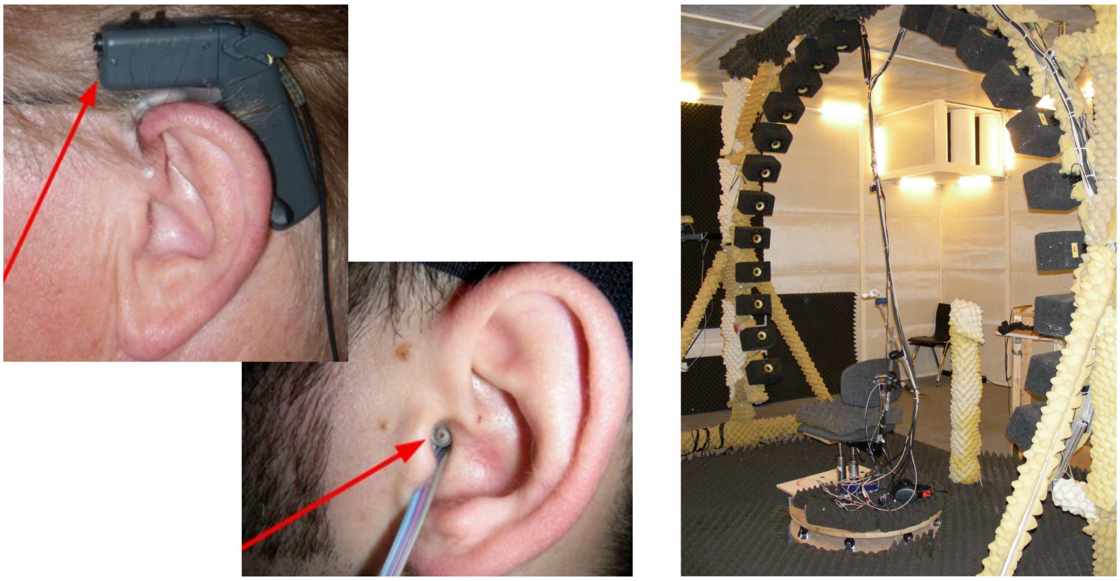


Figure 10: Behind-the-ear microphone (left), in-ear-microphone (middle), Semi-anechoic measurement room at ARI (right), (from: [Psy12])

framework (see below). In order to minimize the time-variance introduced by movements of the subject, a head tracking system is used to monitor the position of the subject's head. In case of exceeding a certain valid range of movement, the measurement for that particular azimuth is immediately repeated, which usually happens about three times per subject. This results in a total measurement time of about 20 minutes, which is not much considering the high spatial resolution.
In the post-processing stage, the influence of the measurement system is minimized by spectral division with the equipment transfer function derived from a reference measurement without a subject. The measurement results are available in three versions:

– long HRTFs: 50 ms
– short HRTFs: windowed to 256 samples corresponding to a length of 5.33 ms @ 48 kHz sample rate
– directional transfer functions (DTFs): windowed to 256 samples

For each subject, anthropometric data of head, shoulder and pinna are recorded. The parameters are based on those used in the CIPIC HRTF database (see *Section 3.1*), but are expanded by nine additional values which describe the pinna in great detail.[36]
A software package "ExpSuite", which was developed at ARI and is compatible to the ARI HRTF format, is available. It is based on Matlab and offers tools for measuring and viewing HRTFs, as well as a player for virtual sound positions utilizing HRTF data.

Data storage Both data and metadata are stored in MAT files according to the "ARI HRTF format v2" documented in [MM11]. All information of one subject is stored in a single file, but is split into three parts within the file:

[36] For more information see the measurement protocol sheet: `http://www.kfs.oeaw.ac.at/research/experimental_audiology/hrtf/protocoll.pdf` (in German)

- *stimPar* is a structure the fields of which contain general information about the dataset such as the sampling rate, the bit rate that was used for the measurements or the Subject ID.
- *hM* is a matrix witch contains the actual impulse responses. Its dimensions are N x M x R (length of IR x number of measurement positions x number of channels).
- *meta* is another structure containing all information that changes with each measurement position. The field *pos* is a matrix where the azimuth and elevation angles in three different coordinate systems (see paragraph about *Geometry* above) plus the excitation channel are listed for each measurement. Its dimension is M x 7 (number of measurement positions x 7). Other fields hold additional information for each single measurement like the amplitude of the excitation signal or the latency of the IRs relative to the general system latency.

3.5 MARL-NYU

MARL-NYU is a file format for storing HRTF datasets which was developed in 2011 by the Music and Research Laboratory (MARL) at New York University (NYU). Its goal is to create a standardized HRIR repository which allows to unite HRTF datasets from different sources into a single uniform database. Currently, it consists of contributions from the following databases:

- LISTEN (*Section 3.2*)
- CIPIC (*Section 3.1*)
- FIU (see corresponding paragraph in *Section 3.6*)
- KEMAR-MIT[37]

The MARL-NYU database is available for download on their website [AR12]. The download includes a documentation of both the file format and the repository [AR11a].

Geometry The positions of the sound sources in the measurement setup are simply described by azimuth and elevation angles and the distance between head and sound source. Azimuth angles range between $-180°$ and $+179°$, and elevation angles range between $-90°$ and $+90°$[38]. The actual angular range and resolution of the data changes, depending on which database it originates from.

Standardization process Existing and arguably also future HRTF databases are measured under different conditions using different setups and settings. This results in variations of various parameters and outside influences between different databases. For example, the filter length, the excitation signal or the angular resolution of the measurement points might change. Naturally, reflections from the room and non-linearities

[37] An HRTF data set of a KEMAR mannikin with 710 source positions that was created at MediaLab Institute of Technology of Massachusetts in 1995. For more information see: Gadner, B. and Martin, K. D. (1995). HRTF Measurements of a KEMAR. Journal of the Acoustical Society of America, 97(6):3907-3908.

[38] This definition of spherical coordinates is referred to as "navigational" in *Table 1*, *Section 2.2*

introduced by the measurement equipment will change, as well.

However, the goal of MARL-NYU is not only to create a common storage format but also to create a uniform HRTF database. For this purpose it is necessary to eliminate such differences. The potential of obtaining such a uniform database was evaluated by applying a standardization process as described in [AR11b], p. 3-5.[39]

The first step is to assure that all datasets use the same sample rate by applying a resampling algorithm where necessary. For evaluation purposes, a sample rate of 44.1 kHz was used. Furthermore, it is defined that the azimuth and elevation angular resolution must be $15°$ (with an elevation angle range of $-45°$ to $+90°$). In case that the original measurement locations do not match this grid, adjacent impulses responses/filters are linearly interpolated to acquire results for all grid points. The issue of varying filter lengths is solved by zero-padding where necessary and subsequent transformation into the frequency domain, resulting in 256-point HRTFs. Finally, normalization is applied by subtracting the mean value and dividing by a standard deviation.

According to [AR11b], this process resulted in a standardized repository that did not show any strong database-related differences. This could be very valuable for the research field of HRTF individualization. On the other hand, it is not unlikely that the interpolation between measurement introduces significant errors.

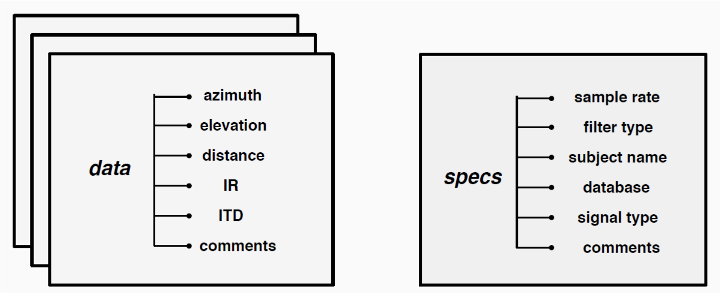


Figure 11: Outline of the MARL-NYU file format (from: [AR11a], p. 2)

Data storage All data and metadata of one subject are stored in a single MAT file. This MAT file consists of a *data* array and a structure called *specs*, as shown in *Figure 11*.

Each element of *data* is a structure which contains all information relevant to one specific measurement location: Azimuth and elevation angles, the distance from the sound source to the head, the IRs or transfer functions, and the corresponding interaural time difference (ITD). For each new measurement location, a new struct is appended to the *data* array. All information that applies to the whole dataset is stored in separate fields of the *specs* struct. For example, this includes the sample rate, the subject's name or the database from which the dataset originates.

For easier interaction with the database, a set of six Matlab® functions is provided.

[39] This process was just performed for evaluation. It has not been applied to the data available on the MARL-NYU website [AR12].

Three of the functions require a specific azimuth-elevation location:

- *findIR*: Returns the impulse responses (for both ears), ITD values and sample rate of that location
- *viewIR*: Plots the impulse responses in time and frequency domain
- *soundIR*: Plays back white noise convolved with the filters corresponding to the given location

There are two search functions and one export function available:

- *findSubject*: Returns *data* and *specs* for a given subject name
- *findDatabase*: Returns a cell array of file names which belong to a given database
- *exportAudio*: Allows to export the impulse responses contained in a given MAT file to a series of WAV-files

3.6 Others

After the more detailed description of databases in the previous sections, further databases and measurement setups will be briefly discussed in this section to provide a more complete picture. This includes some special cases, that need to be considered for a standardized file format.

NAGOYA This collection of HRTF data was published by Takeda Laboratory at Nagoya University, Japan. It consists of several separate datasets from the time between 1999 and 2010. Measurement specifications such as the sample rate and the length of the impulse responses vary. In fact, the impulse responses are mostly stored as plain text numbers in DAT files. The azimuth and elevation angles belonging to a certain impulse response are described in the folder structure and file names.[40] Unfortunately, the quality of the measurements cannot be assessed because the available documentation is very poor.

Florida International University (FIU) The Digital Signal Processing Laboratory at Florida International University, USA measured HRTFs and anthropometric data of 15 subjects in 2004. The database is not publicly available but requires a user account on their website [FIU04] which has to be requested via email. However, the data is included in the MARL-NYU database (see *Section 3.5*).
With twelve azimuth and six elevation angles - all in all 72 source locations - the resolution is rather low compared to other databases. For the measurements, the HeadZap system from AuSIM 3D[41], which is an HRTF measurement system designed for operation in rooms with less controlled acoustics such as offices, was used with Golay-codes as an excitation signal. The HRTFs are represented as 256-point minimum phase impulse responses with corresponding ITDs. At the sample rate of 96 kHz, this equals a length of about 2.7 ms. The anthropometric data of the subjects were retrieved from 3D head

[40] Also see their website: `http://www.sp.m.is.nagoya-u.ac.jp/HRTF/`

[41] See `http://www.ausim3d.com/products/AuWeb_headzap.html`.

scans which is quite a sophisticated approach compared to the mostly tape-based "do-it-yourself" methods used in other databases. Interestingly, only four parameters (head circumeference, ear width and length, concha volume) were extracted from the scans.
The elevation angles of each measurement, the impulse responses and the corresponding ITDs are stored in plain text files. Each file contains the measurement data for one azimuth angle. Within the file, all data and metadata are stored as plain text numbers in a certain order, so some calculations are required to access a particular piece of data. Basically, this is a simple but efficient data format for storing results from a fixed measurement setup. However, it is inconvenient and not very flexible, as even adding one measurement point would not be possible.

Fast near-field HRTF measurements using reciprocal method The Faculty of engineering at Toyama Prefectural University in Imizu, Japan has performed reciprocal HRTF measurements in the near-field in 2010. Miniature loudspeakers were inserted into a dummy-head's ear canals and 36 microphones were placed in a circle with a radius of 20 cm around the head to capture the impulse responses. This avoids the problem that the loudspeakers of a conventional setup can not be considered as point sound sources in the near-field. Furthermore, the reciprocal method is fast, as multiple impulse responses can be measured at once. Therefore, it is also more robust against time variance than slower methods.
On the other hand, the frequency range is limited by the performance of the miniature speakers. Results from the paper indicate sufficient signal-to-noise ratios in the range between 400 Hz and 20 kHz. Also, the sound pressure levels are restricted when using this method with human beings instead of a dummy in order to prevent harm to the subject's hearing sense.
For the measurement, optimized Aoshima time-stretched pulses (OATSP) were used as an excitation signal at a sampling rate of 48 kHz. The choice of OATSP is a bit bizarre as the corresponding paper[42] explicitly suggests it for "measurement of very long impulse responses" which is definitely not the case with HRIRs. On the other hand, time-stretched pulses in general perform pretty well in the practical tests carried out in [SEA02] (not specifically for HRTFs, though), with a SNR of 15-20 dB higher than MLS and IRS, and only 3 dB lower than sweeps. One pulse sequence was 1.37 s long, but was repeated 20 times per measurement resulting in a total measurement time of almost 30 s. This is a good result for acquiring HRTFs in the full azimuthal space with a resolution of 10° (about 0.75 s per HRTF). However, it has to be argued that the MESM (see *Section 2.4*) is even more effective as it allows to simultaneously measure 22 HRTFs in less than 8 seconds (about 0.36 s per HRTF) with significantly higher SNR in a greater frequency range.
No information about the way of data storage is provided in the paper [MH10].

[42] Y. Suzuki, F. Asano, H.-Y. Kim, and Toshio Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses", J. Acoust. Soc. Am. Vol. 97(2), pp. 1119-1123, 1995

ROMEO-HRTF ROMEO-HRTF is part of a robotic project at Telecom ParisTech, France[43]. Romeo is a humanoid robot intended to serve as an assistant for people who suffer from loss of autonomy. It has 16 microphones built into its head. The measured HRTFs of the robot will be used to improve the sound source localisation and separation in its algorithms.

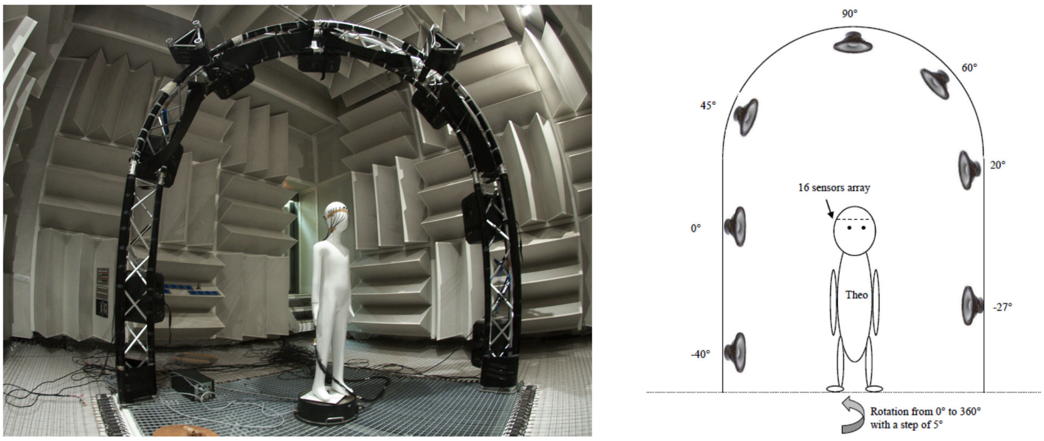


Figure 12: Measurement setup in the aneochic room for acquisition of the 16 HRIRs from the robot dummy "Theo" (left), schematic overview thereof (right), (from: [MG11])

In *Figure 12*, the measurement setup for acquisition of the HRIRs from a dummy of the robot is shown. It is used to cover source positions with 72 azimuth angles ($5°$ angular resolution) at different 7 elevation angles (corresponding to the 7 loudspeakers; increments varying from $20°$ to $30°$), resulting in 504 measurement points. The azimuth angles are realised by rotating the entire dummy via a turntable. For the measurement, Golay-codes were used as an excitation signal at a sampling rate of 48 kHz. The results are available as both HRIRs and HRTFs which are stored in MAT files:[44]

The HRTFs are stored as complex spectra in a matrix of the dimension 16 x 504 x 513 (number of receivers/microphones x number of measurement locations x number of samples). In the same file, azimuth and elevation angels for each measurement are stored as vectors. Furthermore, three numeric variables contain the frequency range (0 - 24 kHz) and the number of frequency bins (513). Storage of the HRIRs is split up in multiple MAT files. Each file contains the IRs for one measurement position, which is described in the file name.[45] Within the file, one vector with a dimension of 360 x 16 (number of time samples x number of receivers/microphones) holds all impulse responses.

This database certainly cannot be compared to conventional databases due to its special

[43] See http://www.tsi.telecom-paristech.fr/aao/en/2011/03/31/romeo-hrtf-a-multimicrophone-head-related-transfer-function-database/ and the corresponding paper [MG11]. For general information about the Romeo project see the website http://projetromeo.com/index_en.html

[44] The data storage is described in a readme-file, which is available online: http://perso.telecom-paristech.fr/~maazaoui/Sons/readme.txt

[45] For example the file "Theo_el020 _az090.mat" contains the impulse responses for a source position of $90°$ azimuth angle and $20°$ elevation angle.

purpose. However, it is an interesting example of an HRTF database that exceeds the usual setup. A common storage format or archive should also be able to handle such more complex configurations.

Universität Oldenburg Modern hearing aid devices typically process signals from 2-3 microphones per ear. This enables the implementation of sophisticated algorithms such as beamforming and sound source localization. In 2009, the institute for "Medizinische Physik" (medical physics) at University Oldenburg, Germany, performed a number of measurements, the results of which are intended to serve the evaluation of such multi-channel hearing aid devices. A detailed description and documentation is available in [HEA$^+$09].
For the measurements, a head and torso dummy was used. In addition to the common in-ear microphone at the exit of the ear canal, three microphones were placed in an empty hearing aid casing behind the ear, resulting in 8 channels for both ears.
A set of 8-channel HRIRs was recorded for two distances (80 cm and 3 m) in an anechoic chamber. Regarding source positions, three azimuth angles ($-10°$, $10°$, $20°$) and three elevation angles ($-180°$, $5°$, $+180°$) were used. Furthermore, 8-channel directional room impulse responses were captured in four realistic environments: 2 offices, a university courtyard and a cafeteria. The loudspeaker and head positions and orientations were chosen so as to get realistic and meaningful results for each environment. Additionally, ambient background noise and some typical sounds such as telephone ringing in the office were recorded for more realistic evaluation possibilities.
For obtaining the impulse responses, Modified Inverse Repeated Sequences (MIRS) (discussed in *Section 2.4*) were used as an excitation signal with a sampling rate of 48 kHz and a resolution of 32 bit. The usage of MIRS ensured immunity against harmonic distortion while exhibiting a white spectrum which is less disturbing to people in public sites than other signal types such as sine sweeps. High SNR values of 86 - 105 dB could be achieved, depending on the measurement location. All impulse responses were windowed to a length of 2400 samples (corresponding to about 50 ms), which is probably appropriate for the BRIRs recorded in realistic environments but which is arguably too long to get rid of all reflections in the HRIRs from the anechoic room. In addition, level normalization was applied to the impulses responses to eliminate the level differences between the different measurement locations.
A download link to the measurement results is provided via e-mail upon registering on their website[46]. The data is available either as multi-channel WAV files or as MAT files. The impulse response of each source position is stored in a separate file resulting in a great amount of files which are rather inconvenient to handle. Within a MAT file, metadata such as the azimuth and elevation angle, sample rate and bit resolution are contained as variables. The actual data is stored in a matrix with a dimension of N x 8 (number of time samples x number of receivers).
Considering this database is a another good test for the flexibility of a common storage format because of its uncommon microphone number, different measurement positions and additional ambient recordings. For this reason, it has been used to evaluate the

[46] `http://medi.uni-oldenburg.de/hrir/`

potential of the proposed Spatially Oriented Format for Acoustics (SOFA) in *Section 4.4*.

Aalto University In 2012, the School of Electrical Engineering at Aalto University, Finland, published a database consisting of HRIRs from 21 subjects measured in an anechoic chamber:[47] 240 source positions were captured at two different distances (0.68 m and 1.35 m). A tracking system based on infrared cameras was used to obtain the actual head position and orientation during the measurement process. The impulse responses are stored in a MAT file, which contains four matrices for left/right ear and near/far loudspeaker position, respectively. Each matrix has a dimension of 8192 x 240 (number of time samples x number of measurement positions). A second MAT file contains four 240-element vectors with the actual azimuth and elevation angles that were measured by the tracking system.

[47] See `http://www.acoustics.hut.fi/publications/papers/aes133-hrtf/`.

4 Standardization of measurement and storage

A new HRTF storage format should certainly be compatible to existing formats while being open and flexible enough for future developments, trends and special configurations exceeding the "classic" HRTF measurement setup. In this chapter, the main issues associated with the creation of a common storage format are revealed by summarizing the examination of some of the more important existing databases in the previous *Chapter 3*. An example from the Universität Oldenburg dataset is converted to the proposed new storage format "Spatially Oriented Format for Acoustics" (SOFA) to practically show how some of the raised issues are addressed in SOFA.

4.1 Geometry issues considering existing databases

Describing the geometry of a specific HRTF measurement setup is relatively simple as shown in *Section 2.2*.[48] All databases of free-field HRTFs presented in this thesis use two angles and the source-head distance ($\widehat{=}$ radius) with the subject's head located in the origin of the coordinate system for specifying measurement point locations. However, the databases' setups and file formats differ in many details:

- The number and the location of measurement points, the radius and the exact definition of the coordinate system vary for each database. In the CIPIC database (*Section 3.1*), the source positions cannot be changed, so for conversion to their data format, existing measurement points must be adjusted to fit their grid.[49] The other databases mostly use vectors containing the angles, so adoption to a different grid would be possible.
- Not all databases explicitly specify the radius (like the Listen HRTF database, *Section 3.2*).
- In addition, the definitions of coordinate systems also vary: For example, in the CIPIC database, horizontal-polar coordinates (also referred to as interaural-polar) are used, in the MARL-NYU format, navigational coordinates are used, in the Listen HRTF database, geogrpahic sphercial coordinates are used and in the ARI format all three of them are used at once to be more flexible. As conversion between these coordinate systems is a rather simple calculation, allowing the user to choose the coordinate system which fits the purpose of their database best would be sensible.

To sum it up, with regard to standard HRTF measurements under free-field conditions, the ARI and the MARL-NYU file formats contain the most complete geometrical description.

When taking other setups, such as BRIR measurements with variable source positions (like T-Labs and Universität Oldenburg, see *3.6* for the latter), into account, additional considerations are necessary:

[48] The issues related to the description of the measurement room were already discussed in *Section 2.2* about the geometry of HRTF measurement setups and thus will not be covered any more in this chapter.

[49] See the following note on the ARI website: `http://www.kfs.oeaw.ac.at/research/experimental_audiology/hrtf/ariandcipicerror.html`

- Generally, the loudspeakers are no longer necessarily located on a sphere surface but may have arbitrary positions in the room. Different sound incidence directions might be obtained by rotation of the subject and this process can be repeated for various source positions.
- The subjects' positions might vary as well, which means that the geometrical description has to be extended by a "listener" with a corresponding position. When considering the case of three microphones mounted in a hearing-aid device casing per ear (like in the Universität Oldenburg's database), their positions should all be uniquely described. Still, as their relative position to the subject does not change, it would be inconvenient to re-specify all six microphone positions when the subject moves. Thus, positions of microphones (or ears) should be defined relative to a subject's "anchor point"; moving the subject is then simply described by moving this "anchor point".
- A listener which acts as an anchor point for relative receivers (microphones, ears) requires both a location and a unique definition of a direction in which it "looks". This allows to describe the rotation of the listener to simulate different sound incidence directions at the same location without source movement. In some cases, like when the hearing-aid device microphones have a relevant directional pattern,[50] it might be necessary to even specify a direction for the receivers (relative to the source).
- In most rooms, defining arbitrary source and listener positions for BRIRs is more practical in Cartesian coordinates (like in the T-Labs database); however, an additional rotation of the subject can more simply be described in spherical coordinates. Being able to specify positions and orientations of all objects each in different coordinate systems would therefore be practical.
- Generally, it is possible that apart from source and listener position and orientation, other parameters change as well between the measurements. For example, the T-Labs format defines an additional azimuth angle between head and torso. The possibility that two measurements are performed for the same positions and are only distinguished by a different head against torso angle (or any other additional parameter), must be taken into account. To give an example, in a case where measurements with three different head against torso angles are performed for each source/listener position and direction, it would be sufficient to just provide the information of all measurement positions (vector of length M) plus the information about the three head against torso angles (vector of length 3) which are used for each of the measurement positions. However, this approach complexes the association of the stored measurement data (impulse responses) with the information about the measurement process (source/listener positions/orientations, additional parameters). Practically, a simpler solution would be to repeat each measurement position three times in the geometry description (vector of length $3 \cdot M$ and store the additional parameter(s) with separate entries for each measurement position (additional vector(s) of length $3 \cdot M$).

With regard to a common storage format, the challenge is to find a general description which is flexible enough to handle all special requirements of each individual setup including possible future applications without becoming overly complicated.

[50] For example, the signals from cardioid capsules can be processed to obtain a variable beamforming pattern. This helps the hearing impaired to focus on specific sound sources like a conversation partner in background noise.

The issue of how to realise the different source positions was already raised in *Section 2.2*: Basically, the loudspeaker and/or the subject can be moved/rotated. Some laboratories move and rotate the loudspeakers, while others use fixed loudspeakers and rotate the subject. The question is whether these differences should be reflected when describing the measurement setup in a storage format. In fact, a literal description of the measurement setup is possible in a flexible storage format. However, this would result in unique geometry characterizations for each laboratory which does not really coincide with the idea of a common storage format to alleviate data exchange. Therefore, for sake of simplicity it is suggested to find a simple common description for standard setups such as free-field HRTFs. In any case, accompanying documentation of the measurement should contain accurate sketches and any necessary facts about of the actual setup and the measurement process.

Different measurement position grids in existing databases A comparison of the measurement positions in some of the databases presented in this thesis is shown in *Figure 13*. The sophisticated spatial resolution of the CIPIC and ARI databases is clearly evident and also their varying resolution depending on the azimuth angle. Of the remaining databases, only Listen HRTF and Aalto capture a rather complete range of azimuth and elevation angles. In the other cases, only one or a few elevation angles were used, partly because of special applications (like hearing-aid device evaluation in the Oldenburg database).

Even if the different source position grids can successfully be specified in a file format, it is questionable whether each laboratory using their own grid is the optimal solution. For some research purposes - in particular with regard to HRTF modelling - it is important to have access to data from many subjects. Thus, merging data originating from different laboratories to obtain a larger number of subjects would be required. However, it is difficult to use HRTF data when the source positions are not identical, because even relatively small changes in the source direction might result in significant changes of the HRTF features. For certain spatialization applications, interpolation might help to overcome the problem and obtain uniform spatial sampling (as proposed by Andreopoulou and Roginska in [AR11b]), but for many research purposes this is not applicable. To give an example, Hölzl [Höl12] combined the ARI and Listen HRTF databases for the coinciding source directions to increase the number of subjects available for his research. ARI has a high resolution of 1550 measurement points and Listen HRTF has 187 measurement points, but only 44 points could be used in his work. Thus, the introduction of standardization schemes for the usage of measurement points in addition to the creation of a standard file format might be worth considering. However, at the time of writing, this seems to be a goal that is not yet realisable.

It is interesting to observe, that of all the databases presented in this thesis (most of which are listed in *Table 5*), only two (CIPIC and ARI) clearly meet the spatial resolution requirements discussed in *Section 2.2*. Partly, this can be related to particular purposes, which do not require as high resolutions. However, for most laboratories the benefits of acquiring HRTFs at more than 1000 source positions for high spatial resolution seem not to justify the increased effort. The future creation of more and larger multi-purpose HRTF databases with sufficient source positions would thus be important.

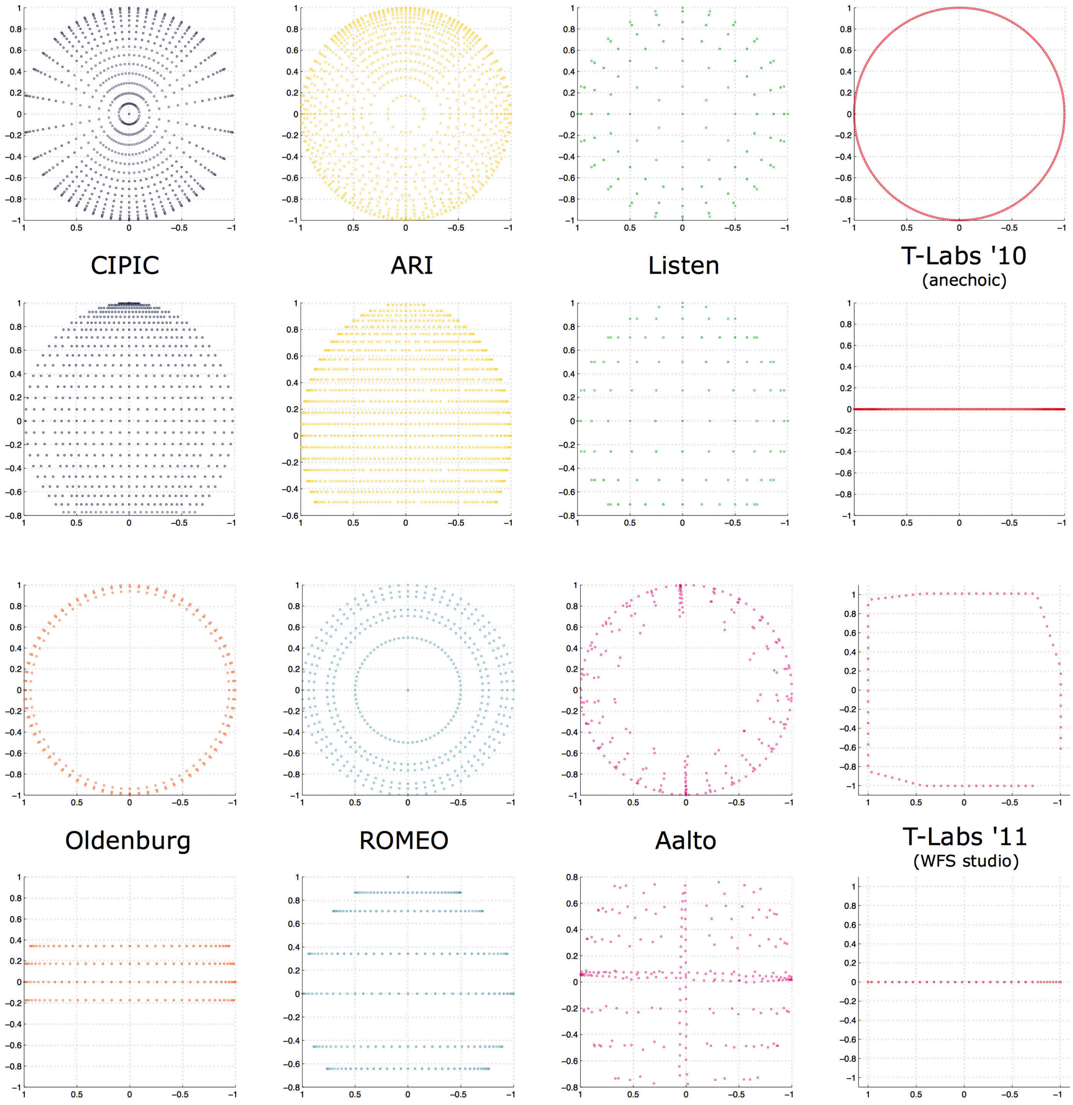


Figure 13: Measurements point grids of various HRTF formats, all normalized to the same radius. (upper figure: front view, lower figure: ground plan)

4.2 Measurement

All institutes acquiring the databases presented in this thesis more or less followed the basic principles of HRTF measurement introduced in *Chapter 2*. Of the more important databases, Listen HRTF and T-Labs were both measured in anechoic rooms, ARI in a semi-anechoic room while CIPIC does not give any explicit information - judging from pictures it seems to be a usual laboratory room with some acoustic absorption pyramid foam. Regarding the choice of system identification method, it is surprising that as shown in *Table 4*, the group of binary sequences is quite commonly used despite the advantages of the sweep-based methods shown in *Section 2.3*. Using sweeps arguably results in improved quality of the measurements due to higher signal-to-noise ratios and possible rejection of non-linearities, while the advanced variant MESM even allows to significantly speed up the measurement process without any disadvantages.

Type	Name	Database
binary sequences	Golay-codes	CIPIC, ROMEO
	OATSP	Toyama University (reciprocal method)
	MIRS	Universität Oldenburg
sweeps	linear sweep	T-Labs
	logarithmic sweep	Listen HRTF
	MESM	ARI

Table 4: Overview of the system identification methods used for acquisition of the databases presented in this thesis (where given)

Wherever documented, the measurements were performed with in-ear microphones using the blocked-meatus technique. For the Listen, ARI and Aalto databases, head-tracking was used to control the actual position of the subject's head. The first two just prevented invalid measurements, whereas in the Aalto database, the measured positions are stored in the database which can clearly be seen in *Figure 13*. Although this is great additional information, it would be important to also include the originally targeted angles for easier access of a specific HRTF, which they did not do, though.

The post-processing stage is also similar for most databases and includes removing the room and equipment transfer function as well as windowing in the time domain. Some databases use quite long windows - in particular Aalto and T-Labs with 8192 and 2048 samples, but also Listen and ROMEO with 512 samples. Usually, about 200 samples (corresponding to about 4.5 ms @44.1 kHz) contain all significant data, otherwise undesired noise or reflections from the room might be introduced.

For future measurements, the MESM should be more often used, as it currently appears to be the most advanced method for HRTF acquisition. Head-tracking should also become a standard so as to guarantee a certain accuracy regarding the spatial positions. Also, the recording of anthropometric data (for example such as in the CIPIC and ARI databases) is recommended to support the research efforts in HRTF individualization. The approach to retrieve the anthropometry from 3D head scans which was implemented for the FIU database is an innovative idea in this domain.

4.3 Data storage in a standardized file format

The examination in *Chapter 3* shows that, occasionally, plain text files or WAV files are used for storing measurement results, which might be an appropriate solution for the likes of simple spatialization tasks. However, for scientific purposes, MAT files have

become the standard solution in the majority of cases. Some basic patterns like storage of the actual impulse responses or filters in a data matrix along with some vectors and variables containing additional information about the source positions and measurement setup (metadata) are similar throughout all databases.
Still, as shown in *Table 3*, the particular dimensions of this data matrix are different for all databases (except Listen HRTF and Aalto). Storing the results from each receiver/ear in separate matrices is the most common approach. The MARL-NYU format even utilizes one matrix per measurement point resulting in an array of matrices. Both cases have the advantage of simple 2-dimensional data matrices, however, larger and more complex setups immediately result in a confusingly high number of matrices. To the contrary, ARI and ROMEO use a third matrix dimension for the receivers (ears) which allows to get by with a single matrix independent of the complexity of the setup. It might be a matter of personal taste, but a single 3-dimensional matrix always seems to be simpler to handle than multiple 2-dimensional matrices.

- R ... number of receivers (for example ears/microphones)
- N ... number of time samples/frequency bins (windowed versions are considered when available)
- M ... number of measurement points
- Az ... number of azimuth angles
- El ... number of elevation angles

Database	matrix dimension	R	N	M	Az	El	number of matrices
CIPIC	Az x El x N	2	200	1250	25	50	2 (one per ear)
Listen	M x N	2	512	187	1-24	10	2 (one per ear)
T-Labs '10	N x M	2	2048	360	360	1	2 (one per ear)
ARI	N x M x R	2	256	1550	18-91	22	1
MARL-NYU	N x R	2	varying				M
ROMEO	R x M X N	16	513	504	72	7	1
Oldenburg	N x R	8	varying				M
Aalto	N x M	2	8192	240	12-72	7	4 (per ear & distance)

Table 5: Overview of measurement points and dimensions of the data matrix in the MAT files of different HRTF formats

Additional information about the measurement (metadata) is available in all databases, but their organisation varies. For example, the meatdata are usually stored in variables, vectors and matrices on a "global" level within the MAT file. In the LISTEN format, though, the "global level" only comprises of two structs for the left and right ear, respectively, which both contain the corresponding impulses responses and the metadata. This means that all metadata is stored two times at different locations which is redundant and confusing. In the CIPIC format, the source positions are not even included in the MAT files but the vectors for the azimuth and elevation angles are described in Matlab® code-style in the documentation [CIP98]. The T-Labs, ARI and MARL-NYU formats are the most accomplished formats in terms of flexibility and expandability. However, the latter two are more or less restricted to describing the classic free-field HRTF setup, so considering more complex geometrical setups as described *Section 4.1* becomes

increasingly complicated if not impossible, depending on the exact purpose. The T-Labs format is quite flexible regarding the geometry, however, the number of receivers is basically hard-coded to two regarding both the geometry description and data matrix.
Although MAT files are widely used for scientific research purposes, it is questionable whether it is the best choice as the basis for a standardized file format. On the other hand, development and maintenance of a custom file container which handles the binary file input/output operations requires a lot of effort.[51] Using an existing file container that takes care of these tasks while providing close and simple interaction with commonly used software such as Matlab®, Ocatve or custom C++ and Java applications might thus be the optimal solution. As larger HRTF databases quickly reach considerable data sizes, included file compression would be a convenient feature. netCDF[52] is an example of such a file container and is therefore considered as the most promising choice for the upcoming HRTF file format SOFA at the time of writing.
To sum it up, the storage formats of most databases are sufficient for their particular purpose; some are more flexible and allow storage of various setups, but none is general enough to serve as a basis for a internationally standardized format for storing HRTFs, BRIRs and similar kinds of directional audio data acquired utilizing arbitrary setups. However, this is necessary if a commonly accepted and possibly even officially confirmed agreement on a standard HRTF format is to be found. Therefore, various research institutes have joint forces and are working on a such a standard, at the time of writing, which was proposed in an AES convention paper by Majdak et al. [MIC$^+$13].

4.4 Exemplary representation of directional audio data in SOFA

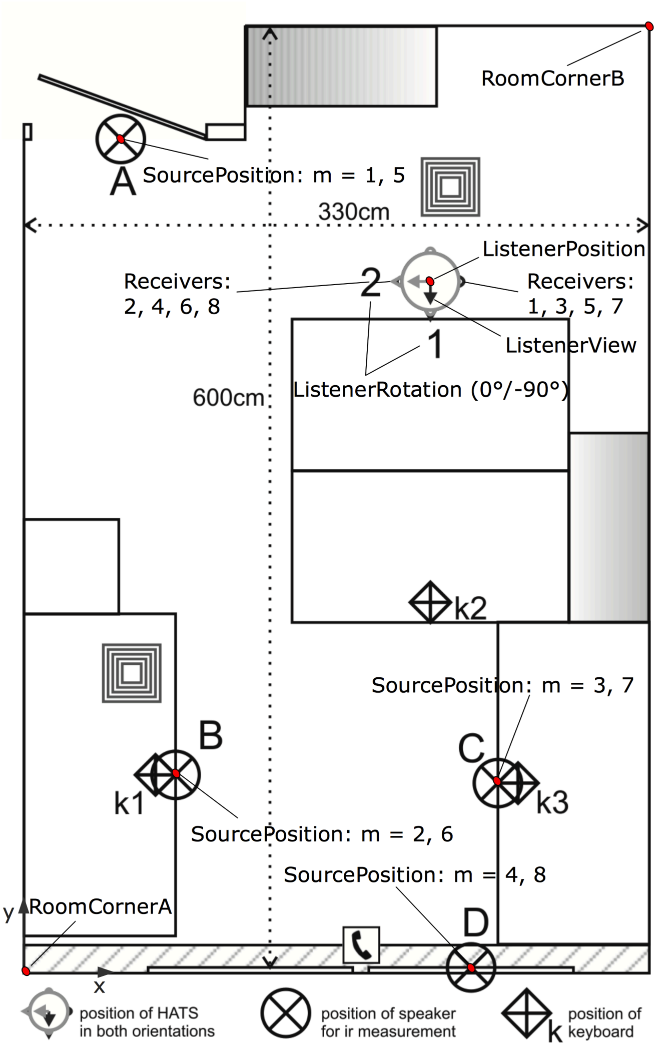
The Spatially Oriented Format for Acoustics (SOFA), proposed in [MIC$^+$13], is a collaborative effort to create an HRTF file format that can serve as the basis for the future HRTF data exchange format which will be officially approved by the Audio Engineering Society.[53] It addresses and provides solutions for most of the issues and difficulties raised in *Sections 4.1* and *4.3*. In order to evaluate the potential and flexibility of the approaches specified in SOFA, the conversion of data from the non-standard database acquired at Universtät Oldenburg (presented in *Section 3.6* of this paper) to SOFA is considered.[54]

[51] The openDAFF format mentioned in conjunction with the T-Labs database (*Section 3.3*) is an attempt in this direction. However, compatibility and support seem not to be on a larger scale and its geometry description does not seem to be flexible enough (measurement point resolution must not vary within a dataset, according to [MIC$^+$13]).

[52] See `http://www.unidata.ucar.edu/software/netcdf/`

[53] A corresponding standardization project "AES-X212" was already approved by an AES sub-committee and assigned to the working group SC-02-08 on Audio File Interchange. See `http://www.aes.org/standards/meetings/init-projects/aes-x212-init.cfm`.

[54] Note that at the time of writing, the SOFA specifications were still subject to change. However, the underlying principles and approaches are already well-established and evaluated. For the current specifications see the file section on the SOFA wiki page at SourceForge: `http://sourceforge.net/projects/sofacoustics/files/`.



RoomCornerA: (0, 0, 0)
RoomCornerB: (6, 3.3, 2.7)

SourcePosition@A: (0.52,5.27,1.8)
SourcePosition@B: (0.79,1.25,1.1)
SourcePosition@C: (2.52,1.23,1.1)
SourcePosition@D: (2.38,0,1.8)

ListenerPosition: (2.16,4.40,1.1)
ListenerView: (0,-1,0)
ListenerUp: (0,0,1)
ListenerRotation (azimuth):
0° for looking straight
-90° for looking over right shoulder

Receivers:
1/2: in-ear mics
3/4: front behind-the-ear mics
5/6: middle behind-the-ear mics
7/8: rear behind-the-ear mics

m	SourcePos & ListenerRot	m	SourcePos & ListenerRot
1	A 1	5	A 2
2	B 1	6	B 2
3	C 1	7	C 2
4	D 1	8	D 2

m ... measurement index

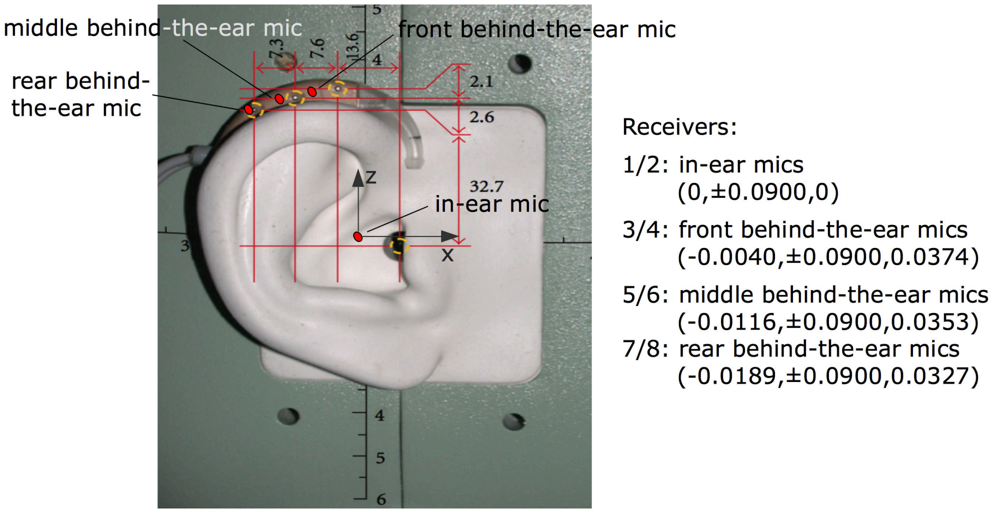
Heights (z-coordinate) are not given and were assumed from the description in the documentation for the exemplary purpose

Figure 14: Sketch of "Office II" where head-related room impulse responses were acquired at four loudspeaker positions and for two orientations of the head-and-torso simulator. The corresponding designations of the objects and conditions in SOFA and their positions and orientations are indicated. (adapted from: [HEA$^+$09])

The conditions of the measurements at "Office II" are shown in *Figure 14*. The room type in SOFA is set to "shoebox" which means that it is defined by the positions of two opposite corners (*RoomCornerA* and *RoomCornerB*) in an arbitrary global coordinate system.[55] Additional parameters such as reverberation time can easily be stored as metadata attributes. The measurements were performed using a loudspeaker and a head-and-torso simulator (HATS) with an in-ear microphones and three behind-the-ear microphones in both ears, respectively. In SOFA, this corresponds to a *Source* with a single *Emitter* (single one-way loudspeaker) and a *Listener* (HATS) with 8 *Receivers* (microphones). The positions of the *Source* and the *Listener* are specified in the same

[55] Covering the details of the SOFA specifications is not the purpose of this thesis and will only be explained where absolutely necessary. For more complete explanations and definitions please refer to the link to the specifications above.

Cartesian coordinate system as the room corners. Measurements were performed at four different source positions with the HATS "looking straight ahead" and "over the right shoulder", respectively, resulting in a total of 8 measurements. The positions of the *Emitter* and the *Receivers* are specified in a local coordinate system (see *Figure 15*), which is always coupled with the *Source* or *Listener* position. This way, the orientations of the HATS can be varied by changing the azimuth angle of *ListenerRotation* and all receivers are also rotated correspondingly. The *Source*/*ListenerView*- and *Up*-vectors define the basic orientation of *Source* and *Listener*.



Receivers:

1/2:	in-ear mics (0,±0.0900,0)
3/4:	front behind-the-ear mics (-0.0040,±0.0900,0.0374)
5/6:	middle behind-the-ear mics (-0.0116,±0.0900,0.0353)
7/8:	rear behind-the-ear mics (-0.0189,±0.0900,0.0327)

Figure 15: Positions of the in-ear and behind-the-ear microphones at the right ear of the head-and-torso simulator given in a local coordinate system with the y-axis and the interaural axis coinciding. Each of the microphones corresponds to one *Receiver* in SOFA. (adapted from: [HEA$^+$09])

The actual impulse responses and some additional information such as the sampling rate were directly read from the MAT files provided by the department for "Medizinische Physik" at Universität Oldenburg and then stored in SOFA. The data were spread across 8 files (one for each measurement) and were stored in N x R matrices (number of samples x number of receivers) which were re-composed to a M x R x N matrix (M ... number of measurements) for storage in SOFA. The geometrical parameters had to be manually derived from the sketches and descriptions provided in the documentation of the data ([HEA$^+$09]) as they are not contained in the original MAT files. These parameters are usually either given as 1 x 3 or M x 3 (in this case: 8 x 3) vectors/matrices, depending on whether a parameter changes or remains constant throughout the measurement data stored in the file.

In this section, it is shown that the results of special measurements for hearing-aid device research can easily be stored in SOFA with relatively little effort, even though an arbitrary example was chosen. The specifications of SOFA are flexible enough to also allow a relatively detailed and complete description of the geometrical measurement situation within the same file despite the non-standard configuration. This demonstrates the high potential of SOFA as a common storage format in the context of directional audio data.

5 Conclusions

In this thesis, the acquisition and storage of head-related transfer functions (HRTFs) and similar directional audio data has been discussed with regard to standardization, particularly of the file format. It also serves as a synopsis of current trends and possibilities in the field of HRTFs and beyond. This is important for both the standardization process and for alleviating access to the topic for readers who want to quickly get a deep insight in this field for whatever reason.

In order to demonstrate the motivation for standardization efforts, the role and possible applications of HRTFs in spatial hearing and the growing field of virtual environments have been shown. The foundations for understanding the issues related to standardization were laid by discussing the description of the geometrical aspects of measurement setups and the measurement process itself. The most important impulse response measurement methods and their individual strengths and weaknesses in the field of HRTF acquisition were discussed so that assessment of the measurement quality is possible. This is valuable for both evaluating existing or planning future measurements. The potential of sweep-based measurements - in particular the Multiple Exponential Sweep Method - due to high achievable signal-to-noise ratios and immunity to distortion was explained and promising alternatives such as time-variant system identification were shown. Existing databases with different purposes were examined in order to provide a complete picture of the diversity of real-world scenarios and to give the interested reader ideas what to consider when setting up an own measurement. A need for more high quality standard HRTF databases with high spatial resolution and a large number of subjects has been made evident.

The theoretical and practical observations from the previous parts were summarized and issues related to standardized measurements and data storage were raised. This includes a presentation of the requirements and approaches to a flexible description of the geometry of measurement setups. The successful exemplary representation of arbitrary directional audio data in SOFA has proved the potential of this proposed common storage format. To sum it up, this thesis provides an extensive introduction to the current issues related to HRTF measurement, storage and the standardization thereof.

References

[ADTA01] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," Ocotber 2001.

[AE09] C. Antweiler and G. Enzner, "Perfect sequence LMS for rapid acquisition of continuous-azimuth head related impulse responses," in *WASPAA*, 2009, pp. 281–284.

[AR11a] A. Andreopoulou and A. Roginska, "Documentation for the MARL-NYU file format, Description of the HRIR repository," Ocotber 2011, available: http://marl.smusic.nyu.edu/projects/HRIRrepository/content/documentation.pdf (date of access: November 2012).

[AR11b] ——, "Towards the creation of a standardized HRTF repository," in *131st Convention of the Audio Engineering Society*, October 2011.

[AR12] ——. (2012) Head-related impulse response repository. [Online]. Available: http://marl.smusic.nyu.edu/projects/HRIRrepository (date of access: November 2012).

[ATVE12] C. Antweiler, A. Telle, P. Vary, and G. Enzner, "Perfect-sweep NLMS for time-variant acoustic system identification," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 517–520.

[BS09] E. Berdahl and J. Smith. (2009) Impulse response measurement using golay complementary sequences. [Online]. Available: http://cnx.org/content/m15947/latest/ (date of access: March 2013).

[CIP98] CIPIC Interface Laboratory, "Documentation for the UCD HRIR files," Ocotber 1998, available: http://earlab.bu.edu/databases/collections/cipic/documentation/hrir_data_documentation.pdf (date of access: January 2013).

[CIP12] CIPIC (Center for Image Processing and Integrated Computing), University of California. (2012) The CIPIC HRTF database. [Online]. Available: http://interface.cipic.ucdavis.edu/sound/hrtf.html (date of access: January 2013).

[DH93] C. Dunn and M. Hawksford, "Distortion immunity of MLS-derived impulse response measurements," in *Journal of the Audio Engineering Scoiety Vol. 41, No. 5*, may 1993.

[Ear08] EarLab Boston University. (2008) Earlab CIPIC HRTF Search. [Online]. Available: http://earlab.bu.edu/databases/collections/cipic/Default.aspx (date of access: January 2013).

[Far00] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," 2000, dipartimento di Ingegneria Industriale, Università di Parma, available: http://www.nvo.com/winmls/nss-folder/electro1acoustics/Measuring%20impulse%20resp%20and%20distortion%20with%20swept%20sine%201341AES00.pdf (date of access: March 2013).

[FIU04] FIU DSP Lab, Florida International University. (2004) HRTF/Anthropometric measurement database. [Online]. Available: http://dsp.eng.fiu.edu/HRTFDB/main.htm (date of access: December 2012).

[HEA$^+$09] H.Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," in *EURASIP Journal on Advances in Signal Processing, Volume 2009, Article ID 298605, 10 pages*, June 2009.

[Höl12] J. Hölzl, "An initial investigation into HRTF adaption using PCA," 2012, project thesis at Institute for Electronic Music, University for Music and Performing Arts Graz, available: http://iem.kug.ac.at/fileadmin/media/iem/projects/2012/hoelzl.pdf (date of access: March 2013).

[IRC03] IRCAM (Institut de Recherche et Coordination Acoustique/Musique), Paris. (2003) Listen HRTF databse. [Online]. Available: http://recherche.ircam.fr/equipes/salles/listen/index.html (date of access: January 2013).

[Maj10] P. Majdak, "Algorithmen in Akustik und Computermusik," 2010, script for corresponding lecture at Institute for Electronic Music, University for Music and Performing Arts Graz.

[MBL07] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," in *Journal of the Audio Engineering Society, Vol. 55, No. 7/8, 2007 July/August*, April 2007.

[MG11] M. Maazaoui and Y. Grenier, "Romeo-HRTF: A multimicrophone head-related transfer functions database," 2011, available: http://www.tsi.telecom-paristech.fr/aao/en/2011/03/31/romeo-hrtf-a-multimicrophone-head-related-transfer-function-database/ (date of access: October 2012).

[MH10] N. Matsunaga and T. Hirahara, "Fast near-field HRTF measurements using reciprocal method," in *20th International Congress on Acoustics, ICA 2010*, August 2010.

[MIC$^+$13] P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, H. Ziegelwanger, and M. Noisternig, "Sptaially oriented format for acoustics: A data exchange format representing head-related transfer functions," in *134th Convention of the Audio Engineering Society*, May 2013.

[MM01] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," in *Journal of the Audio Engineering Scoiety Vol. 59*, december 2001.

[MM07] R. Martin and K. McAnally, "Interpolation of head-related transfer functions." Air Operations Division, Defence Science and Technology Organisation, Australia, 2007.

[MM11] P. Majdak and M. Mihocic, "ARI HRTF format v2," Ocotber 2011, available: http://www.kfs.oeaw.ac.at/research/experimental_audiology/hrtf/ARI%20HRTF%20format.pdf (date of access: January 2013).

[MPC05] P. Minnaar, J. Plogsties, and F. Christensen, "Directional resolution of head-related transfer functions required in binaural synthesis." Journal of the Audio Engineering Society Vol. 53, No. 10, 2005.

[MSHJ95] H. Møller, M. F. Sorensen, D. Hammershoi, and C. B. Jensen, "Head-related transfer functions of human subjects," in *Journal of the Audio Engineering Society Vol. 43, No. 5*, 1995.

[Psy12] Psychoacoustics and Experimental Audiology group at Acoustics Research Institute, Vienna. (2012) ARI HRTF database. [Online].
Available: http://www.kfs.oeaw.ac.at/index.php?option=com_content&view=article&id=608&catid=158&Itemid=606&lang=en (date of access: November 2012).

[SEA02] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," in *Journal of the Audio Engineering Scoiety Vol. 50*, december 2002.

[TAV10] A. Telle, C. Antweiler, and P. Vary, "Der perfekte Sweep - Ein neues Anregungssignal zur adaptiven Systemidentifikation zeitvarianter akustischer Systeme," in *Proceedings of German Annual Conference on Acoustics (DAGA)*, 2010, pp. 341–342.

[Wes11] W. Weselak, "Akustische Messtechnik," 2011, script for corresponding lecture at Signal Processing and Speech Communication Laboratory, University of Technology, Graz.

[WGRS11] H. Wierstorf, M. Geier, A. Raake, and S. Spors, "A free database of head-related impulse response measurements in the horizontal plane with multiple distances," in *130th Convention of the Audio Engineering Society*, May 2011.

[WHGS12] H. Wierstorf, K. Helwani, M. Geier, and S. Spors. (2012) Impulse response measurements. [Online].
Available: https://dev.qu.tu-berlin.de/projects/measurements/wiki/Impulse_Response_Measurements (date of access: January 2013).

[Zaa10] J. Zaar, "Vermessung von Aussenohruebertragungsfunktionen mit reziproker Methode," 2010, project thesis at Institute for Electronic Music, University for Music and Performing Arts Graz, available: http://iem.kug.ac.at/fileadmin/media/iem/projects/2010/zaar.pdf (date of access: March 2013).

[Zah99] P. Zahorik, "Limitations in using golay codes for head-related transfer function measurement," in *Journal of the Acoustical Society of America, March 2000*, 1999.