Technische Universität Berlin
Audio Communication Group
Institut für Sprache und Kommunikation

An augmented acoustics demonstrator with realtime stereo up-mixing and binaural auralization

In fulfilment of
the requirements for the degree of
Master of Science

December 2015

Supervisor:
Prof. Dr. Stefan Weinzierl
Dr. Alexander Lindau

Author:
Raffael Tönges
[redacted]

Acknowledgement

Vielen lieben Dank an Vivi, Gerri, Mama & Papa, Benni, Berlab, Alexander Lindau, Fabian Brinkmann, Stefan Weinzierl und die Musik im Kopf, die raus möchte.

Eidesstattliche Erklärung

Hiermit versichere ich gegenüber der Fakultät I der Technischen Universität Berlin, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Ausführungen, die anderen veröffentlichten oder nicht veröffentlichten Schriften wörtlich oder sinngemäß entnommen wurden, habe ich kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Fassung noch keiner anderen Prüfungsbehörde vorgelegen.

Berlin, den 03. Dezember 2015

Abstract

This work documents the implementation and evaluation of a demonstrator for realtime stereo up-mixing and binaural auralization. The ability of the demonstrator to mixing in the acoustic surroundings (binaural microphones) is also implemented. For this purpose the author realised an *Apple iOS 8* application, which provides a multi-channel audio interface with two channel types. The media channel supports the reproduction of local audio files (MP3, AAC, etc.) and several filter types, such as the USC 10.2 filter (twelve channel up-mixing). The additional channels add artificial room information to create the impression of a virtual listening space. Furthermore, a headphone compensation filter was implemented, which can provide any impulse response (in the defined form) for filtering. Additionally, two microphones were mounted to a consumer headphone, one on each side of the head (hear-through headset), to send binaural acoustic input to the microphone channel of the application. The evaluation showed, that such a system works with modern *Apple* smartphones. The technical evaluation lead to good results taking into account the given limitations (processor, battery). Due to the size of the utilised headphones the binaural recordings are spectral coloured depending on the position of the microphones. Also unnatural ITDs could be produced, which then influences the precise localisation of surrounding acoustic sources.

Zusammenfassung

Diese Arbeit beschreibt die Umsetzung und Evaluation eines Demonstrators für Echtzeit-Stereo-zu-Mehrkanal-*up-mixing* mit binauraler Auralisation und dem kontrollierten Einmischen von Umgebungsgeräuschen (binaurale Mikrofone). Hierfür wurde eine *Apple iOS 8* Applikation implementiert, welche ein Mehrkanalaudiointerface mit zwei Kanaltypen bereitstellt. Der Dateikanal unterstützt die Wiedergabe von lokalen Audiodateien (MP3, AAC, etc.) und verschiedene Filtertypen, wie z.B. dem USC 10.2 Filter (Zwölf-Kanal-*up-mixing*). Über die zusätzlichen Audiokanäle werden künstlich generierte Raumschallinformationen wiedergegeben, um einen virtuellen Hörraum zu schaffen. Zusätzlich ist ein Kopfhörerkompensationsfilter implementiert, welcher beliebige Impulsantworten als Filterfunktion verwenden kann. Zusätzlich wurde ein handelsüblicher gut dämpfender Kopfhörer mit einem Mikrofon an jeder Kopfseite erweitert (*hear-through headset*), um die Umgebungsgeräusche kontrolliert über den Mikrofonkanal der Applikation einmischen zu können. Die Evaluation ergab, dass ein solches System für moderne Smartphones der Marke *Apple* realisierbar sind. Die technische Evaluation führte zu guten Ergebnissen der Applikation im Rahmen der Limitierungen (Prozessor, Akku). Die binaurale Aufnahme wird durch die Montage an den großen Kopfhörern abhängig von ihrer Position spektral verfärbt. Auch können sich unnatürliche ITDs ergeben, was die präzise Lokalisation von umliegenden Quellen beeinträchtigt.

Contents

List of Figures

List of Tables

Glossary

AH	Artificial head
AR	Augmented reality
BRIR	Binaural room impulse response
BRTF	Binaural room transfer function
BS	Back surround
CPU	Central processing unit
DAR	Direct/ambient ratio
DRR	Direct-to-reverberation energy ratio
ETC	Energy time curve
FABIAN	Fast and automatic binaural impulse response acquisition
FC	Front centre
FFT	Fast Fourier transformation
FL/FR	Front left / front right
GUI	Graphical user interface
HATS	Head and torso simulator
HL/HR	High left / high right
HRIR	Head related impulse response
HRTF	Head related transfer function
HpTF	Headphone transfer functions
ILD	Interaural loudness difference
ITD	Interaural time difference
ITDG	Initial time delay gap
LFEL/LFER	Low frequency effect left / low frequency effect right
LMS	Least mean square
MVBNAP	Multiple-wise vector base non-negative amplitude-panning
PAD	Primary ambient decomposition
PCA	Principal component analysis
RIR	Room impulse response
SAQI	Spatial audio quality inventory

SC	Spectral colouring
SL/SR	Surround left / surround right
VBAP	Vector-based amplitude panning
WL/WR	Wide left / wide right

1. Motivation and scope

When *Sony* released the *Walkman TPS-L2* [1] in 1979, mobile music culture started. In 1984, based on this new progress in technology Hosokawa [4] summarised several aspects of individual mobile music listening, such as controlled singularisation, construction or destruction of the meaning of the perceived surrounding and the isolation from the noisy urban soundscape. As the *Walkman* was an invention to be used in everyday life, the *Hot-Line* button was implemented [5]. Pressing this button, the volume of the audio signal was lowered and superimposed by the signal of a microphone, which was built into the enclosure of the *Walkman*. Thus, it allowed the listener to receive the acoustic surroundings without having to remove the headphones. In later models the *Hot-Line* button was removed and mobile music listening became more and more a synonym for singularisation and acoustic seclusion of listeners.

Nowadays, augmented reality is an area of great interest in research and commerce and hear-through headsets are investigated. These headsets are either acoustically transparent per se or they have microphones mounted on the enclosure to control the degree of transparency. Part of the presented work is the development of a hear-through headset as a demonstrator for acoustic augmentation. A customer headphone with noticeable attenuation was used and two microphones were attached to it, one on either side of the head on the enclosure of the headphone. Thus, in comparison to the *Hot-Line* function of the *Walkman* the recorded sound is a binaural signal, containing important acoustic informations about the perceived sources e.g. localisation cues.

Moving from hifi stereo sound at home to mobile stereo sound was first described as "devolution" by Hosokawa [4], as the audio quality of the early compact cassette players was comparably poor. This poor quality was mainly caused by the simple headphones, the cheep technical parts mounted in the player and finally the poor audio quality of the compact cassettes. A comparably big revolution in mobile music was the development of the MP3 codecs, which produces small audio files and thus allows to store a large number of songs, in conjunction with the commercial availability of port-

[1] The *Sony Walkman TPS-L2* was the first commercially available portable cassette recorder and player sold from 1st July 1979.

able digital audio players in the late 1990s and early 2000s. The quality of the portable audio players and the headphones has improved since the *Walkman*, but the quality of the available MP3s was initially quite poor. Throughout the last fifteen years a lot of effort was put into the development of high quality audio codecs and the MP3 has constantly improved as well. In Pras et al. [6] no significant difference between CD quality and MP3 files with $256 kbit/s$ or $320 kbit/s$ could be found. In 2003 more smartphones than feature phones were sold globally [7]. Smartphones usually offer the ability of audio reproduction. According to Lepa [8] in 2012 33% of the German population above fourteen listened to music via smartphone.

One big difference between the hifi stereo system and portable music remains ever since, the listening space. When loudspeakers are used for audio playback, the listening space has a huge influence on the sound. The sound has to travel through space, some frequencies are amplified or attenuated, the room has a specific reverberation time and the loudspeakers are usually fixed to one position. Thus, if a listener is in the sweet-spot, a virtual stage emerges between the two loudspeakers and the room additionally colours the spectrum of the sound. When headphones are used for audio playback, the listening space does not exist. The sound does not travel through space and the room surroundings. The listener himself has no influence on the perceived sound.

Binaural synthesis is the technical approach to create a virtual listening space. The frequency modulation of the listeners body, her head and pinna on a sound propagated from a certain position relative to the listeners position, can be measured and used to filter audio content. Thus, the impression is evoked that sound heard through the headphones was coming from this very position. Further, acoustical characteristics of a real or artificial room can be simulated and added to the audio signal, to make the listener believe, the sound she is listening to, is replayed in this very room. Binaural synthesis is a computing intensive technique, as it can involve a huge number of calculations, depending on the plausibility and authenticity intended.

It is assumed that modern smartphones could compute an up-mixing algorithm from stereo to USC 10.2 surround, where the additional channels are used for additional signals, containing early reflections and diffuse sound, which are usually present in a real listening space. As main part of the presented work, an application was implemented that involved realtime stereo up-mixing and binaural auralization. Further, as the hear-through headset is constructed, the application has the ability of mixing the up-mixed audio with the sound coming from the microphones. Thus, the application will allow to “integrate” the audio signal into the real acoustic surrounding. This way it is hoped to provide a means to both preserve the meaning of the surrounding real environment

and prevent from singularisation and acoustic isolation.

This work is preliminary for an externally founded follow-up research project. The focus lies on the investigation of feasibility of a comprehensive mobile realtime stereo up-mixing application and the integration of realtime binaural recordings over microphones.

This work is structured as follows: In the second chapter the state of research on spatial hearing, binaural synthesis, stereo up-mixing and augmented reality headphones is reviewed. In the third chapter informations on commercially available mobile applications regarding binaural technology and augmented reality are provided. Furthermore, a short hardware review on augmented reality headphones is included. In the fourth chapter the hardware and software concept of the demonstrator is described. In the fifth chapter the encountered problems during the implementation phase are addressed and the developed solutions are revealed. The sixth chapter contains the technical and informal perceptual evaluation of the demonstrator. In the seventh chapter the conclusion of the work is drawn and the future perspectives are provided.

In the entire work the female gender is used as representative for all genders.

2. State of research

Although many multi-channel formats for audio files are available and are gaining more importance in scientific and commercial applications, the stereo formats (WAV, AIFF, FLAC, MP3, AAC, etc.) have remained the standards for audio reproduction until now. If reproduced via loudspeakers, a virtual stage will be perceived between the two speakers, if the listener is located in the sweet spot. The acoustic characteristics of the room lead to an integration of the listener in the acoustic scene, while the location of the musicians remains fixed to the virtual stage. When audio is reproduced via headphones, the sound does not travel through space, but is "injected" directly into the listeners ears. Thus the listening space and the filter characteristics of the listeners morphology is lost. Without any further processing the virtual stage is therefore placed between the listeners ears - in her head.

This phenomenon can be avoided and a virtual listening space with any desired acoustic characteristics can be modelled when using binaural synthesis (Sec. 2.2). To understand binaural synthesis first binaural hearing has to be examined (Sec. 2.1). In addition, stereo signals can be up-mixed (Sec. 2.3) and binaurally auralised to gain a more natural result, containing reflections from virtually added boundary surfaces. Thus, a perceived integration into a surrounding acoustic environment can be achieved.

A major interest of actual audio research is augmented audio. For augmented audio binaural synthesis is essential as sounds can be localised outside the head and arranged spatially somewhere around the listener. Further, using headphones the applicability is not limited to a certain location. To ascertain an adequate augmented audio scenario, the headphones must meet a number of criteria. The quality of the headphones (measured by the frequency response) must be high, in order to produce a plausible virtual acoustic reality. As an augmented audio scenario is usually a superposition of the real acoustic scenery and an artificially generated environment, the headphones should be able to "leak" the acoustic surroundings (Sec. 2.4).

2.1. Spatial hearing

Humans possess two ears. A listener, who listens to a sound source, receives two distinct signals; one at each ear. These two signals can differ in loudness (interaural loudness difference, ILD), time of arrival (interaural time difference , ITD) [1], phase (interaural phase difference, IPD) [2] and spectral colouration (SC), depending on the location of the source relative to the position of the listener and her head orientation. Hearing with two ears and the ability to process and relate these two signals in the auditory system is called binaural hearing.

The location relative to a listener can be defined by the three parameters azimuth angle Φ (horizontal plane), elevation angle Θ (median plane) and distance r. The centre of the listeners interaural axis defines the centre of the coordinate system as shown in Fig. 2.1. The position $p(\Phi,\Theta,r) = p(0°,0°,1)$ refers to the location right in front of the listener in the distance of $1m$.

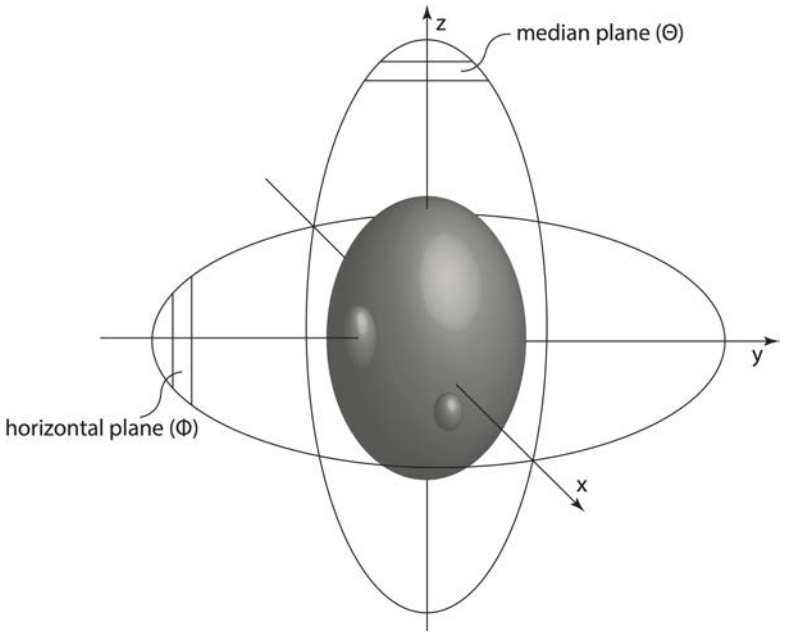


Figure 2.1.: Coordinate system for head related locations.

In a free-field situation, meaning that the sound emitted by a source is not reflected at any boundary surface, the sound waves of a source shifted towards the right side

[1] The ITD can be divided in two subclasses, the fine-structure ITD and the envelop ITD [9, p. 388].

[2] In abrupt signals are considered the ITD can be measured as the difference between the times of arrival of the signal at both ears. In periodic signals the phase shift between the two signals (IPD) defines the ITD.

of the listeners head have an higher amplitude and arrive earlier at the right ear than at the left ear. Further, because the head produces an acoustic shadow at the left ear, the head acts like a low-pass filter. Thus, the resulting left-ear signal is received more bump. Additionally, the torso, the head and especially the pinna reflect and diffract the sound wave depending on the angle of incidence [10, p. 86]. To measure these acoustic differences between the two ear signals, a pair of head related impulse responses (HRIR), or in the frequency domain the head related transfer functions (HRTF), can be recorded and compared for a specific source position. These HRIRs can be recorded, e.g. as a non-individual set for various positions using an artificial head (AH) [3] or a more advanced head and torso simulator (HATS) [4]. AHs and HATSs are more or less detailed replica of a human head (and upper body) with microphones placed in their ears. Further, individual HRIRs can be measured by placing tiny microphones into the ear canal of a subject and record the arriving sound of a wideband audio source. Usually *sweeps* [5] are used as excitation signal, because they are well controllable regarding their frequency properties. Often HRIR sets are recorded, containing as many source positions as possible, on a sphere around the subject in a constant distance ([12], [13], [14]).

The indicators ILD and ITD are binaural cues for spatial hearing. Further monaural cues exist for spatial hearing, such as loudness/intensity, initial time delay gap (ITDG, Sec. 2.1.1) and direct-to-reverberant energy ratio (DRR) (Sec. 2.1.2). SC can be a monaural or interaural cue for spatial hearing, analog to perceived monaural loudness and the ILD. The interaural relation of both ear signals regarding their spectrum is specified as interaural cross correlation (IACC).

When reconstructing or emulating acoustic events (Sec. 2.2) in a virtual environment, the quality of the acoustic image can be investigated regarding distinct features. A set of spatial audio evaluation criteria for virtual acoustic environments is defined by Lindau et al. [15] - the spatial audio quality inventory (SAQI). In Tab. 2.1 the identified criteria are listed. An virtual acoustic environment is ideally not distinguishable from a real acoustic environment and thus, underlays the same criteria for localisation, temporal behaviour, etc. The relevant categories of this set regarding the evaluation of the implemented solution are explained further in this section and are used to explain the acoustic features as they are.

[3] Examples for AHs are the *Neumann KU 100*, the *Sennheiser MKE 2002* or the simpler *AKG D 99 C "Harry"*.

[4] Examples for HATSs are the *FABIAN* from *Technische Universität Berlin* [11] and the *Brüel & Kjær 4128D*.

Category	Auditory quality
Timbre	Tone colour bright -dark, High-/Mid-/Low-frequency tone colour, Sharpness, Roughness, Comp filter coloration, Metallic tone colour
Tonalness	Tonalness, Pitch, Doppler effekt
Geometry	Horizontal/Vertical direction, Front-back position, Distance, Depth, Width, Height, Externalisation, Localisability, Spatial disintegration
Room	Reverberation level, Reverberation time, Envelopment of reverberation
Time behaviour	Pre-/Post-echoes, Temporal disintegration, Crispness, Speed, Sequence of events, Responsiveness
Dynamics	Loudness, Dynamic range, Dynamic compression effects
Artefacts	Pitch/Impulsive/Noise-like artefact, Alien source, Ghost source, Distortion, Tactile vibration
General	(Overall) Difference, Clarity, Speech intelligibility, Naturalness, Presence, Degree-of-Liking, Other

Table 2.1.: The complete set of categories and auditory quality features of SAQI. The set is a collection of features used to investigate the quality of virtual acoustic environments.

2.1.1. Geometry

ILDs, ITDs and the SC are the main cues for the acoustic localisation process. However, ILD and ITD are the prime indicators. For sounds below $1600Hz$ ($\lambda = 22,9cm$) the length of a sound wave is greater than the diameter of an average head. That causes a significant delay in arrival of the sound on either of the both ear canals, while the ILD is relatively small. Therefore, the ITD results in being the most important cue for the localisation of sounds below $1600Hz$. For sounds above $1600Hz$ the delay is less significant, but the ILD is crucial. Thus, the ILD is recognised as the main indicator for localisation of sound above $1600Hz$ [16].

The structure of a human torso and head is rather symmetric to the median plane. Hence, areas of equal ILDs and ITDs can be found on either side of the head. These areas are called *cones of confusion* (Fig. 2.2). Assuming a static head position, a listener can not determine the location of a sound source in these areas only through the two main indicators ILD and ITD. Hence, in addition the SC is used for localisation. Sounds above $2000Hz$ of a source originated behind a listener e.g. reach the ear canals with $2-3dB$ less amplitude at both ears than of a source originated in front of the listener.

Thus, the resulting spectral difference superinduces a cue for front/back discrimination. Further, the elevation of a sound source is identified through the resulting spectral variations caused by angle dependent frequency accentuations and attenuations caused by the shape of the pinna [16].

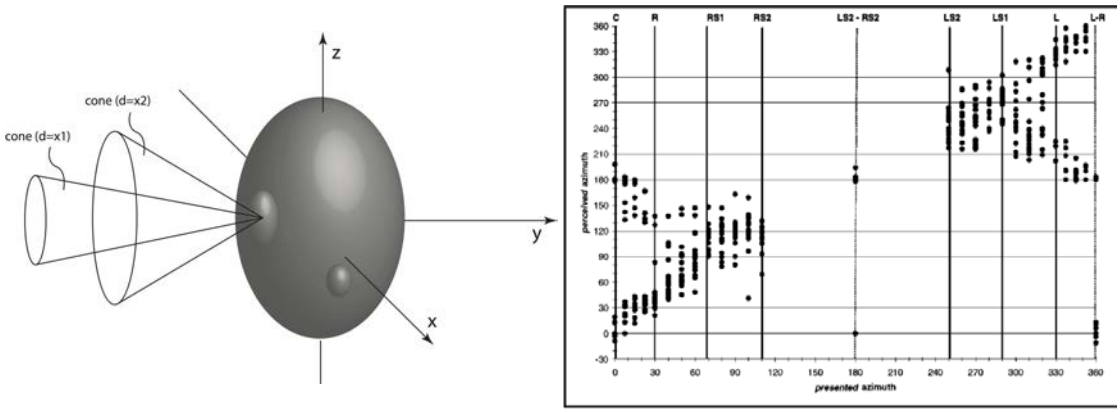


Figure 2.2.: The areas of equal ITD are called cones of confusion. In these areas ITDs cannot be consulted for localising sources. The left figure shows two cones of confusion. The right scatter plot [1, p. 676] shows front/back confusion during localisation tests using static binaural synthesis. In this case all source positions in the frontal hemisphere are located behind the head.

As different cues serve as indicators for the location of a sound source depending on the pitch of the sound and the plane (horizontal/median) in which the source is shifted, the localisation accuracy differs. Since a resolution of $\pm 0.75^\circ$ in the horizontal plane and $\pm 9^\circ$ can be found for a listener [12].

To improve the acoustic localisation, especially to differentiate between locations in the *cones of confusion* or in a reverberant room, minor head movements are performed to gain small variations in ILDs and ITDs ([17, pp. 22-25], [18]). Due to the head movements the source travels more towards one ear, resulting in a louder signal that arrives earlier at this very ear. These minor variations are sufficient to determine the location of a sound source with a higher precision than by interpreting the SC only.

A significant indicator for distance is the loudness of a signal [19]. The further away a source is located, the quieter its sound is heard. The distance can also be analysed by the DRR if boundaries are present to reflect the sound waves ([20], [21]). A longer distance results in a higher energy of the reverberation over the direct sound. During outdoors propagation the spectrum of a sound source changes depending on distance, humidity and temperature. This is due to the fact that the air functions as a low-pass

filter with a varying steepness and overall absorption controlled by the forementioned three parameters for distances above 15 m ([21], [22], [23]).

Without a visual stimulus connected to the acoustic event however, distances are usually underestimated [24].

The ability to locate a virtual (and real) source is one of the main approaches in augmented acoustics solutions.

2.1.2. Room

If a sound source is surrounded by boundaries the sound waves reflect on the surfaces of these boundaries. Depending on the surface properties the direction of the sound is changed or spread and the intensity of the reflecting sound wave is attenuated compared to the entering sound wave. Thus, when a listener receives the sound of this source, she receives a mixture of the direct sound and the reflected sound. The reflections can be distinguished in early reflections and late reflections or reverberation (Fig. 2.3). As explained in Sec. 2.1.1 the DRR is used as cue for distance estimation. But it is also a cue for further environmental features (structure, size, volume, nature of boundaries, etc.). A long reverberation time, e.g. is a cue for a big room and/or highly reflecting boundary surfaces. Furthermore, the DRR is an aesthetic feature that influences the evaluation of a received source. Various evaluation criteria are collected for room acoustics that are based on DRR or solely the reverberation level and time [1, pp. 185-205]. The ITDG can be taken as a cue for room impression too. The larger the ITDG, the larger the listening space [25].

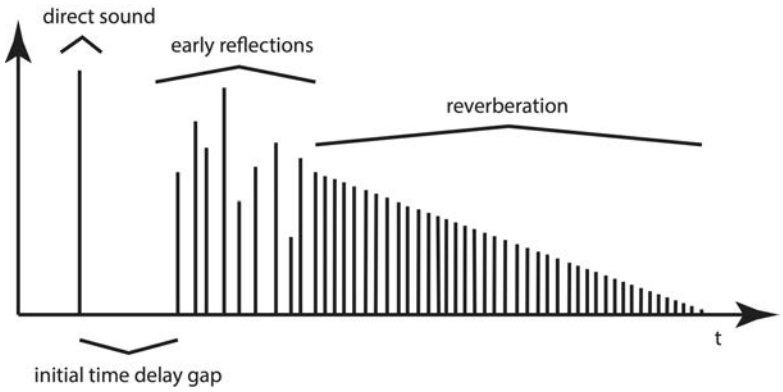


Figure 2.3.: Exemplary reflectogram of a room.

A good room impression is important for the plausibility and a positive evaluation of the described demonstrator.

2.1.3. Timbre

The timbre of a sound is mainly defined by the spectrum and envelope of energy over time. In this work the timbre is mainly influenced by the choice of HRTFs and headphone compensation filters.

2.1.4. Artefacts

Notable artefacts in sound, such as unnatural noise, pitch or distortion, can diminish the evaluation of the sound. These artefacts take place during reproduction, e.g. through discontinuities, transmission errors (analog/digital), high amplitudes, etc.

In the presented work, artefacts did mostly occur during the rendering process (binaural synthesis, headphone compensation).

2.2. Binaural synthesis

If headphones are used, the sound does not travel through a large space. Depending on the physical headphone design (extra-aural, circum-aural, supra-aural, in-ear, etc.) the sound passes the pinna or is even directly injected into the ear canal. Further, the sound moves with the head (distance, rotation). This unnatural behaviour of a static spatial relation between the listener and sound sources, the lack of room reflections and the missing typical reflections/absorptions of torso, head and pinna lead to the perception that the sounds origin in the head (in-head localisation) [26].

One approach to avoid in-head localisation is binaural synthesis or binaural auralization. When a mono sound is binaural auralized, it is convolved with the pair of HRIRs that belong to the desired source direction. Thus, the sound is modulated, as if it was propagated from this very direction in a free-field situation. As the sound is convolved

with the pair of HRIRs, the result is a binaural signal, which contains the two signals that arrive at the two ears. These signals are potentially different in loudness, time of arrival and SC.

For a given stereo signal, binaural auralization leads to the same effect than reproducing the same sound over loudspeakers that are setup for stereo reproduction, when the required HRIRs are applied. In this case both channels are convolved with the appropriate pair of HRIRs of the corresponding positions of the stereo loudspeakers. As an example: A listening situation with two loudspeakers (SL, SR) at the positions $\pm 30^{\circ}$ shall be simulated. To gain the left channel of the binaural signal, the left channel of the stereo signal is convolved with the HRIR of SL for the left ear and added to the left stereo channel convolved with the HRIR of SR for the left ear and vice versa (Fig. 2.4).

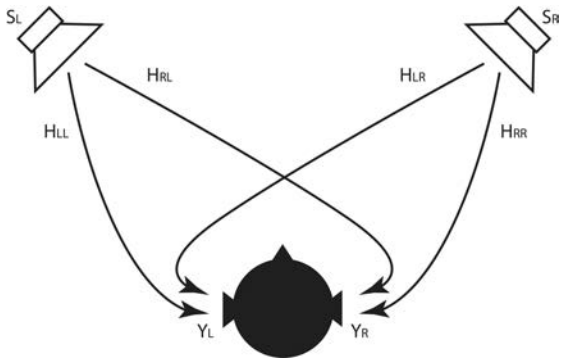


Figure 2.4.: Crosstalk of multiple sound sources such as stereo loudspeakers.

Non-individual HRIRs as well as individual HRIRs can be recorded. For the filtering process both HRIRs can be used, with varying results. HRIRs differ between subjects depending on the morphology of the listeners' torso, head and pinna. If HRIRs are used that do not match the listers experience, the results will not sound natural and often lead to a poor impression of externalisation and the diminished ability to localise a source in the virtual listening space. The experiments presented by Møller et al. [27] and Møller et al. [28] expose the lack of localisation, especially in the median plane and the front/back confusion, when using non-individual HRIRs. Further, as mentioned by Lindau et al. [15], timbre is one of the factors used for evaluating the quality of spatial audio. The choice of the HRIR set used for binaural synthesis has a major influence on the timbre of the received sound and thus lead to inter-individual differences in perception of the simulated scene.

According to the possible improvements achieved when using individual HRIRs, various techniques were investigated to measure individual HRIRs. In theory, individual HRIRs are measured by reproducing a suitable signal from every direction on a sphere around a listener in a free-field environment and recording the arriving signals at each ear canal. To save time often fixed positions on the sphere are used, e.g. every 5° [29]. Afterwards, the resolution is increased by interpolation. A comparatively timesaving modern technique is presented by Fuß [13] and Fallahi [14] to achieve a HRIR set of any desired resolution. A subject is seated on a rotatable chair with the centre of the subjects interaural axis in the centre of a big vertically upstanding metal ring, with loudspeakers attached to cover every segment of 10 degrees distance. The signals reproduced are exponential sweeps according to the multiple exponential sweep method (MESM) as presented by Majdak et al. [30] and optimised by Dietrich et al. [31]. Tiny microphones are placed in the ear canals of the subject with which the arriving signals are recorded. The records are further processed and interpolated to obtain a computer readable HRIR database with any desired resolution. The microphones used for measuring the HRIRs are encapsulated in silicone and thus blocking the ear canal [12]. Therefore, additional compensations are necessary [32]. Further, when entering the ear canal a signal will not be subject to further modulations, so all information about the listener-source relation in distance and angle are already present when measured at this position [12].

2.2.1. Binaural room impulse response

HRIRs contain the direction information of a source - receiver relation in a free-field situation. In natural listening situations a sound is often not only modulated by the listeners body, head and pinna, but also by the surrounding structures. Rooms attenuate or amplify frequencies depending on their structure, geometry, volume and other conditions such as flow of air, temperature, etc. [33, pp. 20-27].

The received sound in a room, can also differ depending on the position of source and listener (if not in the diffuse-field). When sound waves impinge on a surface, the sound wave reflects. Depending on the characteristics of this surface the degree of reflection respectively absorption and the angle of reflection vary. Further, these characteristics are frequency-depended. The simplest example for a perceivable location-dependent sound field is a stationary wave of an adequate low frequency in a rectangular room. When walking along the sound wave, the peaks and troughs of the wave can be spotted

by the higher and lower amplitude at different locations in the room [33, pp. 38-42].

Further, in a reflecting room the sound energy retains longer in this limited volume than it would in a free-field or heavily absorbing room. Thus, the sound of a source can still be heard after the source stopped transmitting (reverberation). The reverberation time, usually denoted as T_{60}, describes the time in which the sound pressure level descends by 60 dB [33, p. 239]. Another characteristic of a room is the quality (density, intensity) of early reflections. Early reflections influence the perception of tone (side walls, ceiling, ground, front wall, rear wall: comb filter effect) and volume (mainly side-walls: envelopment) of a sound source.

The influence of a room on the sound of a source at one position and received at another position - the room impulse response (RIR) - can be measured or simulated. When measuring the RIR usually a omnidirectional source (e.g. dodecahedron loudspeaker) is placed at a few representative positions (e.g. on stage) and a microphone (or HATS measuring binaural RIRs) is placed at an adequate position in the room, e.g. in the sweet spot of a concert hall. Further, an acoustic simulation of a room can be obtained with computer programs such as *CATT Acoustic* [34], *EASE* [35] or *Raven* [36], where the deterministic image source model [37] or the stochastic ray-tracing algorithm ([38], [39]) is used to determine the RIR at a particular position based on a model of the room and the acoustic characteristics of the boundary surfaces. The image source model is an approach where reflections are modelled by repeatedly mirroring the room on each of its boundaries and preserving the relative position of the sound source inside the room. The result is a mesh of mirrored rooms, each containing an image of the sound source. Each of these sources propagate its sound straight towards the receiver in the original room, while its sound is attenuated according to the characteristics of the walls it has to pace. Hence, the reflections of the original source are simulated. The result is an estimation of the real sound field. Ray-tracing on the other hand, describes a model in which a source radiates a defined amount of particles. The receiver is defined with a certain volume. When a particle reaches the receiver at one point, its path is traced to estimate the intensity after the travel. During the travel through the modelled room, the particles are reflected on the surfaces and again modulated according to the characteristics of the surfaces.

If such a sound field in a room shall be simulated, the HRIRs are not sufficient, as no information about the room are contained. The fusion of HRIRs and RIR are called binaural room impulse responses (BRIRs) or binaural room transfer functions (BRTFs)

in the frequency domain. BRIRs contain the information on how a source at one position of a room would be received at another position of the same room, considering early reflections and/or late reflections. In addition to measuring or simulating RIRs of existing rooms, artificial BRIRs can be constructed.

As investigated by Sandvad and Hammershøi [40] relatively short HRIRs lead to acceptable filtering results. A length of 1.5s at a sampling frequency of 48 kHz is sufficient to gain a detection probability of less than 0.08, that subjects were able to distinguish between a reference and the filtered signal. The HRIR refers to a record of a reflection-free impulse, thus a relatively short filter length is plausible. A room, however, is among others characterised by its reverberation time. Thus, a BRIR filter has to be at least as long as the reverberation time of this very room, if the late reflections are to be considered. Schroeder et al. [41] defines BRIRs to consist of typically between 20,000 and 200,000 filter coefficients (0.42 - 4.2 s, 48 kHz), so that a realistic simulation can be achieved. A BRIR containing 200,000 filter coefficients would lead to a latency of 4.2 s (48 kHz). In real-time systems very short latencies are required. Even more when event-bound processing is applied (e.g. dynamic binaural systems, Sec. 2.2.2). Thus, sequential block-wise filtering can be used for long BRIRs to reduce the resulting latency. Efficient algorithms for sequential filtering are presented by Gardner [42], García [43], Hurchalla [44] and Wefers and Vorländer [45].

As BRIRs can improve the listening experience (externalisation, envelopment, distance perception, [46], [24]), in this work BRIRs respectively BRTFs will be used for binaural synthesis.

2.2.2. Dynamic binaural synthesis

Binaural synthesis should lead to perceiving sound outside the head. Further, a precise auditory localisation in the virtual listening space is intended, when using headphones. Binaural synthesis can be applied by using HRIRs/HRTFs and BRIRs/BRTFs. The results gained by using HRTFs for static binaural synthesis is known to be poor ([47], [48], [49]) as sound reproduction in an anechoic environment is simulated. Using BRTFs lead to more natural listening experience due to the addition of reflections. However, both procedures still result insufficient for a reliable front/back discrimination and externalisation. One reason is found to be the static spatial relation between sound source and listener, when processing head movements, which leads to an unnatural hearing

situation. When the sound moves with the head, the brain relates the sound received with the headphone, which is sitting on the head, and therefore the brain locates the sound in the head.

An approach to resolve this problem is dynamic binaural synthesis. The head movements of the listener are tracked using motion detection sensors (head tracker). Thus, by continuously applying the appropriate BRTFs according to the actual head orientation, a sound source can be fixed to one location in the virtual listening space. The spatial resolution of a BRTF set differs according to limitations given by storage/memory, processing capabilities during measurement, measuring instruments, etc. To gain a higher resolution and to avoid discontinuity artefacts, interpolation takes place to handle the transitions between discrete BRTFs.

As mentioned before minor head movements are often processed to improve ones localisation ability. This is found to be a reason why localisation results can be improved with dynamic over static binaural synthesis ([50], [51]).

2.2.3. Headphone compensation

For binaural synthesis an ideal headphone would provide a perfectly linear frequency response so that the headphone itself does not further modulate the reproduced signal in any way. In practice, frequency responses of headphones are not entirely linear. That is accounted for by preferences of companies, the quality of the technical components and other reasons. As an example: Depending on the design philosophy of the company, headphones are often free-field or diffuse-field equalised [32, pp. 197-198], which means that a “linear” frequency response is achieved in a free-field (sound coming from a source in a certain distance right in front of the listener) or a diffuse-field (the sound coming from every direction with the same energy) situation.

To reduce the SC introduced by the headphone, compensation filters can be applied to the system. This can be done by compensating with headphone transfer functions (HpTF) or to some extent with diffuse-field compensation.

Headphone compensation with HpTFs

The frequency response of a headphone can be compensated with the inverse HpTF for the headphone [32]. The compensated frequency response is defined as:

$$H_c(\omega) = H_{HP}(\omega) \cdot H_f(\omega) \tag{2.1}$$

Where $H_c(\omega)$ is the compensated frequency response, $H_{HP}(\omega)$ is the frequency response of the headphone and $H_f(\omega)$ is the frequency response of the filter. In many cases a naive inverse frequency response filtering will impair the frequency response, due to steep notches and further variations in the frequency response depending on the position of the headphone on a listeners head.

Due to the physical design of the headphone (extra-aural, circum-aural, supra-aural, in-ear, etc.) small differences in positioning the headphone on the listeners head have a rather great impact. As shown by Paquier and Koehl [52], the variability in positioning the headphone on the listeners head, leads to audible differences throughout all tested headphones and excerpts. Tested were circum-aural and supra-aural models. Common differences were drifts of the notches and peaks ($\pm 20dB$) in higher frequencies after repositioning the headphone, as a result of pinna resonances and absorptions. The critical frequency is found by Møller et al. [53] to be 7 kHz, above which the variation in pressure division ratio alters too much to find a reasonable compensation curve to flatten the frequency response entirely. Algazi et al. [54] even found the critical frequency to be 6 kHz, from which naive inverse frequency response filtering should be avoided because of the big variances. Further, below 200 Hz SC can be found due to the leakage effect accounted for by gaps between headphone cushion and head surface [55]. These two effects can be seen at Fig. 2.5.

To estimate a reasonable compensation filter, headphone transfer functions can be measured and their inverse can be approximated using a (frequency-dependent) weighted minimum-phase filter ([56], [12]). To avoid extreme amplification in the regions that underlay the variance in positioning of the headphones ($f < 200Hz$ and $f > 6kHz$), the low and high frequencies are compensated with less amplitude or even left out. This is achieved by utilising a regulation function (e.g. low-shelf and high-shelf) when creating the filter to define the degree of compensation per frequency. This way strongly

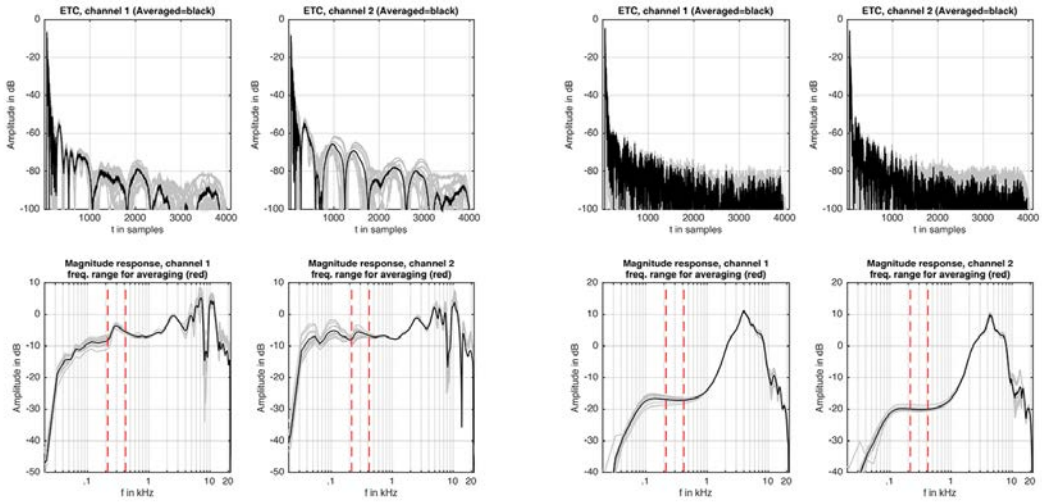


Figure 2.5.: The energy time curves (ETC) and frequency responses of the ten measurements (gray) and mean ETC and frequency response (black) of *Beyerdynamic DT 770 Pro* (left) and *Apple iP6 ear-plug* (right) are displayed.

amplified frequency peaks through extreme compensation can be avoided, which could result in audible ringing artefacts [56]. An example of a compensation filter with applied regulation function can be seen at Fig. 2.6.

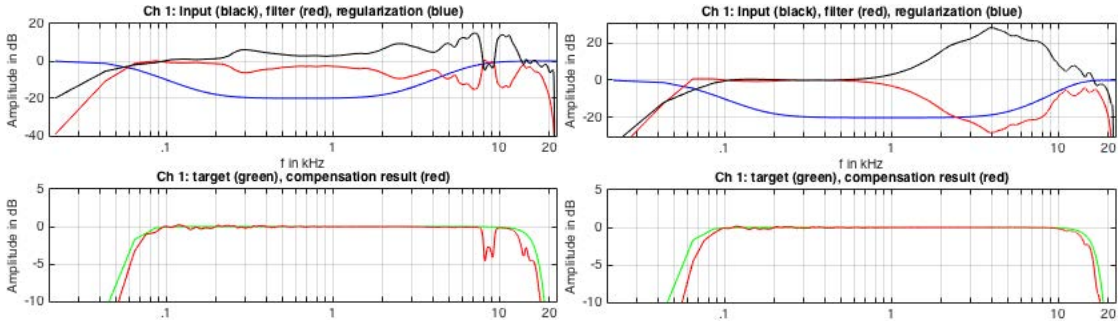


Figure 2.6.: In the two upper plots the mean frequency response (black), the inversion filter (red) and the regulation curve (blue) are shown. In the two lower plots the compensation result (red) is shown according to the desired frequency response (green). All figures display the left channel of *Beyerdynamic DT 770 Pro* (left) and *Apple iP6 ear-plug* (right)

The variances in the HpTF does not depend on the position of the headphones on a listeners head only but also on the structure/topography of the listeners head (leakage effect, low frequencies) and her pinna shape (resonances and absorptions, high frequencies). To compensate the frequency response of the headphone considering the inter-

individual differences, individual HpTF can be measured. Therefore tiny microphones as proposed by Brinkmann [12] are inserted into a subjects ear canals and HpTF measurements are repeatedly processed. During the measurements the headphone is repositioned to gain information about the possible variations due to headphone placement on the listeners head.

Brinkmann [56] shows that individual headphone compensation lead to a better linearisation of the frequency response. In a follow-up study it was found that audio was better evaluated by listeners, when the signals were compensated with the HpTFs of the "recording head" [57]. Thus, it is supposed that for individual recordings a test would yield the same result. However, empirical results explaining this behaviour could still not be found.

The need of individual HpTF to improve localisation is questionable as the differences will only be in SC. Location cues are found in changes in SC, of which the tendencies still might be identifiable, even if the frequency response is different than the one of the real sound field, because the SC of the headphone is constant. The bigger difference might be recognised in discriminating between a real and a simulated sound field, when no compensation filter is applied, simply because of the SC. But it should be mentioned that a listener can anyhow only discriminate between these two sound fields, if they are directly comparable (both sources sound at the same time or right after each other, [58]) or the source is perfectly known, like the voice of a familiar person. In any other situation the coloured sound field could still appear plausible or even natural.

Diffuse-field compensated HRTFs

As headphones often have an even diffuse-field sensitivity level, the usage of diffuse-field compensated HRTFs would lead to a good result for many types of headphones [59]. The remaining weakness of using compensated BRIRs rather than compensate for the actual frequency response of the applied headphone is the headphone specific SC.

If headphones are free-field equalised a compensation will not be necessary if the headphone specific SC is accepted, as HRTFs are recorded in anechoic rooms.

2.3. Stereo up-mixing

In the signal chain of audio playback two acoustically relevant spaces exist. The first space is the recording space (*there-and-than*), which appears often as a mixture of the real studio room acoustics and the eventually applied virtual room acoustics through effects (reverb, etc.). The second acoustically relevant space is the listening space (*here-and-now*). During loudspeaker playback these two acoustic spaces merge to one perceived room impression. As mentioned before the listening space is missing, when listening to audio via headphones without any further processing. Using BRTF filtering, a virtual listening space can be applied.

If HRTFs are used for filtering a stereo signal, the impression of speakers located in a free-field can be achieved. As no reflections from side-walls, the ceiling or rear-wall reach the listeners' ears, the result seams still unnatural [60]. To enhance the listening experience, stereo can be up-mixed to a virtual multi-channel signal matching any surround setup (5.1, 10.2, 20.2, etc.). The surround channels can than be fed with a modulated version of this signal to generate a certain degree of room impression.

To prepare a signal for up-mixing, first it has to be analysed to identify the different components of the signal. Which components to extract, depend on the later applied mixing of the scene. Holman [61] discusses two main approaches for audio mixing. The *direct/ambient* approach gives the listener the impression of being in the audience, e.g. in the sweet-spot of the virtual listening space. Therefore the direct sound component is distributed to the front channels to preserve the stereo image. The additional channels are fed with weighted de-correlated ambient components of the signal to achieve the desired room impression. Further, a high degree of externalisation and the impression of envelopment is achieved by the de-correlation of the distinct signals ([62], [24]).

The other main approach of multi-channel mixing is the *in-the-band* approach, where the identified direct components, representing the distinct sound sources, such as the instruments of a band, are placed on any speaker/speakers around the listener. Thus, the impression arises, that the listener herself is part of the band. In this approach the ambient component is panned to all surrounding speakers to enhance the impression of a natural listening space.

Both approaches have their use cases depending on the desired design. In this work the

direct/ambient approach is used and therefore further examined. Within the performed up-mixing algorithm a virtual USC 10.2 surround setup [6] is generated. The setup is shown in Fig. 2.7. This setup is scalable to other common speaker setups such as surround 5.1 and stereo, as it contains the same channels and extends them. The left respective right channels are the front channel (FL/FR), high channel (HL/HR), wide channel (WL/WR), low frequency effect channel (LFEL/LFER) and surround channel (SL/SR). The unique occurring channels are the front centre (FC) and the back surround (BS) channel.

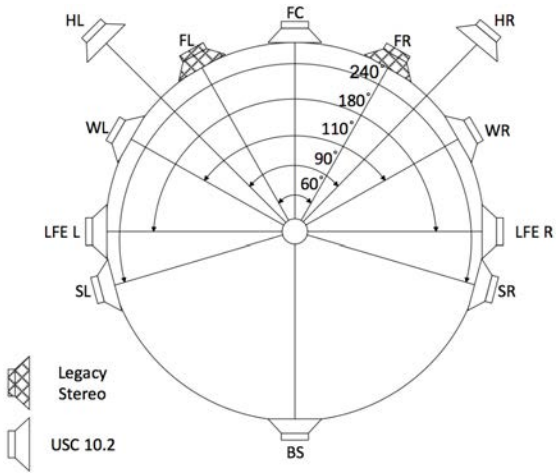


Figure 2.7.: USC 10.2 configuration [2] (naming of FR and FL corrected by author).

The signal is first analysed and than distributed to the virtual channels. Thereafter, the signals are re-synthesised to create the binaural signal. The particular steps of the implemented up-mixing algorithm with applied efficiency modifications proposed by Lee et al. [2] are visualised in Fig. 2.8 and further explained in the following sections Sec. 2.3.1 and Sec. 2.3.2.

All calculations are computed in the frequency domain to profit from the time saving of fast Fourier transformation (FFT) [7] in combination with repeated multiplications and additions instead of multiple convolutions in the time domain.

[6] UCS 10.2 refers to the surround standard defined by the University of Southern California (USC) and *THX Ltd.* [63]

[7] Further informations about FFT on signals can be found e.g. in [64] throughout chapter 8 and 9.

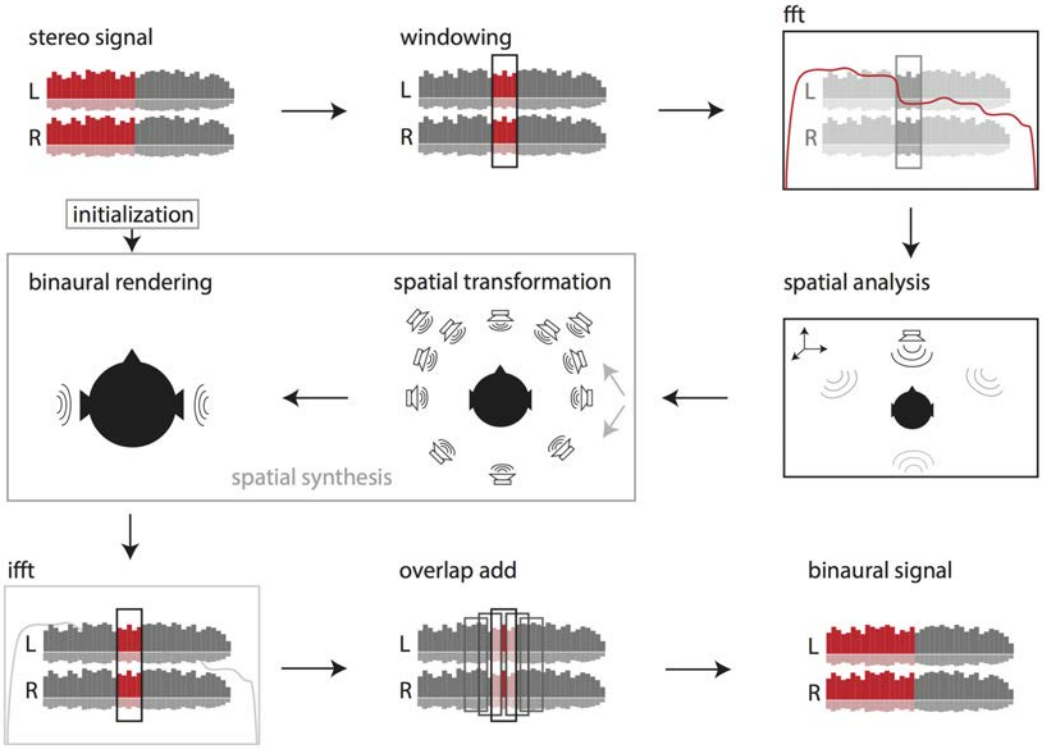


Figure 2.8.: The used algorithm is presented. Audio is represented as stereo audio waveform. The audio file is windowed and FFT is used to convert from time to frequency domain. Than a PCA based spatial analysis is processed. The principal components are placed on front speakers and enhanced by artificial room information placed on the surround channels. The up-mixed signal is binaural synthesised. Later it is transformed to time domain via IFFT and the audio waveform is created via overlap-add method.

2.3.1. Spatial analysis

The first step of stereo to multi-channel up-mixing is to analyse the *there-and-than* scene of the audio material according to the direct and ambient component. Therefore a primary ambient decomposition (PAD) is processed on the two channels of the stereo signal and later the location of the direct component in the stereo image is estimated in the form of panning gains. The necessary information are all acquired through inter-channel relationships, as described in Breebaart and Schuijers [65] and Faller and Breebaart [60]. These inter-channel relationships are analysed and further converted to interaural relationships for the binaural signal, which is synthesised later.

Primary-ambient decomposition

For PAD the vector-based spatial analysis is processed ([66], [67]). It relies on the model of a signal consisting of a direct and an ambient component, as shown in:

$$X_i = g_i S[k,l] + A_i[k,l] \tag{2.2}$$

k and l are the frequency and time frame indices. $g_i, i = L,R$ denotes the left and right panning gain. $S[k,l]$ is the direct component and $A_i[k,l]$ present the left and right ambient components. To carry out these calculations it is assumed that the two channels are uncorrelated and that the ambient components A_L and A_R in each channel have equal levels, so that $E\{S[k,l] \cdot A_i[k,l]\} = 0$ and $E\{A_L[k,l] \cdot A_R[k,l]\} = 0$. These assumptions can be made for typical stereo recordings ([68], [69]). Further the panning gains g_L and g_R are normalised to meet the assumption $g_L^2 + g_R^2 = 1$.

This simple equation for a signal containing the direct and the two ambient components can further be rewritten in matrix notation.

$$\begin{bmatrix} \hat{S} \\ \hat{A_L} \\ \hat{A_R} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{bmatrix} \begin{bmatrix} X_L \\ X_R \end{bmatrix} = \mathbf{W} \begin{bmatrix} X_L \\ X_R \end{bmatrix} \tag{2.3}$$

The indices for frequency and time frame (k, l) are omitted to avoid confusion. $\hat{S}$ denotes the estimated primary component and $\hat{A}_i$ denotes the estimated left and right ambient component, as mentioned before. The elements of the weighting matrix $\mathbf{W}$ are defined through the direct/ambient ratio (DAR) [8] and panning gains, which are estimated through principal component analysis (PCA) ([67], [70], [71], [72], [73]).

To find the DAR and the panning gains first the covariance matrix of the two stereo channels is computed.

$$\mathbf{R} = \begin{bmatrix} |X_L|^2 & X_L X_R^* \\ X_L^* X_R & |X_R|^2 \end{bmatrix} \tag{2.4}$$

The DAR and panning gains can then be estimated from $\mathbf{R}$ [73]:

$$\mathbf{R} = \begin{bmatrix} \mathbf{v_1} & \mathbf{v_2} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \mathbf{v_1} \\ \mathbf{v_2} \end{bmatrix} \tag{2.5}$$

where λ_1 and λ_2 represent the eigenvalues ($\lambda_1 > \lambda_2$) and $\mathbf{v_1}, \mathbf{v_2}$ are the eigenvectors. $\mathbf{v_1}$ in this case is the eigenvector to the bigger eigenvector λ_1. The energy of the direct sound is given by subtracting the smaller from the larger eigenvalue, while the smaller eigenvalue represents the energy of the ambient component respectively [71].

$$\begin{aligned} \lambda_1 &= \sigma_S^2 + \sigma_N^2, \\ \lambda_2 &= \sigma_N^2 \end{aligned} \tag{2.6}$$

Further the panning gains for the direct sound are defined by the elements of the eigen-

[8] The DAR is related to the before mentioned DRR. The DAR does not necessarily involve reverberation but diffuse sound.

vector to the bigger eigenvalue λ_1.

$$\begin{aligned} g_L &= v_{1,1}, \\ g_R &= v_{1,2} \end{aligned} \tag{2.7}$$

For the direct/ambient approach the whole channel is used for calculations and is not further divided in sub-bands. The computation of the whole channels diminishes the processing time, as the covariances of the given stereo channels are scalars and thus the covariance matrix for a stereo signal has the dimension $2x2$. If the covariance of sub-bands are calculated the covariance matrix will result in a three dimensional matrix $2x2xX$, with X representing the number of sub-bands. Because this simplification is made, signals with differently panned components (instruments) are expected to be confusing for this algorithm, as the panning gains will point at the component with the larger bandwidth and higher energy. Thus, the whole stereo signal is panned towards the found direction. To avoid fast changes in panning the covariance matrix is computed recursively [2].

$$\mathbf{R} = \rho\mathbf{R} + (1-\rho)\begin{bmatrix} |X_L|^2 & X_L X_R^* \\ X_L^* X_R & |X_R|^2 \end{bmatrix} \tag{2.8}$$

Without concerning about the DAR an equation containing only the panning gains would look like this:

$$\begin{aligned} \hat{S} &= \hat{g_L} X_L + \hat{g_R} X_R \\ \hat{A_L} &= X_L - \hat{g_L}\hat{S} \\ \hat{A_R} &= X_R - \hat{g_R}\hat{S} \end{aligned} \tag{2.9}$$

But it is found ([69], [67]) that with the PCA-based PAD the primary component will be over estimated if the signal is not panned. To correct this estimation error, Goodwin [67] proposes a modified PCA, introducing $\gamma_S = \frac{\lambda_1 - \lambda_2}{\lambda_1}$ and $\gamma_N = \frac{\lambda_2}{\lambda_1} = 1 - \gamma_S$ as coefficients

to weight the direct sound energy and the ambient sound energy based on the eigenvalue ratio. This results in the weighting matrix:

$$w = \begin{bmatrix} \gamma_S \cdot g_L & \gamma_S \cdot g_R \\ 1 - \gamma_N \cdot g_L^2 & -\gamma_N \cdot g_L \cdot g_R \\ -\gamma_N \cdot g_L \cdot g_R & 1 - \gamma_N \cdot g_R^2 \end{bmatrix} \tag{2.10}$$

Considering the limitations mentioned by Merimaa et al. [69], again this calculation will over estimate the primary component if the signal is not panned. To compensate the PCA noise introduced by the provided algorithm, the direct component factor needs to be $\gamma_S = 0$, if the value drops below a predefined threshold α. This threshold α is an empirical value.

2.3.2. Spatial synthesis

During spatial synthesis inter-channel relationships are converted to interaural relationships of the later rendered binaural signal. First the panning gains are used to distribute the primary component on the front speakers and a low-pass filtered version is send to the subwoofers. Later the ambient signals are generated for the surround channels. Further, all left ear signals are added up and all right ear channels are added up to gain the two channels of the final binaural audio.

All equations containing HRTFs and BRTFs are calculated per channel. During the following calculations the index for the two channels are omitted for clarity. The factors such as attenuation and delay are the same for the left and right channel.

Vector-based amplitude panning

Many surround setups contain a centre speaker. This is due to surround sound being a standard for movie theatres, where the centre speaker is used to ensure the dialogue coming from the front, regardless the location of the seat in the theatre. The existing stereo image, thus, has to be projected on the three channel setup (left channel, centre

channel, right channel). Therefore vector-based amplitude panning is computed ([74], [75], [76]). The following calculations consider a two dimensional setup, but it could also be upscaled to three dimensions.

First the position of the loudspeakers used for the stereo image are defined by the unit vectors pointing from the receiver towards the loudspeakers.

$$\mathbf{l}_i = \begin{bmatrix} cos(\alpha) \\ sin(\alpha) \end{bmatrix}, \tag{2.11}$$

In this equation the index i refers to the position of the loudspeaker and α refers to the angle between the listeners perspective (facing straight forward) and the loudspeaker, as shown in Fig. 2.9 for a two loudspeaker setup. The vectors $\mathbf{l}_i$ pointing to any number of loudspeakers are combined in the matrix $\mathbf{L} = [\mathbf{l_1 l_2} ... \mathbf{l_n}]$. Further the vector $\mathbf{p}$ is defined, pointing at the virtual source identified in the given stereo image. The direction of $\mathbf{p}$ is calculated as a linear combination of the loudspeaker vectors of the stereo setup weighted with the panning gains (Sec. 2.7).

$$\begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} g_L \mathbf{L}_{1,1} + g_R \mathbf{L}_{1,2} \\ g_L \mathbf{L}_{2,1} + g_R \mathbf{L}_{2,2} \end{bmatrix}, \tag{2.12}$$

All vectors are unit-vectors and fulfil the requirement $||\mathbf{x}|| = 1$.

When the direction of the virtual source ($\mathbf{p}$) is determined, the panning gains for the extended stereo setup (left, centre, right) can be calculated as:

$$\mathbf{p}^T = [g_1 g_2 g_3]\mathbf{L}_{123} = \mathbf{g}\mathbf{L}_{123}. \tag{2.13}$$

As $\mathbf{p}$ is already known, the equation is rearrange to calculate $\mathbf{g}$. Therefore $\mathbf{L}$ has to be inverted. Jeon et al. [76] proposes using the *Moore-Penrose* inverse matrix introduced by Courrieu [77], which is defined as $\mathbf{L}^+ = (\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T$. Thus, the panning gains $\mathbf{g}$ can

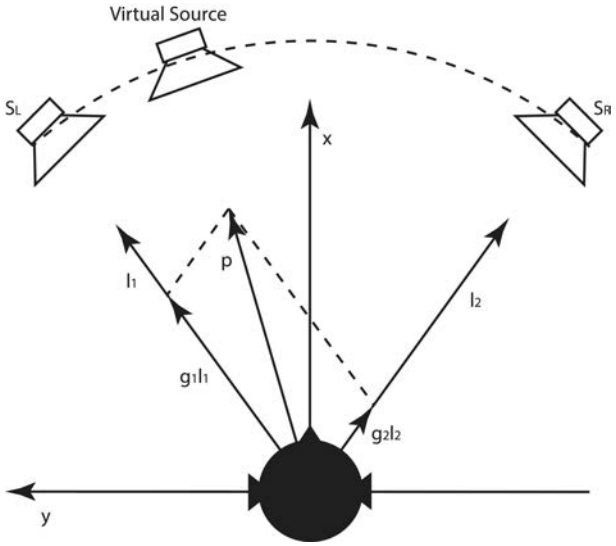


Figure 2.9.: The normalised vectors $\mathbf{l}_i$ point towards the loudspeakers. The normalised vector $\mathbf{p}$ points towards the virtual source.

be calculated as:

$$\mathbf{g} = \mathbf{p}^T \mathbf{L}_{123}^{+}. \tag{2.14}$$

The so calculated panning gains result in curves as shown in Fig. 2.10. The weakness about these panning gains appears in the possible negative panning gains near the far ends of the stereo image, which can result in an out-of-phase effect. Further, these panning curves lead to the detent effect, when a listener moves closer towards one virtual loudspeaker or if the virtual loudspeaker setup rotates so that $\alpha_1 \neq \alpha_2$. In such an unsymmetrical setup the virtual source travels towards the closer loudspeaker, regardless the intended spatial image. To avoid this behaviour, Jeon et al. [76] presents a method for multiple-wise vector base non-negative amplitude-panning (MVBNAP) providing an angle-dependent weighting of the matrix $\mathbf{L}$ [9] as:

$$\mathbf{L}_{123}^{'} = \mathbf{W}(\Theta)\mathbf{L}_{123} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \delta(\Theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{L}_{123} \tag{2.15}$$

[9] The equation for the angle-dependent weighting matrix was corrected from the author and differs from the original equation found in [76].

where $\delta(\Theta)$ is the angle-dependent function:

$$\delta(\Theta) = \left(0.5 + 0.5cos\left(\frac{2\pi\Theta}{2|\Theta_0 - \Theta_p|}\right)\right)^{\varphi}, p = 1 \; or \; 2 \tag{2.16}$$

Θ is the angle between the virtual source vector $\mathbf{p}$ and the listeners perspective. Θ_0 is the angle between the centre loudspeaker and the listeners perspective, Θ_p is the angle between loudspeaker vector one respectively two and the listeners perspective. φ is a fitting value, with different empirically found best values depending on the setup ($\varphi = 0.82$ for $30°, 0°, -30°$ and $\varphi_1 = 0.48$, $\varphi_2 = 0.97$ for $40°, 0°, -20°$, [76]). The introduced weighting matrix $\mathbf{W}$ does not only prevent from the detent effect taking place, but also prevents the panning gains from becoming negative at the far ends of the stereo image (Sec. 2.11).

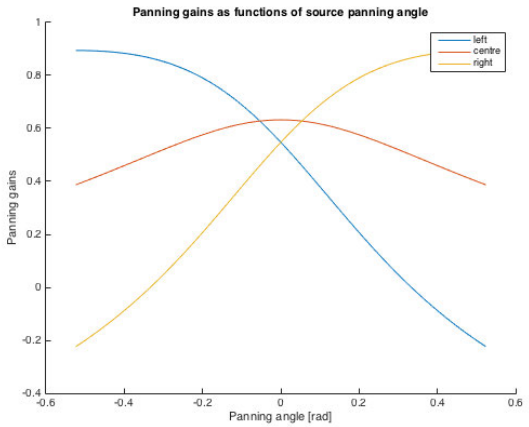


Figure 2.10.: The panning gains for all three speakers (left, centre, right) are plotted. The centre function is seen to be relatively flat, what leads to the detent effect for unsymmetrical loudspeaker setups. Further the panning gains for the bigger angles are negative.

The positions of the loudspeakers the stereo image is mapped on (left, centre, right), are defined through the HRTFs belonging to this very positions for auralization. The resulting front left and front right channels (Y_{FL} , Y_{FR}) contain an ambient amount when rendered. The resulting centre channel (Y_{FC}) only contains the direct component. The calculated ambient signal contained in the left and right channel is weighted by $\frac{1}{\sqrt{N_A}}$, where N_A is the number of channels used for propagation the ambient signal. In the USC 10.2 setup $N_A = 9$, because all channels are used as for the propagation of ambient signals, but the centre and the LFE channels.

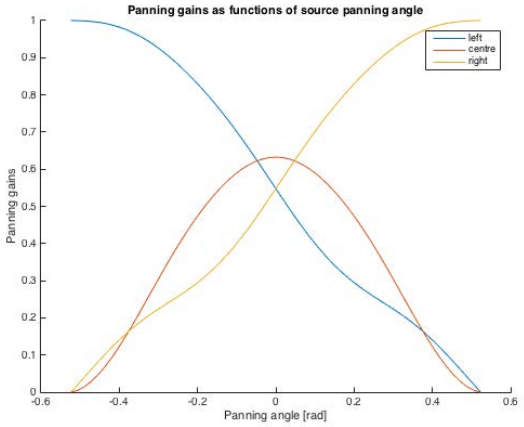


Figure 2.11.: The panning gains for all three speakers (left, centre, right) are plotted. The centre function is steeper towards the far angles which prevents the detent effect. Further the panning gains for the bigger angles are positive.

$$
\begin{aligned}
Y_{FL} &= g_L \hat{S} + \frac{\hat{A}_L}{\sqrt{N_A}} \\
Y_{FC} &= g_C \hat{S} \\
Y_{FR} &= g_R \hat{S} + \frac{\hat{A}_R}{\sqrt{N_A}}
\end{aligned}
\tag{2.17}
$$

Side-wall and ceiling reflections

In USC 10.2 setup the side-wall and ceiling refections are aggregated as the wide channel (WL/WR) and high channel (HL/HR) signals. Therefore the HRTFs, which represent the direction of this channels, are converted to BRTFs containing the necessary room reflection informations in form of constant SC. Both types of reflections have specific characteristics, which inhere in the used conversion.

The density and arrival time of side-wall reflections influence the perceived width of a source. Important for the apparent source width is the lateral efficiency [1, pp. 203-204]. The bigger the amount of early lateral reflections, the wider a sound source is assumed. Further the enrichment of the signal through early reflections lead to a smaller IACC with further leads to an increased preference [78]. The IACC can be calculated via the inter-aural cross correlation function (IACF), comparing the impulse responses

of the left and right ear signal, which is defined as [1, p. 200]:

$$IACF_{t_1,t_2} = \frac{\int_{t_1}^{t_2} p_L(t) p_R(t+\tau) dt}{\left[\int_{t_1}^{t_2} p_L(t) dt \int_{t_1}^{t_2} p_R(t) dt\right]^{1/2}} \tag{2.18}$$

Depending on the bounds of integration, three IACCs can be identified (Tab. 2.2):

	t_1	t_2
$IACC_{E(early)}$	$0ms$	$80ms$
$IACC_{L(ate)}$	$80ms$	$500...2000ms$
$IACC_{A(ll)}$	$0ms$	$500...2000ms$

Table 2.2.: IACF with different lower and upper bound for integration, resulting in $IACC_{E(early)}$, $IACC_{L(ate)}$ and $IACC_{A(ll)}$.

The $IACC_E$ and the degree of width perception correlate negatively, equally the $IACC_L$ and the degree of envelopment of sound. These values can be analysed to estimate the preference according to Hidaka et al. [78].

To generate early reflections the HRTFs of the wide channels are attenuated and the phase is delayed. The attenuation and the phase delay is frequency dependent and can be expressed as [10] [2]:

$$D_i = (1-\eta_i) e^{j\omega\tau_i}. \tag{2.19}$$

i is the index for the frequency bin in this and all following equations. The BRTF containing the early wall reflections is given by:

$$BRTF_{w,i} = D_{w,i} \cdot HRTF_{w,i}. \tag{2.20}$$

Ando [79] identifies some requirements for improving a subjective preference using early side-wall reflections, which are summarised by Lee et al. [2] in the following conditions. The walls should reflect the sound from within an angle of $\pm 55^\circ$ and the

[10] The original equation [2] has been changed by the author to complete the requirements for attenuation. Therefore, the factor η which is named attenuation in the paper is replaced by $(1-\eta)$.

side-walls should be asymmetric regarding the location of a listener. To satisfy the requirements, the side-wall reflections should not be symmetrically distributed over the two wide channels. Instead the left channel is processed as such, while the right channel is further shifted towards the high right channel. Thus, the result is an asymmetrical reflection, which still remains within the proposed angle of $\pm 55^{\circ}$. To achieve the shift towards the high channel the result is a superposition of a reverse weighted reflection function. Thus, the equation for the right channel early reflections is:

$$\begin{aligned} BRTF_{w1,i} &= D_{w1,i} \cdot \beta HRTF_{w1,i} \\ BRTF_{w2,i} &= D_{w2,i} \cdot \sqrt{1-\beta^2} HRTF_{w2,i} \\ BRTF_{wr,i} &= BRTF_{w1,i} + BRTF_{w2,i} \end{aligned} \tag{2.21}$$

As the loudspeaker positions used for calculating this algorithm are virtual and thus can be changed arbitrarily, the shift of the right wall reflection channel could also be done by “moving the loudspeaker”.

The ceiling reflections are added using the high channels. Therefore the HRTFs defining the high channel positions are first equally attenuated and delayed as it was shown for the wide channel signals. But as mentioned by Lee et al. [2] distinct ceiling reflections lead to a higher IACC. Thus, the virtual ceilings are defined as diffusors. To gain characteristic diffusor modulations the phase can simply be randomised. Another approach is the quadratic residue diffuser ([80], [81]), which was proposed by Ando [79] for concert hall ceilings. The quadratic residue diffuser has a constant magnitude frequency response, while the phase is randomised.

To create a digital quadratic residue diffuser, first a series of unique numbers with the periodicity N (in the positive range $[0...N]$) is created by:

$$s_n = s^2 \ mod \ N \tag{2.22}$$

Where N has to be an odd prime number. Based on this series a phase randomiser can

be implemented for ceiling diffusion by:

$$D_{QRdiff} = e^{j2\pi s_n} \tag{2.23}$$

Using this diffuser method the ceiling reflections can be calculated through:

$$BRTF_{c,i} = D_{QRdiff} \cdot D_{c,i} \cdot HRTF_{c,i}. \tag{2.24}$$

$D_{c,i}$ is calculated analog to Eq. 2.19 but with inserting the desired values for the ceiling. A different diffusor series (based on different N) should be used for each channel to assure the most possible de-correlation, as the result is constant for the same N.

The attenuation η and delay τ contained in the previous equations can be freely chosen in reasonable ranges for the walls and the ceiling. η for the proposed calculations should have a value between 0 (completely reflecting wall) and 1 (completely absorbing wall). The maximum delay τ is limited by the length of the BRIR.

Further, the ambient component is again weighted by $1/\sqrt{N_A}$ and added to the wall and ceiling channels (analog to Eq. 2.17).

Surround channels

The surround channels are used to simulate the diffuse sound. Therefore the ambient signal is phase delayed and phase randomised. The phase randomisation again is used to de-correlate the signals to enhance externalisation, a perception of wideness and envelopment.

The delay factor for the surround channels is equally calculated as for the side-walls and ceiling D_i (Eq. 2.17) [11]. Signals arriving more than $1ms$ after the first wave front from different directions containing the same content do not affect localisation but lead to a greater perceived envelopment (precedence effect). This effect is used when generating the surround channels adding the delay D_i.

[11] The proposed equation in [2] is corrected by the author. The delay factor is annotated as $exp(j2\pi\tau)$, but it should be equal to the delay presented for phase delay for side-wall and ceiling channels.

The surround channels are not attenuated, but scaled by the number of surround channels. USC 10.2 setup includes three surround channels (left, back, right), but for the later propagation nine channels are used. Further the phase randomisation for the surround channels is processed with random values uniformly distributed over the interval [0,1] [2].

$$D_{Udiff} = e^{j2\pi v_i} \tag{2.25}$$

The three phase randomisation intervals v_l, v_b and v_r should be different to de-correlate the signals. The resulting BRTFs for the left and right surround channels are given by:

$$BRTF_{A,i} = D_S \cdot D_{Udiff} \cdot HRTF_{A,i}/\sqrt{N_A}. \tag{2.26}$$

The rear surround channel reproduces the left ambient and right ambient signal and thus, the result is weighted by 0.5.

$$BRTF_{SB,i} = 0.5 \cdot D_S \cdot D_{Udiff} \cdot HRTF_{A,i}/\sqrt{N_A}. \tag{2.27}$$

Subwoofer

Additionally the direct signal is send to the LFE channels, to achieve a bass boost. The LFE signals are generated by applying a low-pass filter (120 Hz cut-off frequency) to the virtual loudspeakers at $\pm 90^\circ$. The gain of these channels depend on the characteristics of the virtual room and the subjective taste.

Alternative methods for de-correlating signals

In the proposed algorithm the DAR is detected and used to calculate gains for the direct and ambient channels. Thus, the full signal (separated in left and right channel) is reproduced on the twelve channels. The ambient channels have to be de-correlated by

e.g. phase randomisation. Applying the proposed algorithm the two stereo speakers (FL, FR) reproduce the original stereo signal. The center speaker enhances the stereo image by a monophonic mixdown of left and right channel. Thus, the number of speakers is scalable (e.g. 2.0, 5.1, 10.2).

Usher and Benesty [82] proposed an comparable algorithm. But in their case the original signal is explicitly split into a direct and an ambient component. The de-correlated signal part is taken as ambient component and the remaining part as direct component. Thus, the original stereo signal is not preserved. Reproducing a signal which is processed this way on stereo speakers will only contain the extracted direct component.

Alternative to phase randomisation by a uniform distributed interval $[0,1]$ or by applying simulated quadratic residue diffuser other approaches can be used for de-correlating channels to achieve a greater envelopment and *out-of-head* localisation. Choi et al. [83] introduces an approach modelling realistic angle-dependent early reflection simulations ($delay + gain + HRTF = BRIR$). A set of first order infinitive impulse response shelving filters are used to approximate HRTFs for the two sides of the head at $\pm 90°$, as these two directions are used as the only origins of reflection in this algorithm. Thus, the gains and delays contained in the HRTFs can be applied without the high-cost computation involved with actual HRTF filtering.

Schroeder et al. [39] proposes taking advantage of a reverberation model, based on ray-tracing, to de-correlate signals and simulate late reverberation. In this approach real room reverberation is modelled to generate a filter, with which the BRIRs are generated from HRIRs. The further processing involves convolution of the signal with relatively long BRIRs, as the usage of late reverberation lead to long BRTFs.

Borß [84] uses the room geometry to define the direction of early reflections, the reverberation time is defined frequency-dependent and an echo density profile is used for further room characterisation. The model leads to IACCs which can be measured in real rooms and thus, realistic room impressions can be obtained.

Binaural synthesis

As mentioned above the loudspeaker directions are obtained by filtering with the HRTF pair, that belong to the defined direction. One direction is represented by one pair

of HRTFs - one HRTF for each ear. The two channels of the final binaural signal are superpositions of the prior calculated signal representations (BRTFs) of the virtual multi-channel setup, multiplied with the corresponding audio signal ($\hat{S}$, $\hat{A}_l$, $\hat{A}_r$).

The final equations to calculate the channels are [2] [12]:

$$\begin{aligned} Y &= Y_S + Y_{AL} + Y_{AR} \\ Y_S &= (A+B+C)\hat{S} \\ Y_{AL} &= D\hat{A}_L \\ Y_{AR} &= E\hat{A}_R \end{aligned} \tag{2.28}$$

The variables A, B, C, D and E are defined as:

$$A = \sum_{c\in\{FL,FC,FR\}} g_c H_c \tag{2.29}$$

$$B = D_w[H_{WL} + \beta H_{WL} + \sqrt{1-\beta^2}H_{HL}] \tag{2.30}$$

$$C = D_c[D_{QRdiff}H_{HL} + D_{QRdiff}H_{HR}] \tag{2.31}$$

$$D = [H_{FL} + D_{Udiff}(H_{HL} + H_{WL} + H_{LFEL} + H_{SL} + H_{HR} + H_{WR} + H_{LFER} + H_{SR} + 0.5H_{BS})]/\sqrt{N_A} \tag{2.32}$$

$$E = [H_{FR} + D_{Udiff}(H_{HL} + H_{WL} + H_{LFEL} + H_{SL} + H_{HR} + H_{WR} + H_{LFER} + H_{SR} + 0.5H_{BS})]/\sqrt{N_A} \tag{2.33}$$

[12] Some modifications have been made by the author to correct the equations.

These calculations are processed per time frame and per channel. Further, all calculations are realised in the frequency domain. The three indexes (time frame, channel, frequency) are omitted for clarity. The HRTF of each channel is abbreviated as H with an index refering to the channel name.

The variable A must be calculated per time frame, as it contains the input-signal-dependent gain factors. B, C, D and E are calculated per loudspeaker location. If the loudspeakers are fixed to a virtual location, these calculations can be done only once at the start and further treated as constants. If dynamic binaural synthesis is required, all five equations have to be calculated per time frame, respectively per head movement. It is assumed that twelve channel dynamic binaural synthesis exceeds the processing capabilities of the chosen device and thus, was not implemented.

Overlap-add / Overlap-save

The proposed calculations are processed in realtime. Therefore, the audio file is read and passed as stream of audio blocks into the filter routine. Each audio block of the stream is transformed from the initial stereo signal block to the final binaural signal block and attached to the previous audio block, to assure a constant audio stream. If the audio block of length N_a is convolved with a filter impulse response of length N_f, the resulting block contains more samples than the initial block, namely $N_a + N_f - 1$. The same will happen if the filtering is processed in the frequency domain, as in the proposed algorithm. The audio block is first fast Fourier transformed with a desired number of points. In digital audio processing a block length between 128 samples and 1024 samples is common, to keep the latency low. The audio block B_a is zero-padded to the double length of the desired FFT length N_p to avoid circular convolution. After filtering, the resulting iFFT generates a block B_f with the length $2 \cdot N_p$. Thus, an overlapping block $B_o l$ is generated.

The audio stream requires a returning block from the filter routine with the same length as the initial audio block N_a. If the overlapping block $B_o l$ is discarded, the filter decay of the previous block is cut. To preserve the filter decay, the overlap-add or overlap-save method can be used [85].

Overlap-add and overlap-save differ as the one works with the output signal and the other works with the input signal. The overlap-add method involves partitioning the

signal in non-overlapping blocks. These blocks are zero-padded at the end of the signal to the length $N_a + N_f - 1$, filtered and then the overlapping block $B_o l$ of length $N_f - 1$ is added to the following (output) audio block B_f (Fig. 2.12). The overlap-save method describes partitioning the signal in overlapping blocks. The first audio block is supplemented with $N_f - 1$ zeros at the beginning of the block. After filtering the redundant (overlapping) block is discarded (Fig. 2.12). Overlap-add is suggested [85, p. 520], when the input is very long or continuous. Thus, overlap-add is convenient for audio streams.

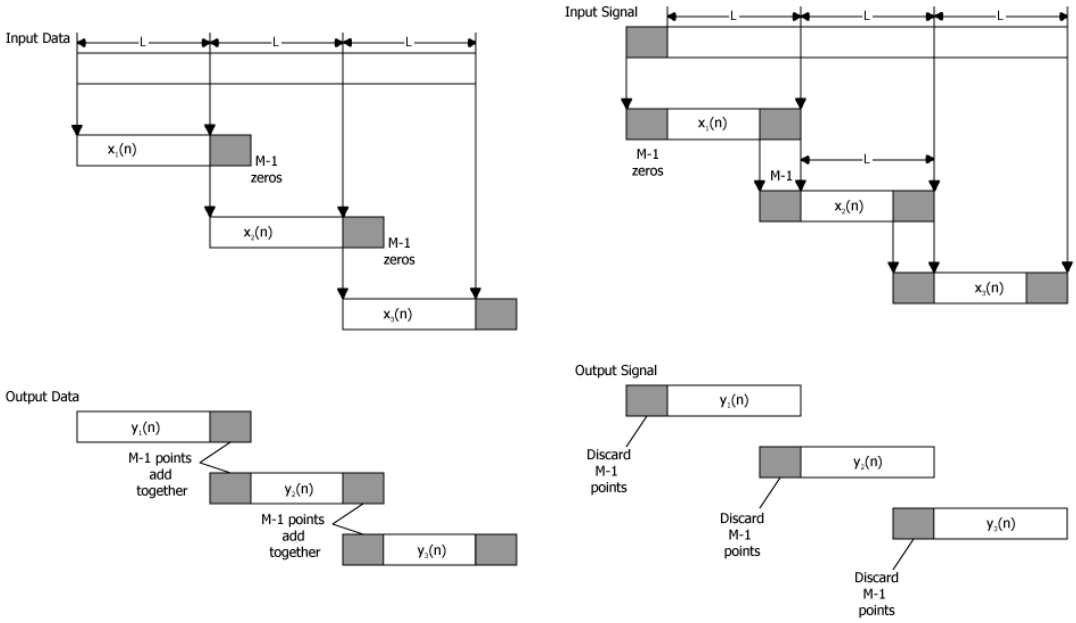


Figure 2.12.: The left figure illustrates the overlap-add method, where the non-overlapping blocks are zero-padded at the end and then added to the following blocks. The right figure illustrates the overlap-save method, where overlapping blocks are taken and the front is discarded to gain sequential non-overlapping blocks [3].

2.4. Augmented reality headphones

Augmented reality is a vast field of research, that can be divided in subclasses depending on the sense or senses addressed by augmentation. While for visual augmented reality various solutions already exist in research and commerce, the field of acoustic augmented reality gained less attention throughout the last decades but has cought up in the last couple of years. Reasons for that can be found when considering accessibility. While every modern smartphone is equipped with a camera that can be

pointed at a street, and information about the neighbourhood is projected on the screen, typical headphones are often not suitable for a valuable acoustic augmented reality. Augmented reality headphones have to be acoustically transparent or even better, to an adjustable extent so that the real acoustic surroundings and added virtual reality can be superimposed.

There are two ways to realise augmented reality headphones. The first way is to build the headphones using acoustically transparent materials, so that the surrounding acoustic environment will reach the ear canal as if no headphone at all is worn. This way a good localisation of the real sources can be achieved as it is received without major colouring as shown by Martin et al. [86]. The second way is to use headphones, which apply a strong attenuation (closed headphones), and mount microphones on the earcaps ([87], [88], [89], [90]). This way the surrounding soundscape is alterable in loudness and other acoustic features. The microphone signals could even be completely disabled, so that an acoustic isolation can be obtained if desired. In augmented reality solutions both ways can be convenient, depending on the use case, but the latter gives more opportunities for configurations.

2.4.1. Localisation

If headphones made from acoustically transparent material are used, a marginally (but significantly) impaired localisation of the real surrounding sources was found, if compared with not using headphones [86]. The degree of degradation depends on the (acoustically transparent) earpiece, as the ILDs and ITDs do not change. The SC changes marginally. This results in a higher front/back confusion rate, during the performed listening tests [86].

As mentioned in Sec. 2.1, ITDs and ILDs are the most important indicators for localisation in the horizontal plane. If the microphones are outside the ear canal - as they are when mounted on the enclosure of the headphones - the ITDs and the ILDs will differ from the real ones. Therefore, a localisation blur will occur in some solutions where microphones are used, depending on the physical size of the headphone. As the microphone position is outside the ear canal, the spectrum of the signal arriving at the microphones will be differently coloured, than signals entering the open ear canal. Tests driven by Homann et al. [90] reveal, that localisation in the horizontal plane will marginally be impaired, if the source is in front of the listener. Localisation behind the

listener and further in the median plane is strongly affected.

In medical research microphone positions for hearing aids have been intensely researched and evaluated regarding their capability of generating spatial cues. Mainly until 1990 it was investigated if there is a significant advantage of having microphones in a binaural meaningful position, such as in the ear canals, compared to positioning the microphones behind the pinna. The results were ambiguous. Westermann [91] found a significant advantage of positioning a microphone in the ear canal, while Leeuw [92] and Noble [93] could not find a significant advantage of a certain position. As discussed by Van den Bogaert et al. [94] comparing these studies is not trivial because these studies address different dimensions of localisation (frontal-horizontal hemisphere, frontal-vertical hemisphere and front/back confusion). Further, the different hearing aids models might process the data differently, which likewise might have an influence on the potential ability of localising sound.

The study presented by Van den Bogaert et al. [94] compared three different types of hearing aids regarding their ability of generating spatial cues. The three types were: in-the-ear, in-the-pinna and behind-the-ear. Two models of the behind-the-ear type were tested, which might be equipped with a different processing unit. The study showed that for a male speaker signal a good left-right identification can be achieved with all three types of hearing aids tested. Front/back discrimination was also reliable for all types, but there were differences between the models. Even one of the tested behind-the-ear device yielded adequate front/back detection. The localisation in the median plane was equally good over all four models (three types). All types led to an error rate between 23.2 ± 4.4 and 30.1 ± 9.0. The study revealed that the accuracy of localisation is reduced by hearing aids in all cases. But as users train on using hearing aids, further spatial cues, such as low frequency spatial cues, get more important and can compensate the loss of higher frequency spatial cues.

Hammershøi and Møller [95] examined different microphone positions and found the position $6mm$ outside the ear canal to be sufficient, because most of the directional information are already contained in the signals recorded at that position.

These studies lead to the impression, that different microphone positions are possible for binaural recording, as long as they are somehow close to the ear canals. If circumaural or supra-aural headphones are required, the microphone positions have to be reconsidered. When mounted on the outside of the earcaps of the headphones, the head

diameter is enlarged. This yields to location errors in the horizontal plane, depending on the source direction, as the virtual ITDs and ILDs differ from the natural ones. The equation to calculate the ITD as function of head diameter and source direction is given as [96]:

$$\hat{\tau} = \frac{a}{c}(sin\theta + \theta). \tag{2.34}$$

Where $\hat{\tau}$ denotes the ITD, a is the head radius, c is the speed of sound and θ the source direction.

As an example: When the head diameter is $18cm$, the microphones are mounted $3cm$ outside of the ear canal, the head diameter is enlarged in total by $6cm$. As is shown in Fig. 2.13, the resulting ITDs grow faster with the angle for a bigger head diameter. In this case the ITDs are magnified by a factor 1.3333. Vice versa the source seems to shift faster towards the sides if the ITDs are bigger than accustomed, which is the expected effect.

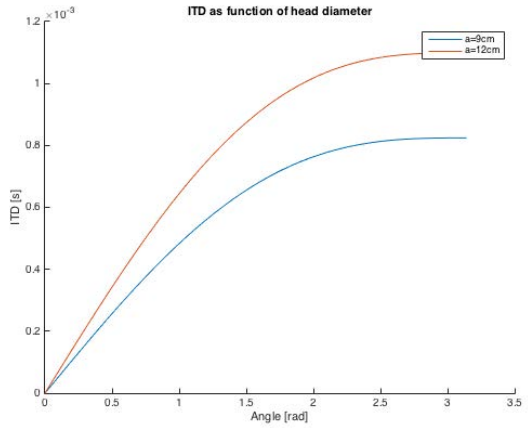


Figure 2.13.: This figure shows the perceived source angle in degrees (x-axis) in relation to the ITD in meter (y-axis).

When the microphones are placed in front or behind the headphones, acoustic shadowing will take place. A suitable position for a sufficient localisation should be evaluated in case of using circum-aural or supra-aural headphones.

2.5. Summary

Usually a sound event is received as two distinct signals - one at each ear. The signals can differ in loudness (ILD), time of arrival (ITD) and spectral colouration. These differences are used by the auditory system as cues for acoustic localisation and further analysis. Further the outer body (torso, head, pinna) influences a sound when traveling from a source to the ear canals. These influences can be recorded as angle-depended HRIRs (free-field), containing the gain and phase factor per frequency. Equally, the structure of the surrounding environment modulates a sound. Thus, in rooms RIR or BRIRs can be measured, again containing the gain and phase factor per frequency regarding a spatial relation between source and receiver.

In comparison to audio reproduced with loudspeakers, audio reproduced with headphones does not travel through space and thus, is not modulated by our body or the surrounding room. If a virtual listening space is desired, HRIRs or BRIRs can be used as filters to add spatial informations. To further enhance a room impression of stereo audio listened to with headphones, the signal can be up-mixed to a desired number of channels. These additional channels are used to add wall- and ceiling-reflections and diffuse-sound. Thus, a high impression of envelopment can be achieved. To reduce the spectral colouration introduced by a headphone, a system can be compensated using a weighted filter based on the inverted HpTF.

For applications addressing augmented acoustics, headphones are needed which are acoustically transparent, either by building them using acoustically transparent materials, or by mounting and wiring microphones to the headphones.

3. Commercial solutions

Several stationary solutions [1] for binaural synthesis were implemented and tested by researchers to assure sufficient low latencies and high quality, which lead to good acoustic scene simulations using *Raven*, *Soundscape Renderer* or other environments. Usually powerful computers and expensive hardware need to be used. To maximise the processing power GPU processing, multi-threading and computer clustering are often included techniques. The major problem of these setups is that the listener is forced to a certain location (near the computer). This situation leads to the thought, that mobile devices could yield sufficient results in binaural synthesis to liberate a listener from stationary hardware and thus, to realise more realistic augmented/virtual acoustic solutions [97].

Furthermore, technology regarding binaural synthesis is increasingly becoming available for commercial use. Several commercial solutions are offered presently, such as the head tracking solution *Beyerdynamic Headzone*® or binaural rendering headphones such as *NEOH* headset and *Jabra Intelligent Headset*. Further *Dysonics* released the software *Rappr*, a five channel up-mixing and binaural auralization solution, and an affordable headtracker called *Rondo Motion*, which is connected via bluetooth. All of these solutions involve either stationary computers or a particular hardware.

3.1. Binaural applications

Solutions for smartphones are of interest, because smartphones are widely spread and their processing power is continuously increasing. Modern smartphones are equipped with relatively powerful processors, big memory capabilities and various sensors, which makes them predestinated for all kinds of modern technologies including binaural syn-

[1] Stationary in this case refers to the hardware used for processing the simulations.

thesis.

In the field of binaural auralization some applications already exist for *iOS* while for *Android* and *Windows Mobile* no relevant applications were found by the author, which is unique for these platforms. Most available applications are demos for binaural rendering techniques and frameworks such as *Two Big Ears - 3Dception* (cross-platform solution) or *MN Signal Processing - Virtual Room*. Further *Accessibility* released the *Augmented Stereo* application, which is a dynamic binaural audio demonstrator using the internal compass of the iPhone to adjust the dynamic source position. Some games are using binaural synthesis or even concentrate completely on the audio layer. Examples for these applications are *Blowback - Die Suche* (*iOS / Android*) and *Ear Games - EarMonsters*. All of these games binaurally render dynamic synthetic sound sources and use either display tapping or even motion sensor data as control commands.

Headquake® Pro released by Sonic Emotion is a music player with binaural filtering. The music contained in the device media library (*iTunes* library) can be reproduced and the two virtual stereo speakers can be symmetrically shifted on the frontal horizontal plane up to $\pm 90^{\circ}$. A non-individual headphone compensation for various headphone models [2] is implemented.

The most advanced mobile binaural auralization application seems to be *Dysonics - Rondify*. It does dynamic binaural auralization of stereo music files provided by the streaming service *Spotify*. The sound is enhanced by reverberation to convey the impression of a surrounding room. This application supports dynamic binaural synthesis, if used together with *Dysonics - Rondo Motion* as headtracker. When tested by the author, this application did not work. The application continuously crashed while initialising without a tracker connected.

3.2. Augmented reality applications

Most mobile applications addressing augmented reality focus on visual augmentation. Often the smartphone can be pointed at an object and additional information is displayed on the screen. Thus, these applications are implemented as daily helpers. Most of the augmented audio applications are games. These games often use the motion

[2] The filters for specific headphone models can be bought from within the application.

sensors (and GPS) to track the direction of the users sight and her position, while synthetic audio is dynamically rendered. The smartphone is mostly pointed towards the desired direction, so no external head tracker is needed. Thus, it is independent of head movements, even though it is assumed that the smartphone is pointing the same direction the user is facing. These games rely on using a somewhat acoustically transparent headphone, so that the real acoustic surrounding is audible. Examples for this kinds of applications are *Re-Sounding - Meltdown* or *Eidola Multiplayer* presented by Moustakas et al. [98]. *Inception - The APP* is an example for an augmented virtuality game. The microphone of the smartphone captures the surrounding acoustic environment, which is then modified and embedded into the virtual soundscape. Both methods of augmentation (augmented reality / augmented virtuality) are subcategories of the main category mixed reality [99].

The *Technische Universität Berlin* and the *Technische Hochschule Aachen* in cooperation currently investigate an orientation system for blind people (*Verbundprojekt OIWOB*). The visual scene around the user is analysed and detected relevant objects are auralized. The next step of such a system could be using augmented reality headphones to augment the real acoustic soundscape by the auralized objects. A reliable helper system for blind people have to be very accurate (head tracking, synthesis) to avoid any potential accidents.

3.3. Augmented reality headphones

Recently binaural recording for consumer use is gaining more interest. Headphones equipped with microphones on both sides were recently released or are announced for a soon release. Examples are *Roland CS-10EM* and *Hooke Audio Verse*. The scope of these headphones involve recording and reproducing audio but usually they do not support simultaneous use of microphones and headphones, due to bluetooth limitations or poor acoustic insulation between microphone and transmitter.

3.4. Summary

Mobile applications addressing binaural synthesis or augmented acoustics are usually games or demonstrators. Only a few audio players with HRTF filtering or up-mixing exist. The quality of the existing mobile binaural synthesis solutions is mainly poor, as no real impression of externalisation or sound envelopment was perceived.

4. Hardware and software concept

As part of this work it was planned to realise an *iOS 8.0* audio player for *iPhone* and *iPad* as practical part of the work, containing realtime stereo up-mixing and binaural auralization. Further, a hear-through headset should be constructed. The combination of both hardware components then should serve as a demonstrator for an enhanced augmented acoustics audio player.

The decision of using *Apple* hardware was also decided to benefit the ZIM project of the *Technische Universität Berlin*, within which a headphone will be developed and connected to an *Apple* mobile device. On this mobile device an application shall run supporting up-mixing and dynamic binaural auralization.

4.1. Hardware concept

The hardware for realising the demonstrator is decided to be an *iPhone*/*iPad*. This decision is due to the good performance of these devices and because additional audio hardware is needed to care for required audio connections (stereo audio input). *Apple* mobile devices support such peripherals.

Further, a hear-through headset is obligatory. Hear-through headsets are not commercially available until now, but there are various ways to construct such a device. The author required controllable acoustic transparency, meaning that headphones with a high degree of attenuation should be used, that have binaural microphones attached (one microphone on each side of the head).

4.1.1. Mobile device

The application should be implemented for *iOS 8* . Therefore, the device running the application has to be one of the following:

1. *iPhone 6 Plus*
2. *iPhone 6*
3. *iPhone 5 S*
4. *iPhone 5 C*
5. *iPhone 5*
6. *iPhone 4 S*
7. *iPad Air 2*
8. *iPad Air*
9. *iPad Mini 3*
10. *iPad Mini 2*
11. *iPad Mini*

The CPU of the devices *iPad 2* and *iPad 3* are assessed not to be powerful enough for running the application, even though they are able to run *iOS 8* .

The chosen device for testing was the *iPad Air* (I. Generation, 2013). It is equipped with a $64-bit$ *Apple A7* system-on-chip dual-core $(1.3-1.4GHz)$ processor, an *Apple M7* motion processor and a *PowerVR G6430* quad-core graphic processing unit. Via the *Lightning Connector* additional hardware can be connected, such as audio interfaces.

4.1.2. Hear-through headset

The hear-through headset was build by adding two microphones to common headphones (Fig. 4.1). Thus, the degree of acoustic transparency of the headphone is controllable. Further, the headphone and the microphones may be exchanged in the course of future development.

Figure 4.1.: The hear-through headset realised as closed-back circum-aural headphone with two microphones mounted on the enclosure.

The demonstrator software runs on an *iPhone*. Thus, any common headphone can be connected. Three main types of headphones are defined in ITU [100], such as circum-aural, supra-aural (supra-concha) and in-ear (intra-concha). All of these types can be build as closed-back or open-back design. Closed-back refers to an enclosure, which suppresses the emission of sound outside the headphones as much as possible.

For the intended system a low sound emission to the outside is necessary to avoid feedback between the microphone and headphone, as the two microphones are placed near or even on the enclosure of the headphones. Thus, closed-back headphones are used. Usually, open-back headphones can provide a more linear frequency response, whereas closed design headphones tend to cause a bass boost. As described above the frequency response of the headphone is compensated in the software.

For the conceptualised demonstrator two peripherals had to be connected to the smart-

phone, which were the headphone and the microphones (considered as one stereo microphone). Both have two independent audio channels and for both connections a high audio transmission quality is required.

Headphones with active noise cancelling [1] could be modified, as they contain already all the required hardware. To use headphones with active noise cancelling, the microphone connections had to be opened and redirected to use the recorded sound with the above mentioned application. Another variant is to simply attach microphones to any suitable (frequency response, attenuation) common headphone.

Headphones

Two headphones were selected for testing with the demonstrator. The first headphone was the circum-aural *Beyerdynamic DT 770 Pro* with 32Ω impedance (Fig. 4.2; data sheet: App. B), which provides a high acoustic isolation and thus, a high attenuation outside the ear pieces. The frequency response covers a wide range and is tolerable for a closed-back headphone.

The second headphone was the *iPhone 6 earPod* (Fig. 4.2) as is delivered with any current *Apple* mobile device. This headphone is inserted into the ear and its enclosure design is open-back. It does not fulfil the closed-back requirement, but was investigated as it is a very cheep variant. Its technical specifications can be found in the *Apple* user forum [101]:

- Impedance: 45 Ω
- Sensitivity: 109 dB
- Frequency: 5 Hz - 21 kHz
- Operating principal: Open air

Because the attenuation of the *iPhone 6 earPod* enclosure is not very high (open-back), the microphones were placed on a hairband close to the ears and not mounted onto the

[1] Active noise cancelling is a procedure, where the surrounding noise is recorded and then eliminated through the phase-shifted noise signal.

enclosure. The frequency responses of both headphones are plotted in Fig. 2.5.

Figure 4.2.: The headphones used for the demonstrator. Left: *Beyerdynamic DT 770 Pro*; Right: *Apple iPhone 6 earPod*.

The headphones should be connected to the smartphone via cable, so that no loss of quality or additional latency is aggregated. Further the headphones remain independent of batteries and work with the smartphone at any time. A bluetooth audio chip such as the *Roving Networks RN52* could be used for audio streaming from the smartphone towards the headphones, as it supports the bluetooth protocol A2DP [2] in sink mode. With minor modifications a headphone jack can be connected to the chip. Thus, the headphone is connectable to the bluetooth chip, which could be wirelessly connected to the smartphone.

Microphone

The microphones are needed to make the headphones acoustically transparent. As the microphones are mounted on the head (either on the headphone enclosure or near the ear with a hairband), small microphones are preferred. Possible models are the electret capacitor microphones *Panasonic WM-61a* (App. B) or MEMS microphones such as the *ADMP 401*.

The positioning of the microphones is of interest as explained in Sec. 2.4. When using the *iPhone 6 earPod* the microphones were mounted on a hairband and positioned close

[2] A2DP (advanced audio distribution protocol) supports high quality stereo audio streaming. Transmitted audio is compressed using various codecs such as SBC (standard), MP3, AAC and AptX. Bitrates up to $345 kBit/s$ are used.

to the ears, shifted towards the front of the head. When using the *Beyerdynamic DT 770 Pro* two microphone positions were examined: Configuration 1) On the enclosure facing to the sides ($\pm 90°$); Configuration 2) On the enclosure facing to the front ($0°$), as shown in Fig. 4.3. As the *Beyerdynamic DT 770 Pro* has relatively big ear-pieces ($4-5cm$ on one side depending on how much the cushion is compressed), configuration 1 is expected to result in unnatural ITDs, as the head diameter is artificially extended (Sec. 2.4.1). In configuration 2 unnatural colouring could occure through acoustic shadowing, when a sound source is located on the rear hemisphere. Further, reflections on the headphones could result in colouring for positions in front and beside a source.

Figure 4.3.: The microphone positions used on the *Beyerdynamic DT 770 Pro*. Left: In configuration 1 the microphones are mounted on the outside of the enclosure facing $\pm 90°$. Right: In configuration 2 the microphones are mounted on the front of the enclosure facing in front.

Speech, traffic, etc. are assumed to be important sound sources, that should be identifiable while using augmented reality headphones. As the microphone signals are mixed with the audio the frequencies above about $8kHz$ are assumed not to contain relevant information for perceiving the surrounding environment. Further, low frequencies were cut to avoid impact sounds distracting the perceived audio. Thus, the audio information of the surrounding acoustic objects was expected to be in the range between $100Hz - 8kHz$.

Apple mobile devices do not provide any native connection for stereo microphones. The four-contact jack for headsets only support a mono microphone connection.

iOS devices compared to other smartphones do not use the micro USB port for charging. The native *Lightning Connector* or the older *Dock Connector* can be extended with the

Apple Camera Connector to obtain a fully operational USB 2.0 I/O port. Thus, USB audio interfaces, which can communicate via the native *CoreAudio* driver, are applicable with *iOS* devices. For the demonstrator the *Behringer UCA 222* was used to connect the stereo microphone to the *iOS* device, because it provides sufficient inputs/outputs and it is USB bus powered.

The *Behringer UCA 222* (Fig. 4.4) is equipped with one stereo line input and one stereo line / headphone output. As the microphone output level is very low, an additional microphone preamp was required for the desired connection. The *Velleman Super Stereo Ear Amplifier Kit - MK 136* (Fig. 4.5) is a stereo microphone preamp assembly kit which converts analog stereo microphone input into line output. The kit was modified by the author to run on a small rechargeable LiPo battery, in order to reduce the weight. The complete hardware connection diagram is shown in Fig. 4.6.

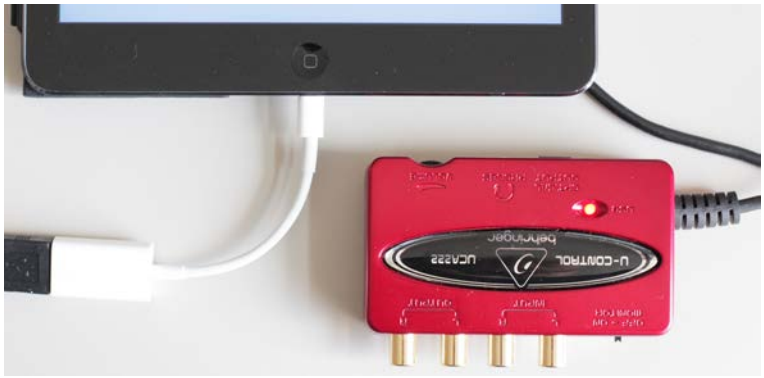


Figure 4.4.: The *Behringer UCA 222* audio interface, which can be connected to *Apple* mobile devices via lightning connector.

Further, a bluetooth connection using *A2DP* was investigated. When bluetooth is used no additional cables are necessary as shown in Fig. 4.7. The bluetooth chip *Roving Networks RN52* provides two microphone inputs and has the ability to stream audio. To gain the ability of streaming audio from the chip to another device, the correct firmware has to be installed, which supports *A2DP* in source mode. The firmware that supports streaming does not support receiving audio. Thus, the function of this chip has to be defined at the start and can not be changed while running. As the headphones were to be connected via cable, the functionality of this chip was sufficient.

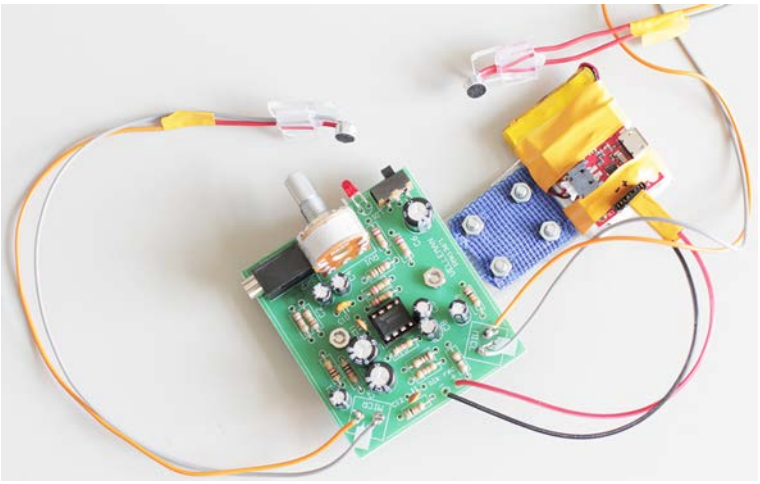


Figure 4.5.: The *Velleman Super Ear* microphone preamp with the two microphones connected and the LiPo battery pack.

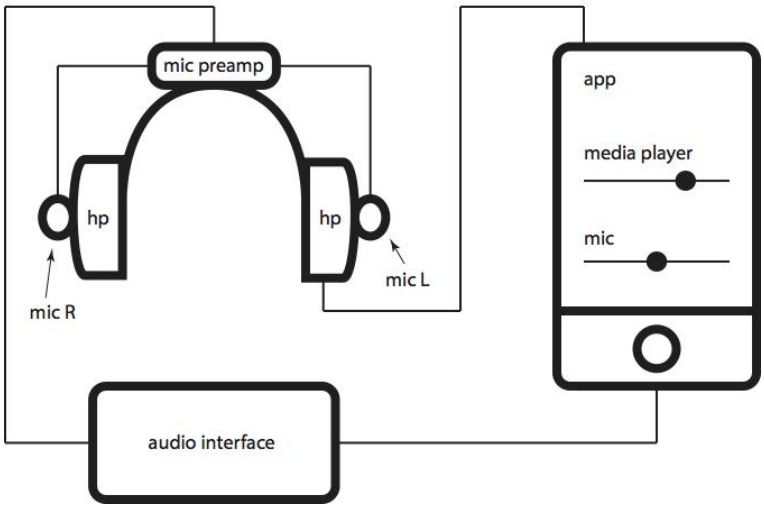


Figure 4.6.: Hardware connection diagram. Input: The stereo microphones are fixed to the microphone preamp, which is connected to the audio interface with a chinch cable. The audio interface is connected to the lightning connector of the mobile device. Output: The headphone is connected to the audio jack of the mobile device.

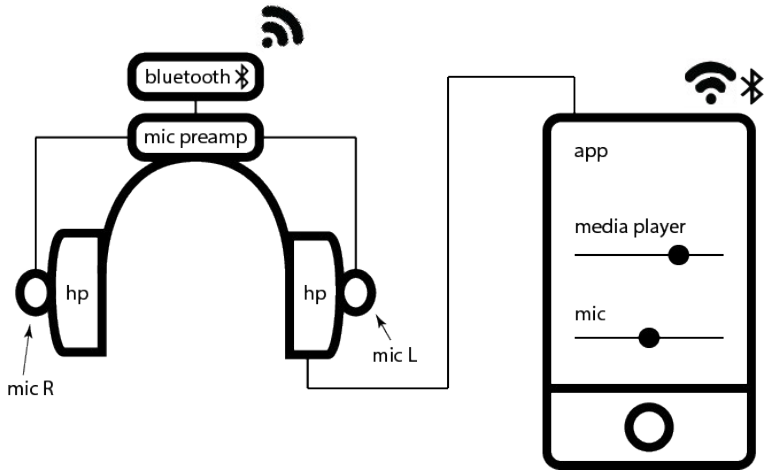


Figure 4.7.: Hardware connection diagram. Input: The stereo microphones are fixed to the microphone preamp, which is connected to the bluetooth streaming chip. The bluetooth streaming chip sends the microphone signals to the bluetooth receiver of the mobile device. Output: The headphone is connected to the audio jack of the mobile device.

4.1.3. Headtracker

Initially the demonstrator was conceptualised including dynamic binaural synthesis. Therefore, a headtracker was build to be physically connected to the headphones detecting the head orientation. The headtracker is based on the *9DOF razor IMU* [102] but is a combination of two independent components, the *Punch Through Design - LightBlue Bean* as processor and transmitter (bluetooth LE) and the *9DOF GY-85* motion board B. Early in the process it was found that the additional work was not applicable in the given time. The headtracker has to send the data on the head orientation with a high transmission rate to the smartphone which then processes the informations. This transmission has to be synchronised to ensure the correct processing according to the head orientation. To establish a well synchronised and stable communication of the interacting peripherals a significant amount of additional work time would have been necessary.

Further, the performance of the chosen test device seemed not to be sufficient for a realtime twelve channel binaural synthesis with direct/ambient component extraction. The amount of vector calculations is strongly increased (Sec. 2.3.2), as all calculations condensed in the otherwise static equations B, C, D and E would have to be processed for each time frame, which involves multiple vector convolutions for each channel.

Further, the time needed for the *openDAFF* database access gains higher importance and would have to be investigated.

In respect to the additional investigation and work time the headtracker integration was not completed.

4.2. Software concept

The application is conceptualised as a modular multi-channel audio player, with internal realtime filtering. All code were be written in native *Objective C* [3]. *C* should be used, when extremely performant programming is required (compensation, up-mixing, binaural auralization). The preferred frameworks are *CoreAudio*, *The AmazingAudioEngine*, *Accelerate Framework* and *ReactiveCocoa*.

CoreAudio is the native *Apple* audio framework. Through *CoreAudio* an application can receive data from the input or send data to the output of the device. Further, an application can interact with the operating system to process audio data, such as reproducing, filtering, mixing, etc..

The AmazingAudioEngine is an open source project to simplify high-level audio functions.

The *Accelerate Framework* is *Apples* native vector and matrix computation framework. Real as well as complex calculations are supported, which is advantageous when working with audio signals in the frequency domain.

ReactiveCocoa is an open source project implementing functional reactive programming. It can serve as a virtual wire between variables, objects or functions to achieve performant and responsive code.

The prototype of the up-mixing and auralization filter was implemented in *Mathworks - Matlab*.

Binaural synthesis involves a great number of calculations. When taking into account

[3] Objective C is the native programming language from *Apple* used for *iOS* or *MAC OSX* implementation.

dynamic binaural synthesis also the head position has to be tracked. Lindau [103] investigated thresholds for minimum detectable total system latencies depending on head movement velocity, head angle and stimulus type and found minimum thresholds between 50*ms* and 60*ms*. This shows that systems implemented for (dynamic) binaural synthesis have to be highly performant and responsive. The applicability of the proposed algorithm including binaural auralization is doubted using an *iPad Air* with the desired filter lengths.

4.2.1. Multi-channel audio player

A modular multi-channel audio player was implemented (Fig. 4.8). The player includes multiple channel objects, which are combined in the multi-channel audio player interface. The volume of each channel is individually controllable. To make a perceptual meaningful loudness control the linear generated values from the volume fader were used as input for the final volume curve function f_{vol}:

$$f_{vol}(x) = \begin{cases} x \cdot t^4/t & x < t = 0.55 \\ x^4 & x \geq t = 0.55 \end{cases} \tag{4.1}$$

A linear function results in minor changes in loudness perception in the two upper thirds of the interval [0,1], while fader movements in the first third result in big changes in loudness. That is because the sound pressure level and the perceived loudness are related logarithmically [1, p. 31]. An estimation for a perceptual balanced loudness curve in a dynamic range of 60dB is found to be $f_{vol}(x) = e^{6.908x}/1000$, which can be approximated as $f_{vol}(x) = x^4$ [104] (Fig. 4.9). The linear function below the threshold is applied by the author for a softer decrease of loudness in the still relevant range. As the player is to be used with headphones and the output of the smartphone is limited, the adjusted amplitudes are not very high and even the very quite amplitudes are audible.

According to their functionality the channel objects can be divided into two classes. The first class is the media channel object. It supports reproducing and filtering of audio files. The channel reads audio files from the hard drive of the smartphone (contained in the *iTunes* library) and provides the typical transport functions (play/pause, previous file, next file, seeking). Filtering can be toggled (on/off) and contains different up-

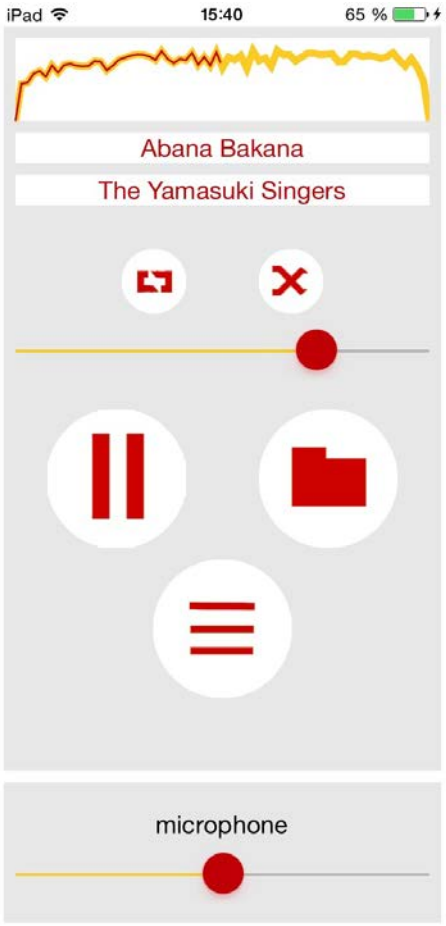


Figure 4.8.: The main view of the music player.

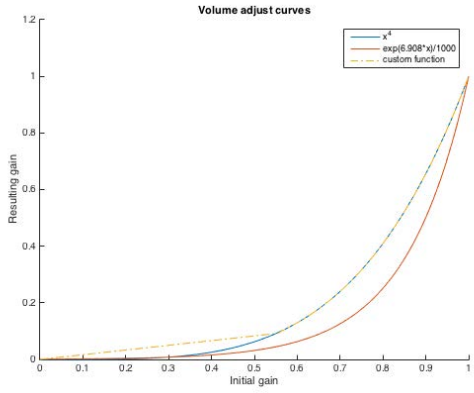


Figure 4.9.: The red curve follows the inverse of the estimated loudness sensation. The blue curve follows the approximation as simple potential function. The yellow curve follows the custom function applied by the author.

mixing setups (10.2 and 3.0) [4], binaural synthesis simulating two speakers at $\pm 30^{\circ}$ in a free-field and headphone compensation for different headphones.

The second channel class is the microphone-through channel object. The main input selected (microphone, audio interface, bluetooth headset, etc.) in the operation system preferences is routed to this channel. If the microphone is mono, it will automatically be routed to both channels of the headphone. If the connected microphone is stereo, the channels will be send to the corresponding headphone channel (L→L, R→R).

The channel object can also fulfil both functionalities (media/microphone-through). If desired, a switch can be applied to switch between these both functionalities.

4.2.2. Filtering

The filtering routine is divided into two sections, the up-mixing and binaural auralization filter and the compensation filter. Both filter types are applied to the signal in the frequency domain.

The audio frames are processed in the audio callback, before the frames are forwarded to the hardware output. Each audio frame contains 256 samples. In the audio callback function, first the signals are transferred into the frequency domain via FFT with 4096 bins to match the frequency resolution of the filters. The *Accelerate Framework* offers FFT functions that ignore the mirrored spectrum above the nyquist frequency for more efficiency (less calculations have to be made, less memory is needed). After the signal spectrum has been calculated the actual filtering is computed. Before the audio is finally send to the output, inverse FFT is used to convert the processed signal into the time domain. To avoid aliasing in the time domain, overlap-add is used. The frequency response contains 4096 bins (including the mirrored spectrum), therefore the inverse FFT results in a time signal of 4096 samples. Thus, 3840 samples have to be stored in the overlap buffer (ring buffer) and added to the signal of the following cycle.

When the application is initiated, all static terms of the algorithm (Eq. 2.30 - Eq. 2.33) are computed and saved for continuous usage.

[4] The 10.2 setups reproduces the ambient component of the signal. Thus, the impression is evoked, that the three speakers (FL, FC, LR) are positioned in a room of a defined characteristic. The 3.0 setup reproduces only the stereo image of the direct component, but mapped on three speakers. Thus, there is no impression of a room.

Up-mixing

The frequency response of the audio frame is analysed regarding its direct and ambient component as described in Sec. 2.3.1. Both components are represented by an independently weighted (gained) version of each channel of the input signal. The direct component is mapped on the extended stereo setup containing the centre channel. Therefore, the input dependent MVBNAP is computed as described in Sec. 2.3.2.

Further, the direct component is used to generate the early reflections from the side-walls and the ceiling. The early reflections arriving from the side-walls, are modulated using a phase delay and an attenuation factor. The ceiling reflections are de-correlated with a modelled quadratic residue diffuser as described in Sec. 2.3.2. When 10 channels are used as virtual channels (the subwoofers are not mentioned, because they do not have any influence on the room impression in this setup) and de-correlated early reflections (according to the explained considerations, Sec. 2.3.2) are added, the achieved IACC function correlates strongly with the theoretical IACC curve for binaural hearing based for perfect diffuse noise as proposed by Menzer and Faller [105] (Fig. 4.10). All loudspeaker representations (HRTFs) are loaded at the initialisation phase as static terms.

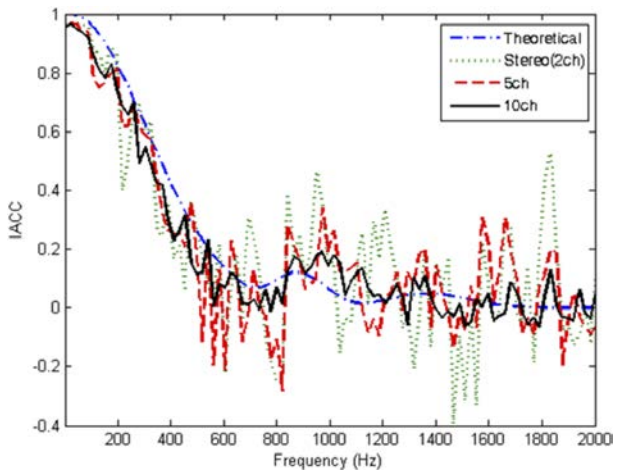


Figure 4.10.: The IACC curve plotted for 2, 5 and 10 channels. Further the theoretical IACC curve for binaural hearing is plotted [2].

The ambient component is distributed to every channel but the centre channel. All signal components representing the ambient signal, are contained in the pre-calculated static terms. The ambient signals are de-correlated via simple phase randomisation with a uniform distribution in the interval $[0, 1]$.

Binaural auralization

As base data for the binaural auralization a HRTF dataset is required. All loudspeaker positions are defined by the corresponding HRTF pair. As mentioned in Sec. 2.2 different HRTFs can be used. For the presented demonstrator the individual HRTF dataset of the author is used, which was recorded at *Technische Universität Berlin* using the technology presented by Fallahi [14] and Fuß [13].

A HRTF dataset can contain a huge number of entries, thus a database with efficient data handling algorithms is needed. The two open source projects addressing HRTF database handling *OpenDAFF* [106] and *SOFA* [107] were evaluated and considered for usage. Finally, *OpenDAFF* was found to be the best candidate for this application, as it can be integrated with minor changes into an *iOS* project. The *SOFA* project contains dependencies, which are not available for *iOS*. *OpenDAFF* is standardised, so that any HRTF dataset can be converted using the provided functions and scripts [5]. Any *OpenDAFF* HRTF dataset created with *OpenDAFF* file format version 1.7 can be used with this demonstrator. [6]

OpenDAFF offers the necessary functions for creating a dataset and searching in this dataset. The azimuth and elevation angle can be used for nearest neighbour retrieval. Nearest neighbour in this case refers to the source location in space and does not take into account the characteristics of the corresponding impulse response. The four corner points belonging to the rectangular framing spherical segment are returned. Each corner point, refers to one entry in the database.

Each virtual loudspeaker is represented by the HRTF pair corresponding to the predefined position. Thus, the superposition of all signals (loudspeakers) contains the direction information personalised for the author. The ETC, magnitude frequency response, ITDs and ILDs of the horizontal plane of the individual HRTF dataset is plotted in Fig. 4.11.

[5] The *OpenDAFF* project contains *Matlab* scripts for data conversion.

[6] The dataset for the presented demonstrator contains HRIRs not HRTFs. These are later transferred to the frequency domain.

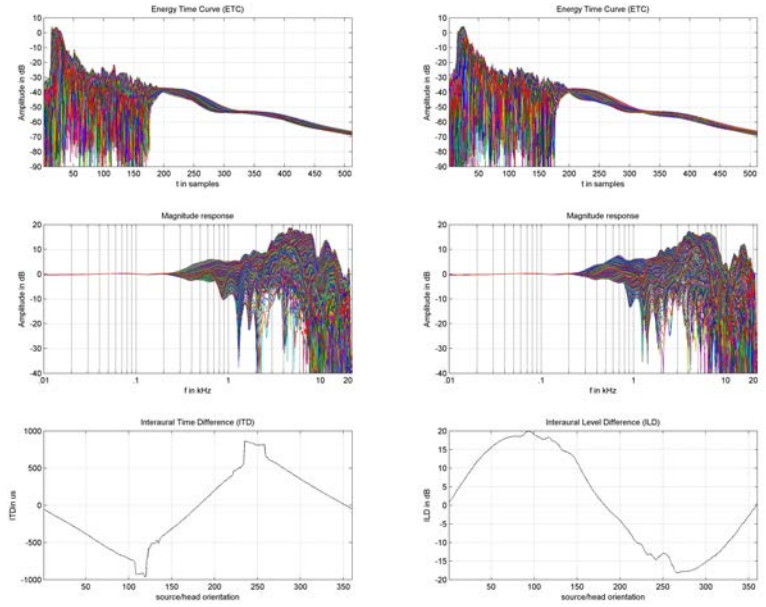


Figure 4.11.: ETC, magnitude frequency response, ITDs and ILDs of the horizontal plane of the individual HRTF dataset of the author. The ETC and magnitude frequency response are plotted per channel and for all directions on the horizontal plane with a resolution of 1°. ITDs and ILDs are plotted as functions of the angle.

Compensation

As part of this work HpTFs of several commercially available closed headphones (6 in-ear, 1 supra-aural, 1 circum-aural) were measured using the *Fast and Automatic Binaural Impulse response AcquisitioN* (FABIAN) HATS [11]. Every headphone measurement was repeated ten times. Between the measurements the headphones were repositioned to discover potential differences in the resulting frequency responses as mentioned before (Sec. 2.2.3). Four headphones were selected as examples for individual headphone compensation for the demonstrator. The corresponding headphone compensation filters can be chosen from the settings view of the application as described in Sec. 4.2.3.

The complete measurement report in German language is found in App. A.

The compensation filters were created using a *Matlab* script by Fabian Brinkmann, Alexander Lindau and Zora Schaerer (*Technische Universität Berlin*, audio communication group) using the methods presented by Brinkmann [56]. The adjusted values are shown in Tab. 4.1. The compensation filters were saved as wav file and later converted

via *Matlab* into one text file (.txt) per channel with every value of the impulse response in a new line (2048 values).

HP	N_f [f]	LS freq [Hz]	HS freq [f]
UE Zinken	100	100	4000
BD DT 770 Pro	100	100	4000
Creative	1000	300	7000
earPods	400	100	8000

Table 4.1.: The headphones (HP) presented are: UrbanEars Zinken, Beyerdynamic DT 770 Pro, Creative inEar (delivered with Creative Zen MP3 player), Apple earPods. The N_f is the normalisation frequency (normalisation factor: 1/3). LS stands for low shelve (filter) and HS stands for high shelve (filter). The gain for the shelving filters are 20 dB in all cases.

The filter length of the inverse frequency filter is set to $N = 2048$ (equals 4096 bins, but the application internally handles the mirrored spectrum). That implies a sufficient frequency resolution of $\Delta f = 10.77Hz$ while causing an adequate delay $t = \frac{N}{2} = 23.2ms$.

The compensation filter routine reads the impulse response of the compensation filter from text files stored in the body of the application. Thus, any desired filter can be added or existing filters can be adjusted. Every impulse response is saved in a separate file, while the both channels are also divided in separate files. The files contain one filter coefficient per line (2048 lines) and a blank line at the end.

In the case of up-mixing both filters (up-mixing and headphone compensation) are applied in the same filter callback handling routine for more efficiency.

In addition to headphone compensation the microphones connected to the microphone through channel of the application can be compensated using a reduced compensation filter contained in the application. A text file is used to store the impulse response of the compensation. The frequency response of each microphone and further the frequency dependent gain differences between the microphones can be compensated. Thus, when the microphones are replaced by other models, the correct impulse response can be added.To avoid a raising of the latency of the microphone channel, in the presented configuration microphone compensation is not added. Noticeable latencies on the microphone channel can be confusing, when the acoustic and visual event is not synchronous.

4.2.3. Settings

The graphical user interface (GUI) of the application provides a settings view to control all the adjustable functionalities of the application. The settings view is divided into two views, the main settings view and the surface characteristics view .

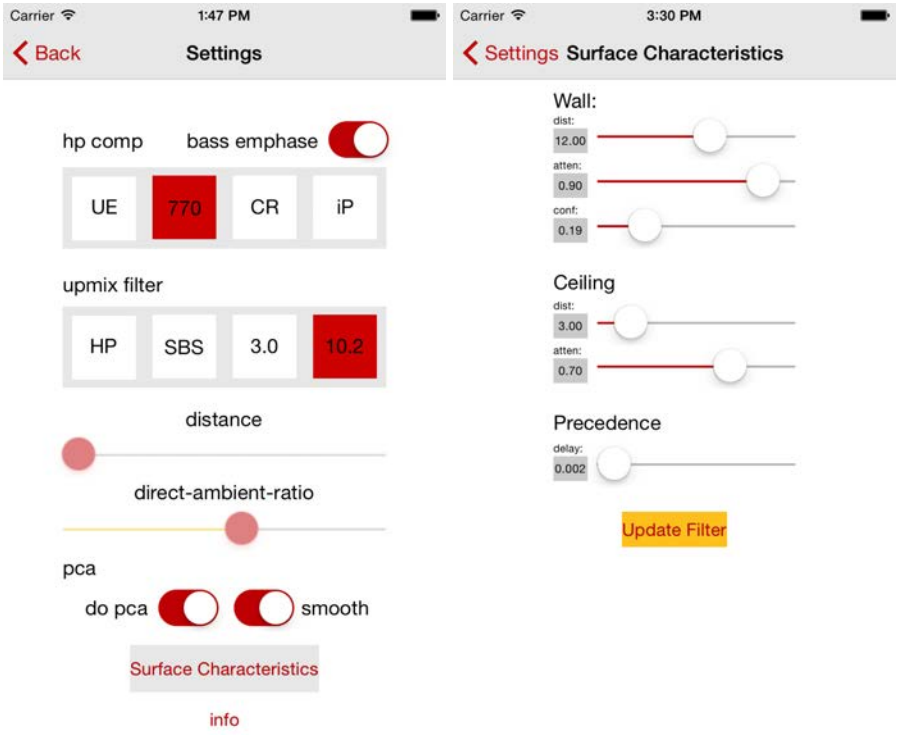


Figure 4.12.: The main settings view (left) and the surface characteristics view (right) of the GUI.

The headphone that is to be compensated, can be selected in the settings view of the GUI. All available headphones are shown in Tab. 4.2.

Name	Description
UE	UrbanEars Zinken
770	Beyerdynamic DT 770 Pro
CR	Creative inEar (Zen MP3)
iP	iPhone earPods (iPhone 6)

Table 4.2.: Selectable headphones for compensation.

Further the desired up-mixing filter can be chosen from within the settings view. All available filter types are listed in Tab. 4.3.

Name	Description
HP	Headphone (no filtering); The signal is reproduced as it is.
SBS	Stereo binaural synthesis, with a stereo loudspeaker set at $\pm 30^{\circ}$ in free-field
3.0	Extended stereo image up-mix.
10.2	USC 10.2 up-mix. The subwoofers (X.2) can be activated with the bass emphase switch.

Table 4.3.: Selectable filter types (no filter, stereo binaural synthesis, up-mixing).

Some of the parameters used during the up-mixing process can also be controlled from within the GUI of the application. These parameters are presented in Tab. 4.4.

Name	Description
τ_w	Side-wall early reflection delay; The delay is controlled via the distance between user and side-wall.
η_w	Attenuation of the side-walls.
β	This parameter controls the shifting of the left early reflection between side-wall and ceiling.
τ_c	Ceiling early reflection delay; The delay is controlled via the distance between user and ceiling.
η_c	Attenuation of the ceiling.
τ_p	Surround channel delay.

Table 4.4.: Controllable parameters from within the GUI for up-mixing behaviour.

If the up-mixing filter is selected (10.2 or 3.0), PCA can be toggled. If PCA is deactivated a fixed direct ambient ratio is adjustable via slider. Further a smoothing function for PCA can be toggled. The smoothing function is further explained in Sec. 5.2.1.

4.2.4. Bass boost

In order to enhance the sound a bass boost is applied by amplifying the subwoofer signals.

In USC 10.2 the subwoofers are fed with a superposition of centre and back surround signal. Further, a low-pass filter with a cut-off frequency of $120Hz$ is applied. To enhance the bass, subwoofer channels are fed with a mono down-mix of the resulting binaural signal and a low-pass filter with a cut-off frequency of $120Hz$. The low-pass is

realised as a 6^{th}-order Butterworth filter [108] and the subwoofer gain is increased by $3dB$ per channel.

4.3. Summary

The aim of this work was to implement an augmented acoustics demonstrator with realtime stereo up-mixing and binaural auralization with its hardware and software components. A mobile device running *iOS 8* was found to fulfil the discovered requirements. The processing power is sufficient high and external audio hardware can be connected. Further, the operating system supports low level audio processing. The headphones *Beyerdynamic DT 770 Pro* and *Apple earPlugs* were decided to be extended with microphones and used as hear-through headsets. The software was to provide a multi-channel audio player to mix virtually up-mixed audio files from the local storage with the binaural microphone signals.

5. Problems and developed solutions

During the implementation and construction phase hardware limitations were encountered, that led to modifications of the initial plans. Further, the proposed algorithm produced audible artefacts.

5.1. Hardware

A smart combination of the involved peripherals was a major concern while working on the demonstrator. The goal was to use as little wired connections as possible, to assure the maximum allowable mobility of the user. But exactly this consideration could not be fulfilled due to hardware and software limitations.

5.1.1. Connecting peripherals

When connecting the peripherals to the mobile device several observations were made. The audio input and output of *Apple* mobile devices are always coupled. This leads to three different cases for audio routing:

1. *Device hardware I/O*: build-in microphone or headset microphone; build-in speaker or headphone (1 IN / 2 OUT)

2. *HSP/HFP (bluetooth)*: connected headset as input and output (1 IN / 1 OUT)

3. *A2DP (bluetooth)*: output (no input available, 0 IN / 2 OUT) [1]

The bluetooth headset protocol (*HFP/HSP*) does only provide a mono input channel with limited bandwidth (narrowband $300Hz - 3.4kHz$, wideband $100Hz - 7kHz$). Two input channels are required for the developed application, as the input signal should be binaural. As audio input and output are always coupled in *iOS* it is not possible to use the four-contact jack to connect a headset (including a microphone) and additionally use the bluetooth protocol *HFP* to receive a second microphone signal. Further, again when the *HFP* protocol is connected only mono output is supported.

The widely supported *A2DP* cannot serve as audio sink on *Apple* mobile devices. Therefore audio streaming from a peripheral towards the mobile device is not possible up to now. Further, when streaming audio from an *Apple* mobile device towards a peripheral any audio input has to be disabled. This behaviour prevents the usage of *A2DP* for audio streaming with the developed application.

Sending raw audio data using a bluetooth data protocol (*L2CAP*) was intended to stream the microphone data to the smartphone, but the *CoreBluetooth* framework for *iOS* only supports *Bluetooth LE*, which has a very limited data rate. Therefore, a bluetooth data connection between a peripheral and an *Apple* mobile device is not sufficient for an intense data stream, such as high quality raw audio data stream.

5.2. Software

During the development and evaluation of the software potential audible artefacts were encountered when implementing the proposed algorithm. Further, due to processing limitations the used BRIRs are short in comparison to BRIRs used in systems running on desktop computers.

[1] For applications using audio the category of the AVAudioSession has to be defined to either *AVAudioSessionCategoryPlayAndRecord*, *AVAudioSessionCategoryPlayback* or *AVAudioSessionCategoryRecord*. When high quality audio streaming via bluetooth is required the category has to be set to *AVAudioSessionCategoryPlayback*, which disables the audio input.

5.2.1. DAR and panning gain smoothing

The PCA is calculated per frame and estimates the DAR and the panning gains by observing the correlation of the two channels. When multiple sources exist in the audio signal, like different instruments of a band, which are panned to different locations in the stereo image, the DAR and panning gains can vary strongly between the frames. If the calculated values are directly set as DAR and panning gains, the resulting signal can contain audible noise, depending on the composition of the input.

In the algorithm the PCA is calculated recursively considering the previous frame. This procedure results insufficient and audible noise is rendered throughout various music excerpts (not limited to a certain style of music).

To avoid audible noise emerge through inconsistencies, which was a quality feature mentioned in SAQI [15], a smoothing function is implemented in the application. The calculated values for DAR and panning gains are compared with the values for these very parameters of the previous audio frame. If a new value is bigger than the previous one, the previous value is increased by 0.004 and set as new value. A step width of 0.004 was found by the author to be a reasonable value for most signals.

The step width could also be adjusted adaptively depending on the absolute difference between the new and previous value. This might lead to a more dynamic signal, even though no detoriations could be found by the author by using a static step width and thus, linear smoothing curve.

5.2.2. BRIRs

The BRIRs created in this application are relatively short. They consist of 4096 samples, which is equivalent to a length of $0.09s$ at the sampling frequency of $44.1kHz$. A typical reverberation time for a concert hall is $2s$. Thus, the constructed BRIRs are not able to simulate a typical reverberation time.

The de-correlation of the early reflections and additional ambient signals rendered by the application produce an impression of wideness. If a realistic reverberation time is desired, this can be achieved by introducing a reverberation effect as proposed by Borß

[84].

5.3. Summary

During the implementation and construction process hardware and software limitations led to the following decisions. Neither the initially planed wireless connection between the two microphones and the mobile device nor a bidirectional audio streaming (microphone/headphone) could be established, due to the audio routing policy introduced by *iOS*. Therefore, a wired connection involving additional audio hardware was realised. The proposed up-mixing algorithm had to be extended by a smoothing function to avoid audible noise, caused by strongly varying gain values and direct respectively ambient factors. The usage of an reverberation effect for realising a realistic reverberation time is suggested to be evaluated in further research.

With raising computation power of smartphones dynamic binaural synthesis with multiple channels including long BRIRs and an external headtracker is assumed be feasible in the next years.

6. Evaluation

After realisation the demonstrator was evaluated by the author based on technical features (processing time, resource consumption, modelling plausibility) and the quality of the rendering algorithm regarding the resulting sound (subjective room impression, envelopment, audible artefacts). In addition the hear-through headset was evaluated informally regarding a listeners ability to localising real audio sources.

6.1. Technical evaluation

To evaluate the application based on the computing performance and the frequency characteristic of the virtual room was analysed and rated.

6.1.1. Performance

The application is a complex construct of multiple objects and routines. All routines and the visual rendering can be classified regarding the central processing unit (CPU) workload and the required virtual memory. The development environment *Xcode* contains a view, which visualises the application workload. Thus, processes and threads can be analysed and rated. An exemplary average output of *Xcode* for the realised application running on an *iPad Air* is shown in Fig. 6.1 and Fig. 6.2.

At the start of the application the visual components and all global variables are initialised ($\sim 3s$ on *iPad Air*). After the initialisation phase of the application (peak 10% CPU workload) the average CPU workload is 1%. When the first audio file is loaded the up-mixing filter is initialised simultaneously. Thus, the high peak (29%) represents the workload for both routines, file access and filter initialisation. When the audio file is

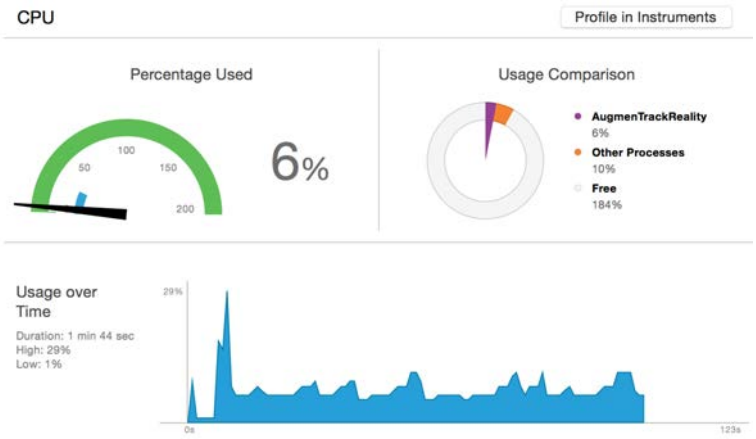


Figure 6.1.: Illustration of Xcode output on CPU workload of the profiled application.

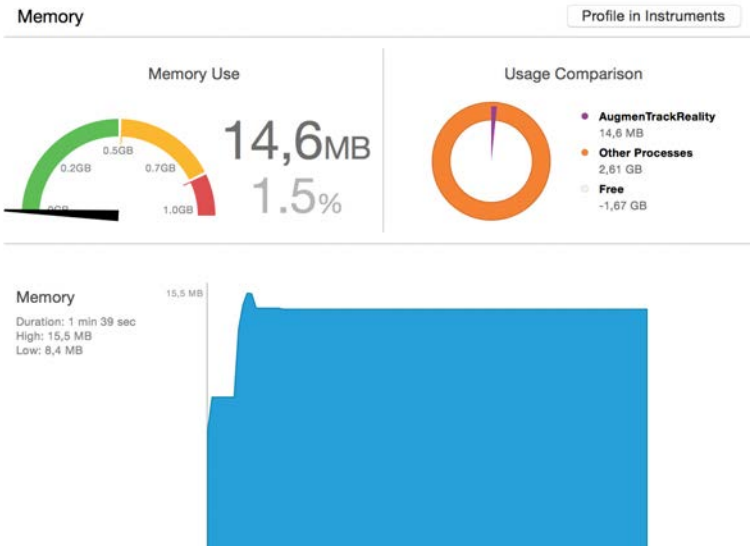


Figure 6.2.: Illustration of Xcode output on memory workload of the profiled application.

rendered the average CPU workload is in the range $5-9\%$. The memory consumption of the applications on average is $11-16MB$ RAM.

The application size is $4.2MB$, containing all dependencies, filtering data, UI images, etc.

Important for real-time applications is the processing time of involved routines. In this application the entire filtering is computed in the audio callback. Each audio frame has 256 samples. Using a sampling frequency of $44.1kHz$ one audio frame has a length of $5.8ms$. The time needed for the filtering routine must be shorter than the time length of an audio frame, to assure a continuous rendering without any dropouts. The time needed for the filtering routine is measured using C timestamps. The measuring revealed an average computing time of the involved filtering routine of $0.5ms$. The calculation of the static terms in the initialisation phase of the filter is measured to be about $12ms$.

All the presented total values can differ depending on the device used to run my application and other applications running simultaneously.

6.1.2. Frequency characteristic of the virtual room

To analyse the frequency characteristics of the virtual room the BRTF of the virtual room was plotted. A $1s$ long sweep with normalised amplitude and 44100 samples covering the frequency range $0Hz-22050Hz$ was used as input signal. Because the sweep is monophonic the PCA estimates the whole signal as being direct component. To analyse the frequency characteristics of the virtual room at different DARs, the DAR was set manually. Three different values were set for γ_S (direct component factor) and γ_N (ambient component factor) to get an overview of the changing room characteristic. The relation between γ_S and γ_N was set to: $\gamma_N = 1-\gamma_S$. The BRTFs for $\gamma_S = 1$, $\gamma_S = 0.5$ and $\gamma_S = 0$ are plotted in Fig. 6.3.

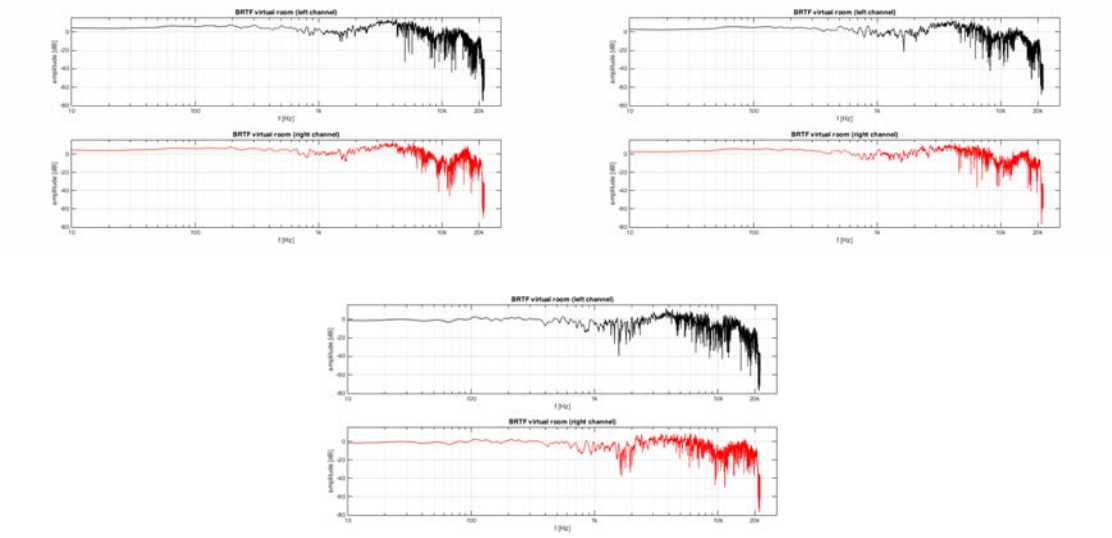


Figure 6.3.: Upper left: BRTFs for $\gamma_S = 1$ and $\gamma_N = 0$. Upper right: BRTFs for $\gamma_S = 0.5$ and $\gamma_N = 0.5$. Lower center: BRTFs for $\gamma_S = 0$ and $\gamma_N = 1$.

6.1.3. PCA algorithm

The PCA algorithm estimates the DAR and the panning gains. Several test signals were used for evaluating the estimation of the panning gains. The signals were either sinus signals or white noise, with a panning angle $-30° \leq \Phi \leq 30°$. The estimated panning gains, the resulting angle and the estimation error can be found in Tab. 6.1.

Further, if two sources which are panned to the two extremes of the stereo image are played right after each other, the smoothing algorithm (Sec. 5.2.1) will lead to a continuous shifting of the second source from the position of the first source to its correct position.

6.1.4. Discussion

The technical evaluation of the demonstrator showed that the software is performant and running stable on the chosen mobile device (*iPad Air*) with an average CPU workload of 7% and average memory usage of 1.5%. Due to the relatively low CPU consumption the battery service life is appropriate. The mean processing time for the filtering is $0.5ms$, which is very low. The initialisation time of the static terms is relatively high, being $12ms$. This initialisation time of the filter does not contain the time needed for

$Signal_{Angle}$	$gain_L$	$gain_R$	Res_{Angle}	Err_{Angle}
S_0	0.707	0.711	$-0.14°$	0.14
S_{-6}	0.623	0.779	$-5.45°$	-0.55
S_{-15}	0.45	0.89	$-14.81°$	-0.19
S_{18}	0.92	0.37	$18.22°$	-0.22
S_{30}	1.0	0.001	$29.98°$	0.02
N_0	0.707	0.711	$0.14°$	0.14
N_{12}	0.991	0.078	$27.97°$	-15.97
N_{-15}	0.08	0.995	$-27.93°$	12.93
N_{-21}	0.04	0.999	$-28.99°$	7.99
N_{30}	1.0	0.002	$29.95°$	0.05

Table 6.1.: In this table the panning gains estimated by the PCA are listed and the complemented by the resulting angle and the error angle. The signals are named with an abbreviation for the signal type and the source angle (left:positive, right:negative) of the signal in degrees. S stands for sinus (1000Hz), N stands for white noise ($20Hz - 20kHz$), SW stands for sweep ($10s$, $20Hz - 20kHz$). $gain_{L/R}$ are the estimated panning gains for the left and right channel. The Res_{Angle} is the angle calculated from the estimated panning gains. Error is the error between estimated angle and real angle in degrees.

the database access, in case a new HRTF has to be loaded from the *OpenDAFF* set. The database access add another $2ms$, when reading twelve channels. This confirms the initial assumption that dynamic binaural synthesis in realtime involving twelve channels is not possible on the chosen device, as the audio frame is $5.8ms$ long. Additional to the raw filtering process the application has to render interface changes, handles data access (read, write, copy, delete, etc.) and unser interactions (hard-/software interrupts). Further, the operation system consumes processing resources.

After visually inspecting the BRTFs plots, the plots vary between BRTFs with high frontal direct sound emission when setting a high direct sound factor and a theoretic diffuse BRTF for a high diffuse sound factor. Investigating the BRIR spectra it can be seen that, the BRTFs show a higher roughness when decreasing γ_S and increasing γ_N. This is caused by the larger amount of diffuse sound and the smaller amount of direct sound. As the diffuse sound is generated by superimposing and de-correlating the ten channels containing direct, reflection and surround signals (the subwoofers are not considered), the rendered signal magnitude spectrum results more corrugated. The third plot shows the measurement of the BRTF in the diffuse-field without the direct sound and early reflections.

It can be noticed, that the lower frequencies are not modulated although a mode pattern for small rooms would be expected. This is because the low frequencies were extrapolated from the recorded HRTFs. A flat response for the low frequencies lead to a more pleasant sound. Further, it can be seen that the frequencies above $15kHz$ (left channel) and $18kHz$ (right channel) decrease. This is caused by the quality of the HRTFs that belong to the directions defined by the loudspeaker positions (Fig. 4.11).

Fig. 6.4 shows the BRTFs of the opera house in Sydney/Australia [109] and the Promenadikeskus concert hall in Pori/Finland [110]. It can be seen that both spectra are more fuzzy. This is caused by the reverberation time in both halls, which is not modelled in the implemented application. Concert halls usually have a reverberation time of approximately $2s$, the Sydney opera house has a reverberation time of $1.4s$ ($100Hz - 8kHz$) [111]. The virtual room introduced by the presented algorithm does not have such a long reverberation time (see Sec. 5.2.2). Through the longer reverberation time even more reflections are superimposed. Thus, the BRTF is further modulated.

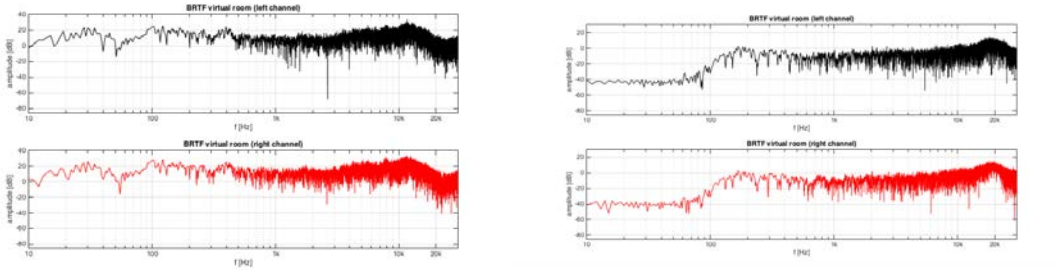


Figure 6.4.: Left: BRTF of the opera house in Sydney/Australia (subject A). Right: BRTF of the Promenadikeskus concert hall in Pori/Finland (binaural, subject 1, position 1)

The estimated source angle matches very good with the real angles for sinus signals. The small estimation errors can be explained by rounding errors and further inaccuracies due to limited bit resolution or potential type conversions in the functions of the used frameworks. As mentioned in Sec. 2.1.1 the minimum human localisation blur was reported to be $0.75°$ in the horizontal plane. Thus, the error should be inaudible.

The angle of the noise source is overestimated when the signal is panned $0° < \Phi < 30°$. The signal panning is achieved by setting the gain of each stereo channel and the inter-channel delay. The panning gains are calculated through an inter-channel correlation. As the white noise signals have a random phase, the virtual source angle is overestimated towards the sides. This leads to the assumption that the presented algorithm is insufficient for the angle detection of panned noisy direct sources. Further,

it has to be considered, that the uncorrelated parts of the signal (such as panned noise) is assumed to be the ambient component.

The application is assumed mainly to be used for reproducing music or speech, which can contain both types of signals noise and superpositions of harmonics. Listening to several music and speech signals the author found that a virtual stage comparable to the one generated by the native stereo loudspeaker assembly can be determined.

6.2. Perceptual evaluation

As the demonstrator is implemented for audio reproduction the perceptual evaluation is of major interest. Because of the limited time for this work and the focus on the technical feasibility and realisation of a fully working demonstrator the perceptual evaluation is performed as a selftest of the author and an informal listening test with six subjects.

6.2.1. Listening test

The six subjects were asked to listen to self chosen songs from a huge playlist containing mostly popular music, rock and jazz. The provided headphone was the *Beyerdynamic DT 770 Pro*. After a while of listening and trying out the different filter types (10.2, 3.0, 2.0 and no filter), the subjects were asked about their experience with the demonstrator and the impression of the virtual listening space.

The main results of this informal survey - while not having evolved from a formal textual analysis - are roughly summarised here: The perceived degree of room impression was rated high for music, when comparing the 10.2 up-mixing filter to the unfiltered stereo sound. The subjects reported that the sound gained wideness. The sound was described as clear and the virtual listening space was assessed to be a large space even though the reverberation time is short. The 10.2 filter was found to be the best applicable filter.

Front/back confusions were not reported.

When the subject mixed the surrounding sound with the audio content reproduced in the application, the impression of externalisation emerged.

6.2.2. Acoustic localisation

In the user test with six subjects, two microphone positions for the *Beyerdynamic DT 770 Pro* and one microphone position for the *iPhone earPods* were informally investigated regarding the ability of localising real sound objects using the hear-through headset only through the acoustic event (closed eyes). The subjects were seated on a chair in a living room of $21m^2$ equipped with two medium sized couches, a wooden dining table with four wooden chairs, two big windows and thin curtains. The walls were made of plastered bricks, the ceiling was made of plasterboard and the ground was covered with laminate. The chair was placed in the centre of the room. The reverberation time of the room was estimated to be around $1.5s$.

Every subject was wearing a sleeping mask and placed the hear-though headset on her head. The headphone jack was directly connected to the microphone preamp (the mobile device application was not used). The author moved a loudspeaker in the room and the subject had to point at the estimated position of the loudspeaker. The test signal was a male speaker talking continuously. The experiment was repeated ten times per subject, microphone position and headphone, while each repetition represented one source direction. This resulted in 30 responses per subject.

The microphone positions when using the *Beyerdynamic DT 770 Pro* were as described in Sec. 4.1.2 (configuration 1 / configuration 2). When the *Apple earPods* were used, the microphones were placed on a hairband about $3cm$ in front of the ears (configuration 3).

It was found that for configuration 1, the sources were thought to be positioned more towards the sides ($\pm 90^\circ$) as they actually wore, for azimuth angles between $\pm 45^\circ - 135^\circ$. For configuration 2 the localisation was adequate. If the source was placed behind the subject, it was described to be more muffled than when it was perceived from the front.

The localisation and the sound when using the *Apple earPods* and the microphones close to the ears was found adequate.

Further, in a final self-test of the author the source was moved around the subject to examine the perceived velocity of the source depending on the azimuth angle. In the case the head diameter was increased (configuration 1) the movement of the source on the sides appeared faster towards the positions at $\pm 90°$.

6.2.3. Discussion

The informal perceptual evaluation of the application shows the tendency that a high level of room impression and sound envelopment is achieved using the presented audio enhancement compared to the unfiltered stereo sound. The 10.2 up-mixing filter was found to be the best of the given filters, as it evoked a strong impression of envelopment and a well perceived integration of the listener into the virtual listening space, which confirms the initial assumption.

The timbre of the sound when using the presented application is mainly influenced by the choice of HRTFs and the quality of the headphone compensation filter. The combination of the personalised HRTFs and the appropriate headphone compensation for the used headphone leads to a clear acoustic and a well defined virtual stage. Thus, individual HRTFs and (non-individual) headphone compensation lead to a good listening experience.

The additional usage of the microphones for mixing the surrounding acoustic environment with the reproduced audio, led to the impression of externalisation. It is assumed that this happens, because it seems somehow plausible for a user, that the audio signal is coming from somewhere in the environment, if the acoustic environment was mixed in. The reproduced audio is integrated into the environment to some extent, instead of creating a completely new and isolated environment, defined by the characteristics of the recording space and a virtual listening space only.

When using the proposed hear-through headset, the acoustic localisation was found to be adequate. The subjects could identify the source position without seeing the corresponding visual events with a maximum error of approximately $\pm 20°$ (near $\pm 45°$ and $\pm 135°$), when the microphones were mounted on the enclosure of the relatively big *Beyerdynamic DT 770 Pro* facing to the sides, as anticipated. The localisation error can be explained through the virtual enlargement of the head diameter and the resulting larger ITDs. The localisation error was smaller, when the microphones were facing to the

front, as the virtual head diameter was only slightly enlarged. For the second configuration the sound coming from behind the subject was perceived to be more muffled. The more muffled sound is caused through acoustic shadowing. If the microphones were fixed to a hairband and positioned next to the ear, no colouration through acoustic shadowing or unrealistic ITDs appeared. The second and third configuration were found to be the best solutions, as the localisation results were the best. Further, an acoustic shadowing is normal for sources sounding from behind the listener. The resulting low-pass filter is much more pronounced compared to the soft low-pass introduced by the pinna when a sound source is placed behind a listener, but the stronger attenuation of the high frequencies in this case even seem to underline the differences in timbre typical for the source - receiver relation.

As mentioned in Sec. 2.4.1, people who have to rely on hearing aids learn to additionally use low frequency direction cues to compensate the missing high frequency direction cues, which are missing due to unnatural virtual ear canal positions (e.g. behind the ear). Binaural headsets are still not common, but are assumed to gain more attention throughout the next years. While a technology is often used to support a users abilities, a user also improves her skills in using this very technology and the possible advantages. Thus, an average user might be able to improve her interpretation capabilities of low frequency direction cues when using a binaural headsets and thus, enhanced her localisation capability of real surrounding sound objects.

6.3. Summary

From informal interrogation of six subjects the room impression introduced by the upmixing filter was assessed as good and described as plausible. Further, an envelopment and an integration of the listener into the virtual listening space was reported. The analysis of the frequency response of the virtual listening space led to the conclusion that the frequency characteristics are plausible and show similarities to frequency responses of real concert halls. The angle of the virtual source was correctly estimated for sinus signals, but strongly overestimated for noise signals (if panning is gained through inter-channel delay).

The hear-through headset deteriorates the ability of acoustic localisation to some extend, but the results were still adequate. As in natural situations a visual feedback is

often present, the quality of the headset is assumed adequate.

7. Conclusion

When stereo audio files are played back via headphones the listener perceives the audio content originating inside her head. Using HRTFs to filter stereo audio signals adds direction information of virtual stereo loudspeakers placed somewhere around the listener in a free-field. If the virtual loudspeaker setup is extended additional sound components typical for the perceived sound in rooms can be added, such as early reflections and diffuse sound. This enhancements of audio can lead to an adequate room impression and envelopment. Further, it was assumed that mixing the up-mixed and binaural auralized audio with the surrounding acoustic environment can lead to the impression of externalisation of the perceived audio signal.

For stationary computers several solutions regarding virtual acoustic (up-mixing, binaural auralization, room simulation, etc.) are available and yield good results. Mobile devices are a daily companion and constantly improved with more performant processing units. To utilize these developments in future, an application for up-mixing and binaural auralization was implemented for *iOS* and an hear-through headset was constructed to investigate the feasibility of a mobile solution of the listed features.

7.1. Summary

Since an processing intensive mobile application addressing audio signal processing was planned, the existing hardware and software had to be reviewed to find an adequate mobile device. This device also had to have the ability of connecting external audio hardware to it. In summary it was decided to use an actual *Apple* mobile device running *iOS 8* and native *Objective C* and *C* for programming, which fulfilled all requirements.

An multi-channel audio player with different types of channels was implemented. The audio player can access the local audio file library and reproduce any file type, that is supported by the operating system. A multi-channel input is supported, too. It serves as a stereo microphone through channel. The audio, that is played back from the local storage, can be filtered in realtime in three different ways. The first filter is a stereo binaural filter using two HRTF pairs for the discrete positions of the virtual loudspeakers in a free-field. The second filter is a 3.0 up-mixing filter, which enhances the stereo image by a virtual centre speaker. The third filter is a 10.2 up-mixing filter, which adds additional channels are added propagating early reflections and diffuse sound.

For hear-through purposes a microphone was mounted on either side of the enclosure of the consumer headphone *Beyerdynamic DT 770 Pro*. For a second configuration the microphones were applied to a hairband to both ears and used *Apple earPods* for listening. The microphones were connected via microphone preamp and audio interface to the mobile device.

The performance of the application and the workload of the mobile device were tested and evaluated as part of this work. Both, the processing time and the workload were very low. The BRTFs of the virtual room were analysed and compared to BRTFs of real concert halls. The BRTFs generated by the application were evaluated as physically plausible.

An informal perceptual evaluation was performed with six subjects, testing the overall experience with a focus on envelopment and room impression and the localisation ability, while the hear-through headset was worn. The listeners rated the 10.2 up-mixing filter to be the best. It led to the best perception of a pleasant virtual listening space.

The usage of individual HRTFs and (non-individual) HpTFs for headphone compensation was found by the author to lead to a well defined virtual stage.

7.2. Perspectives

The use of dynamic binaural synthesis is know to improve the results of virtual acoustic solutions. As it was found that dynamic binaural synthesis using twelve (or ten) channels is not applicable with the present state of mobile technology when using the

presented algorithm, it should be investigated how many channels can be dynamically rendered in realtime. To reduce the number of channels, diffuse sound could be added using light-weight reverberation models instead of discrete channels.

To allow dynamic binaural synthesis the setup has to be extended by a headtracker. A simple do-it-yourself headtracker was built and implemented by the author based on the code of the *9DOF razor IMU* [102]. As hardware a *Punch Through Design - LightBlue Bean* and a nine-degrees-of-freedom sensor board (various models are available, e.g. the board presented in App. B) was combined. The *LightBlue Bean* already processes the data from the sensor board and sends it to the presented application. The data containing the roll, tilt and pan angle, can be used by the application to select the correct HRTF pairs.

With upcoming operation system updates it might be possible to stream high quality multi-channel audio from a peripheral to an *Apple* mobile device. Having the option doing this, would make the additional audiointerface obsolete and a wired connection could be avoided and replaced by a wireless solution.

As the application and the headset are ready for various augmented reality solutions, the application could be extended by the ability of rendering internal sources from any direction, to apply artificial soundscapes or additional acoustic information.

Therefore the future perspective might be a smartphone which allows listeners wireless high quality spatial audio consumption in combination with environmental awareness to avoid acoustic seclusion.

8. Bibliography

[1] Weinzierl, Prof. Dr. Stefan (Ed.) (2008): *Handbuch der Audiotechnik*. 1. Berlin Heidelberg: Springer-Verlag.

[2] Lee, Taegyu; Yonghyun Baek; Young-cheol Park; and Dae Hee Youn (2014): "Stereo Upmix-based Binaural Auralization for Mobile Devices." In: *IEEE Transactions on Consumer Electronics*, **60**(3), pp. 411–419.

[3] Jones, Douglas L. and Ivan Selesnick (2010): *The DFT, FFT, and Practical Spectral Analysis*. Houston, Texas: Rice University.

[4] Hosokawa, Shuhei (1984): "The Walkman Effect." In: *Popular Music - Performers and Audiences*, vol. 4. Cambridge, UK: Cambridge University Press, pp. 165–180.

[5] Sony (1979): "TPS-L2: Stereo-Cassettenplayer."

[6] Pras, Amandine; Rachel Zimmerman; Daniel Levitin; and Catherine Guastavino (2009): "Subjective Evaluation of MP3 Compression for Different Musical Genres." In: *127th AES Convention - Convention Paper 7879*. New York, USA.

[7] AES (2014): "AES Technical Committee. Audio for Telecommunications." Online. URL `http://www.aes.org/technical/at/`.

[8] Lepa, Steffen (2014): "Survey Musik und Medien. Genutzte Audiogeräte 2012." URL `http://musikundmedien.ak.tu-berlin.de/survey-2012/audionutzungsdaten/audiogeraete/`.

[9] Stecker, G. Christopher and Frederick Gallun (2012): "Binaural Hearing, Sound Localization, and Spatial Hearing." In: Kelly L. Tremblay and Robert F. Burkard (Eds.) *Translational Perspectives in Auditory Neuroscience. Normal Aspects of Hearing*. San Diego, USA: Plural Publishing Inc.

[10] Vorländer, Michael (2008): *Auralization. Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. 1. Berlin Heidelberg: Springer-Verlag.

[11] Lindau, Alexander and Stefan Weinzierl (2006): "FABIAN - An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom." In: *24. Tonmeistertagung - VDT International Convention*. Leipzig, Germany.

[12] Brinkmann, Fabian (2011): *Individual headphone compensation for binaural synthesis*. Master's thesis, Technische Universität Berlin, Berlin, Germany.

[13] Fuß, Alexander (2014): *Entwicklung eines vollsphärischen Multikanalmesssystems zur Erfassung individueller kopfbezogener Übertragungsfunktionen*. Master's thesis, Technische Universität Berlin, Berlin, Germany.

[14] Fallahi, Mina (2014): *A system for the fast measurement of individual head-related transfer functions. Simulation and implementation of measurement algorithms*. Master's thesis, Technische Universität, Berlin.

[15] Lindau, Alexander; et al. (2014): "A Spatial Audio Quality Inventory (SAQI)." In: *Acta Acustica united with Acustica*, **100**(5), pp. 984–994(11).

[16] Blauert, Jens (1997): *Spatial Hearing. The psychophysics of human sound localization*. 2. Massachusetts, USA: MIT Press.

[17] Lu, Yan-Chen (2010): *Active Hearing Strategies for Binaural Sound Localisation in Azimuth and Distance by Mobile Listeners*. Ph.D. thesis, University of Sheffield.

[18] Keyrouz, Fakheredine and Klaus Diepold (2006): "A Rational HRTF Interpolation Approach for Fast Synthesis of Moving Sound." In: *IEEE 12th Digital Signal Processing Workshop*. Teton National Park, USA, pp. 222–226.

[19] Zahorik, Pavel; Douglas S. Brungart; and Adelbert W. Bronkhorst (2005): "Auditory Distance Perception in Humans: A Summary of Past and Present Research." In: *Acta Acustica united with Acustica*, **91**(3), pp. 409–420.

[20] Nielsen, Søren H. (1992): "Auditory Distance Perception in Different Rooms." In: *92nd AES Convention*. Vienna, Austria.

[21] Zahorik, Pavel (2002): "Assessing auditory distance perception using virtual acoustics." In: *Journal of the Acoustical Society of America*, **111**(4), pp. 1832–1846.

[22] Harris, Cyril M. (1966): "Absorption of Sound in Air versus Humidity and Temperature." In: *Journal of the Acoustical Society of America*, **40**(1), pp. 148–159.

[23] DIN (1966): *ISO 9613-2. Dämpfung des Schalls bei der Ausbreitung im Freien*. Deutsches Institut für Normung e.V.

[24] Algazi, V. Ralph and Richard O. Duda (2011): "Headphone-Based Spartial Sound." In: *IEEE Signal Processing Magazine*, **33**, pp. 33–42.

[25] Davis, Don (1979): "The Role of the Initial Time Delay Gap in the Acoustic Design of Control Rooms for Recording or Reinforcement Systems." In: *64th AES Convention*. New York, USA.

[26] Sakamoto, Naraja; Toshiyuki Gotoh; and Yoichi Kimura (1976): "On -Out-of-Head Localization- in Headphone Listening." In: *Journal of the Acoustical Society of America*, **24**(9), pp. 710–716.

[27] Møller, Henrik; Michael Friis Sørensen; Clemens Boje Jensen; and Dorte Hammershøi (1996): "Binaural technique: Do we need individual recordings?" In: *J. Audio Eng. Soc.*, **44**(6), pp. 451–469.

[28] Møller, Henrik; Clemen Boje Jensen; Dorte Hammershøi; and Michael Friis Sørensen (1996a): "Using a Typical Human Subject for Binaural Recording." In: *100th AES Convention*. Copenhagen, Denmark.

[29] Zhang, Mengqiu; Wen Zhang; Rodney A. Kennedy; and Thushara D. Abhayapala (2009): "HRTF measurement on KEMAR manikin." In: *Aoustics*. Adelaide, Australia.

[30] Majdak, Piotr; Peter Balazs; and Bernhard Laback (2007): "Multiple Exponential Sweep Method for Fast Measurement of Head-Related Transfer Functions." In: *Journal of the Audio Engineering Society*, **55**(7/8), pp. 623–637.

[31] Dietrich, Pascal; Bruno Masiero; and Michael Vorländer (2013): "On the Optimization of the Multiple Exponential Sweep Method." In: *Journal of the Audio Engineering Society*, **61**(3), pp. 113–124.

[32] Møller, H. (1992): "Fundamentals of binaural technology." In: *Applied Acoustics*, **36**, pp. 171–218.

[33] Möser, Michael (2009): *Technische Akustik*. 8. Berlin Heidelberg: Springer-Verlag.

[34] (2015): URL `http://www.catt.se/`.

[35] (2015): URL `http://ease.afmg.eu/`.

[36] Schröder, Dirk and Michael Vorländer (2011): "RAVEN: A Real-Time Framework for the Auralization of Interactive Virtual Environments." In: *Forum Acusticum*, pp. 1541–1546.

[37] Schroeder, Dirk and Tobias Lentz (2006): "Real-Time Processing of Image Sources Using Binary Space Partitioning." In: *Journal of the Audio Engineering*

Society, **54**(7/8), pp. 604–619.

[38] Kolowski, Andrzej (1985): "Algorithmic Representationof the Ray Tracing Technique." In: *Applied Acoustics*, **18**, pp. 449–469.

[39] Schroeder, Dirk; Phillip Dross; and Michael Vorländer (2007): "A Fast Reverberation Estimator for Virtual Environments." In: *30th International AES Conference*. Saariselkä, Finland.

[40] Sandvad, Jesper and Dorte Hammershøi (1994): "Binaural auralization: comparison of FIR and IIR filter representation of HIRs." In: *96th AES Convention*. Amsterdam, Netherlands.

[41] Schroeder, Dirk; et al. (2010): "Virtual Reality System at RWTH Aachen University." In: *International Symposium on Room Acoustics ISRA*. Melbourne, Australia.

[42] Gardner, William G. (1995): "Efficient Convolution without Input-Output Delay." In: *Journal of the Audio Engineering Society*, **43**(3), pp. 127–136.

[43] García, Guillermo (2002): "Optimal Filter Partition for Efficient Convolution with Short Input/Output Delay." In: *113th AES Convention*. Los Angeles, USA.

[44] Hurchalla, Jeffrey R. (2010): "A Time Distributed FFT for Efficient Low Latency Convolution." In: *129th AES Convention*. San Francisco, USA.

[45] Wefers, Frank and Michael Vorländer (2011): "Optimal filter partitions for real-time FIR filtering using uniformly-partitioned FFT-based convolution in the frequency-domain." In: *14th International Conference on Digital Audio Effects*. Paris, France, pp. 155–161.

[46] Lokki, Tapio; et al. (2004): "Application Scenarios of Wearable and Mobile Augmented Reality Audio." In: *116th AES Convention, Convention Paper 6026*. Berlin, Germany.

[47] Pernaux, Jean-Marie; Marc Emerit; Jerome Daniel; and Rozenn Nicol (2002): "Perceptual Evaluation of Static Binaural Sound Synthesis." In: *22nd International AES Conference on Virtual, Synthetic and Entertainment Audio*. Espoo, Finnland.

[48] Sandvad, Jesper (1996): "Dynamic Aspects of Auditory Virtual Environments." In: *100th AES Convention*. Copenhagen, Denmark.

[49] Hammershøi, Dorte and Jesper Sandvad (1994): "Binaural Auralization, Simulating Free Field Conditions by Headphones." In: *96th AES Convention*. Amsterdam, Netherlands.

[50] Minnaar, Pauli; Søren Krarup Olesen; Flemming Christensen; and Henrik Møller (2001): "The Importance of Head Movements for Binaural Room Synthesis." In: *International Conference on Auditory Display*. Espoo, Finnland.

[51] Hirahara, Tatsuya; Yuki Sawada; and Daisuke Morikawa (2011): "Impact of dynamic binaural signals on three-dimensional sound reproduction." In: *Inter-noise*. Osaka, Japan.

[52] Paquier, Mathieu and Vincent Koehl (2010): "Audibility of headphone positioning variability." In: *128th AES Convention, Convention Paper 8147*. London, UK.

[53] Møller, Henrik; Dorte Hammershøi; Clemen Boje Jensen; and Michael Friis Sørensen (1995): "Transfer Characteristics of Headphones Measured on Human Ears." In: *Journal of the Audio Engineering Society*, **43**(4), pp. 203–217.

[54] Algazi, V. Ralph; Carlos Avendano; and Dennis Thompson (1999): "Dependence of Subject and Measurement Position in Binaural Signal Acquisition." In: *Journal of the Audio Engineering Society*, **47**(11), pp. 937–947.

[55] H., Dillon (1977): "Effect of leakage on the low-frequency calibration of supraaural headphones." In: *Journal of Acoustical Society of America*, **61**(5), pp. 1383–1386.

[56] Brinkmann, Fabian (2010): "On the effect of individual headphone compensation in binaural synthesis." In: *Fortschritte der Akustik: Tagungsband d. 36. DAGA*. Berlin, Germany, pp. 1055–1056.

[57] Lindau, Alexander and Fabian Brinkmann (2012): "Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings." In: *Journal of the Audio Engineering Society*, **60**(1/2), pp. 54–62.

[58] Lindau, Alexander and Stefan Weinzierl (2012): "Assessing the Plausibility of Virtual Acoustic Environments." In: *Acta Acustica united with Acustica*, **98**(5), pp. 804–810.

[59] Huopaniemi, Jyri and Matti Karjalainen (1996): "HRTF Filter Design Based on Audiory Criteria." In: *Nordic Acoustical Meeting*. Helsinki, Finnland, pp. 323–330.

[60] Faller, Christof and Jeroen Breebaart (2011): "Binaural Reproduction of Stereo Signals Using Upmixing and Diffuse Rendering." In: *131st AES Convention, Convention Paper 8541*. New York, USA.

[61] Holman, Tomlinson (2001): "Mixing the Sound." In: *Surround Magazin*, pp.

35–37.

[62] Kendall, Gary S. (1995): "The Decorrelation of Audio Signals and Its Impact on Spatial Imagery." In: *Computer Music Journal*, **19**(4), pp. 71–87.

[63] Holman, Tomlinson (2008): *Surround Sound: Up and Running*. Elsevier/Focal Press.

[64] Oppenheim, Alan V.; Ronald W. Schafer; and John R. Buck (2004): *Zeitdiskrete Signalverarbeitung*, vol. 2. München, Germany: Pearson Studium.

[65] Breebaart, Jeroen and Erik Schuijers (2008): "Phantom Materialization: A Novel Method to Enhance Stereo Audio Reproduction on Headphones." In: *IEEE Transactions on Audio, Speech and Language*, **16**(8), pp. 1503–1511.

[66] Goodwin, Michael M. and Jean-Marc Jot (2007): "Primary-Ambient Signal Decomposition and Vector-Based Localization for Spatial Audio Coding and Enhancement." In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. pp. I–9–I–12.

[67] Goodwin, Michael M. (2008): "Geometric signal decompositions for spatial audio enhancement." In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 409–412.

[68] Avendano, Carlos and Jean-Marc Jot (2004): "A Frequency-Domain Approach to Multichannel Upmix." In: *J. Audio Eng. Soc*, **52**(7/8), pp. 740–749.

[69] Merimaa, Juha; Michael M. Goodwin; and Jean-Marc Jot (2007): "Correlation-Based Ambience Extraction from Stereo Recordings." In: *123rd AES Convention, Convention Paper 7282*. New York, USA.

[70] Avendano, Carlos and Jean-Marc Jot (2002): "Frequency Domain Techniques for Stereo to Multichannel Upmix." In: *22nd AES International Conference on Virtual, Synthetic and Entertainment Audio*. pp. 1–10.

[71] Briand, Manuel; David Virette; and Nadine Martin (2006): "Parametric Representation of Multichannel Audio Based on Principal Component Analysis." In: *AES Convention, Convention Paper 6813*. Paris, France.

[72] Vincent, Emmanuel; Hiroshi Sawada; Pau Bofill; Shoji Makino; and Justinian Rosca (2007): "First stereo audio source separation evaluation campaign: data, algorithms and results." In: *7th Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*. London, United Kingdom, pp. 552–559.

[73] Baek, Yong-Hyun; Se-Woon Jeon; Young-Cheol Park; and Seok-pil Lee (2012): "Efficient Primary-Ambient Decomposition Algorithm for Audio Upmix." In:

133rd AES Convention, Convention Paper 8754. San Francisco, USA.

[74] Pulkki, Ville (1997): "Virtual Sound Source Positioning Using Vector Base Amplitude Panning." In: *Journal of the Audio Engineering Society*, **45**(6), pp. 456–466.

[75] Pulkki, Ville and Tapio Lokki (1998): "Creating Auditory Displays with Multiple Loudspeakers Using VBAP: A Case Study with DIVA Project." In: *Proceedings of International Conference on Auditory Display (ICAD'98)*. Glasgow, United Kingdom.

[76] Jeon, Se-Woon; Young cheol Park; Seok-Pil Lee; and Dae Hee Youn (2011): "Virtual source panning using multiple-wise vector base in the multispeaker stereo format." In: *19th European Signal Processing Conference*. Barcelona, Spain, pp. 1337–1341.

[77] Courrieu, Pierre (2005): "Fast Computation of Moore-Penrose Inverse Matrices." In: *Neural Information Processing - Letters and Reviews*, **8**(2), pp. 25–29.

[78] Hidaka, Takayuki; Leo l. Beranek; and Toshiyuki Okano (1995): "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls." In: *Journal of Acoustical Society of America*, **98**(2), pp. 988–1007.

[79] Ando, Yoichi (1985): *Concert hall acoustics*. New York, USA: Springer-Verlag.

[80] Schroeder, M.R. (1979): "Binaural dissimilarity and optimum ceilings for concert halls: More lateral sound diffusion." In: *Journal of the Acoustical Society of America*, **65**(4), pp. 958–963.

[81] Lee, Kyogu and Julius O. Smith (2004): "Implementation of a Highly Diffusing 2-D Digital Waveguide Mesh with a Quadratic Residue Diffuser." In: *International Computer Music Conference*.

[82] Usher, John and Jacob Benesty (2007): "Enhancement of Spatial Sound Quality: A New Reverberation-Extraction Audio Upmixer." In: *IEEE Transactions on audio, speech and language*, vol. 15. pp. 2141–2150.

[83] Choi, Tacksung; Young Cheol Park; and Dae Hee Youn (2006): "Efficient Out of Head Localization System for Mobile Applications." In: *120th AES Convention, Convention Paper 6758*. Paris, France.

[84] Borß, Christian (2009): "A VST Reverberation Effect Plugin Based on Synthetic Room Impulse Responses." In: *International Conference on Digital Audio Ef-*

fects (DAFx-09). Como, Italy.

[85] Orfanidis, Sophocles J. (2010): *Introduction to Signal Processing*. Rutgers University.

[86] Martin, Aengus; Craig Jin; and André Van Schaik (2009): "Psychoacoustic Evaluation of Systems for Delivering Spatialized Augmented-Reality Audio." In: *Journal of the Audio Engineering Society*, **57**(12), pp. 1016–1027.

[87] Mueller, Florian and Matthew Karau (2002): "Transparent Hearing." In: *Conference on Human Factors in Computing Systems*. Minneapolis, USA, pp. 730–731.

[88] Härmä, Aki; et al. (2004): "Augmented Reality Audio for Mobile and Wearable Appliances." In: *Journal of the Audio Engineering Society*, **52**(6), pp. 618–639.

[89] Tikander, Miikka; Matti Karjalainen; and Ville Riikonen (2008): "An Augmented Reality Audio Headset." In: *11th International Conference on Digital Audio Effects*. Espoo, Finnland.

[90] Homann, Pable F.; Anders Kalsgaard Møller; Flemming Christensen; and Dorte Hammershøi (2014): "Sound localization and speech identification in the frontal median plane with a hear-through headset." In: *Forum Acusticum*.

[91] Westermann, S. (1985): "Comparing BTEs and ITEs for localizing speech." In: *Hearing Instruments*, **36**(2), pp. 20–24.

[92] Leeuw, A.R. (1987): "Speech understanding and directional hearing for hearing-impaired subjects with in-the-ear and behind-the-ear hearing aids." In: *Scandinavian Audiology*, **16**(1), pp. 31–36.

[93] Noble, W. (1990): "A comparison of different binaural hearing aid systems for sound localization in the horizontal and vertical planes." In: *British Journal of Audiology*, **24**, pp. 335–346.

[94] Van den Bogaert, Tim; Evelyne Carette; and Jan Wouters (2009): "Sound localization with and without hearing aids." In: *NAG/DAGA International Conference on Acoustics*. Rotterdam, NL, pp. 1314–1317.

[95] Hammershøi, Dorte and Henrik Møller (1996): "Sound transmission to and within the human ear canal." In: *Journal of Acoustical Society of America*, **100**(1), pp. 408–427.

[96] Woodworth, Robert S. and Harold Schlosberg (1962): *Experimental psychology*. New York: Holt, Rinehard and Winston.

[97] Sander, Christian; Frank Wefers; and Dieter Leckschat (2012): "Scalable Binaural Synthesis on Mobile Device." In: *133rd AES Convention, Convention*

Paper 8783. San Francisco, USA.

[98] Moustakas, Nikos; Andres Floros; and Nicolas Grigoriou (2011): "Interactive Audio Realities: An Augmented / Mixed Reality Audio Game prototype." In: *130th AES Convention*. London, UK.

[99] Milgram, Paul and Herman Colquhoun Jr. (1999): *Mixed Reality: Merging Real and Virtual Worlds*, chap. A Taxonomy of Real and Virtual World Display Integration. Berlin Heidelberg: Springer-Verlag.

[100] ITU (2011/12): *Rec. ITU-T P.57: Arificial Ears*. Geneva, Switzerland: International Telecommunication Union.

[101] User-forum, Apple (2015): URL `http://www.apple.com/shop/question/answers/iphone/what-are-the-technical-specifications-for-the-earpod-headphone/Q9KHDUAF4CDAK79YY`.

[102] Bartz, Peter (2014): "9DOF Razor IMU." URL `https://github.com/ptrbrtz/razor-9dof-ahrs/wiki/Tutorial`.

[103] Lindau, Alexander (2009): "The Perception of System Latency in Dynamic Binaural Synthesis." In: *NAG/DAGA*. Rotterdam, pp. 1063–1066.

[104] Thomas, Alexander (2002): URL `http://www.dr-lex.be/info-stuff/volumecontrols.html`.

[105] Menzer, Fritz and Christof Faller (2010): "Investigations on an Early-Reflection-Free Model for BRIRs." In: *Journal of the Audio Engineering Society*, **58**(9), pp. 709–723.

[106] Wefers, Frank (2010): "OpenDAFF." In: *DAGA 2010*. Berlin, Germany.

[107] Majdak, Piotr and Markus Noisternig (2014): *SOFA - Spatially Oriented Format for Acoustics*. Tech. rep.

[108] Selesnick, Ivan W. and C. Sidney Burrus (1998): "Generalized Digital Butterworth Filter Design." In: *IEEE Transactions on Signal Processing*, **46**(6), pp. 1688–1694.

[109] CARLab (2010): "Listening through different ears in the Sydney Opera House." URL `http://www.ee.usyd.edu.au/carlab/UserFiles/SOH/index.html`.

[110] Merimaa, Juha; Timo Peltonen; and Tapio Lokki (2005): "Concert Hall Impulse Responses - Pori, Finland." URL `http://legacy.spa.aalto.fi/projects/poririrs/`.

[111] Opera, Sydney (2007): *Sydney Opera House - Technical and Production Information*. Tech. rep., Sydney Opera House.

A. Headphone measurment

MESSPROTOKOLL

MESSUNG VON CONSUMER KOPFHÖRERN

Raffael Tönges

1 EINLEITUNG

Im Rahmen der Masterarbeit "An augmented acoustics demonstrator with realtime stereo up-mixing and binaural auralization" werden verschiedene kostengünstige Kopfhörer mit dem Messroboter FABIAN (Fast and Automatic Binaural Impulse response AcquisitioN [1]) gemessen. Es werden die Frequenzgänge von acht Kopfhörern aufgenommen. Darunter sind sechs in-ear Kopfhörer, ein supra-auraler Kopfhörer und ein circum-auraler Kopfhörer.

2 MESSUNG

In diesem Abschnitt werden die Messtechnik, die Referenzmessung und der Messaufbau für die Kopfhörermessungen beschrieben.

2.1 MESSTECHNIK

Die Messungen werden mit FABIAN gemacht. In den beiden blockierten modellierten Gehörkanälen sind Mikrofone vom Typ *DPA 4060* montiert, welche mit einem *Lake People MIC-Amp C360* verbunden sind. Zur Aufnahme der Messdaten am Computer ist der Mikrofonvorverstärker über das Audiointerface *RME Hammerfall DSP Multiface II* an einem Computer angeschlossen. Auf dem Computer läuft Windows 32 bit und die FABIAN Messsoftware.

2.2 REFERENZMESSUNG

Zunächst wird eine Referenzmessung gemacht, um später den Einfluss des Messsystems auf die Messungen kompensieren zu können. Hierfür werden die Ausgänge des Audiointerface an den Mikrofonvorverstärker angeschlossen. Die Ausgänge des Mikrofonvorverstärkers werden wiederum an die Eingänge des Audiointerface angeschlossen. Als Anregungssignal dient ein mittellanger Sweep (1,5 s) der von 20 bis 22000 Hz ansteigt und eine Basserhebung von 20 dB bis 100 Hz hat (Monkey Forest Preset m. Bass Shelf). Er wurde mit 44,1 kHz abgetastet. Die Aussteuerung des Anregungssignals ist -10 dBFS. Die Blockgröße beträgt 256 Samples und die FFT Länge beträgt 16.

Für das System ergibt sich ein Roundtrip-Delay von 2415 Samples. Der Delay wird bei den weiteren Messungen durch das FABIAN Messsystem automatisch kompensiert.

2.3 MESSAUFBAU

Während der Messungen werden die zu messenden Kopfhörer von dem Audiointerface mit dem gleichen Sweep angesteuert, wie er auch für die Referenzmessung verwendet wurde (1,5 s, 44,1 kHz, 20-22k Hz, Bass Shelf 20 dB bis 100 Hz). Das System wird zwischen 29,6 Hz (-3 dB) und 21260 Hz (-3 dB) gemessen. Die Aussteuerung des Sweeps ist bei jedem Kopfhörer individuell gewählt, sodass ein möglichst guter Signal-Rausch-Abstand erzielt wird.

Die Mikrofone sind an den Mikrofonvorverstärker angeschlossen und dieser an den Eingängen des Audiointerface.

Die Mikrofone sind so in den Gehörkanal gesteckt, dass die Membran mit dem Eingang des Gehörkanal abschließt. Die Seitenansicht von FABIAN ist in Abb. 1 abgebildet. Der circum-aurale und der supra-aurale Kopfhörer werden mit dieser Mikrofonposition gemessen. Da die Mikrofone ein Gitter vor den Membranen haben, welches ca. 2 mm vorsteht, werden die Mikrofone von den in-ear Kopfhörern ca. 2 mm weiter in den Gehörkanal hineingedrückt. Innerhalb der Messungen der in-ear Kopfhörer variiert die Position der Mikrofone nicht. In Abb. 2 sind exemplarisch zwei Messkonfigurationen abgebildet.

Abbildung 1: FABIAN mit Mikrofonen im Gehörkanal.

3 MESSERGEBNISSE

In diesem Abschnitt sind die Messergebnisse der einzelnen Kopfhörer angegeben und mit Grafiken veranschaulicht.

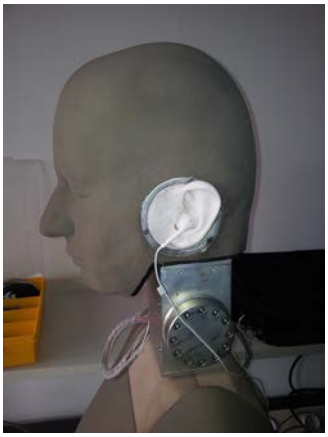
BEYERDYNAMICS DT 770 PRO

Bei diesem Kopfhörer (Abb. 3) handelt es sich um den circum-auralen Kopfhörer *Beyerdynamics DT 770 Pro* mit geschlossener Bauart und 32 Ohm Widerstand. Der Sweep ist mit -10 dB ausgesteuert. Die Energy Time Curve und der Frequenzgang sind in Abb. 4 gezeigt.

URBANEARS ZINKEN

Bei diesem Kopfhörer (Abb. 5) handelt es sich um den circum-auralen Kopfhörer *UrbanEars Zinken* mit geschlossener Bauart. Der Sweep ist mit -15 dB ausgesteuert. Die Energy Time Curve und der Frequenzgang sind in Abb. 6 gezeigt.

(a) FABIAN mit circum-auralen Kopfhörern.



(b) FABIAN mit in-ear Kopfhörern.

Abbildung 2: FABIAN mit verschiedenen Kopfhörern.

3.1 In-ear Kopfhörer

iPhone 6

Bei diesem Kopfhörer (Abb. 7) handelt es sich um einen in-ear Kopfhörer, der mit dem *Apple iPhone 6* ausgeliefert wird. Der Sweep ist mit -20 dB ausgesteuert. Die Energy Time Curve und der Frequenzgang sind in Abb. 8 gezeigt.

iPhone 5S

Bei diesem Kopfhörer (Abb. 9) handelt es sich um einen in-ear Kopfhörer, der mit dem *Apple iPhone 5S* ausgeliefert wird.Der Sweep ist mit -20 dB ausgesteuert. Die Energy Time Curve und der Frequenzgang sind in Abb. 10 gezeigt.

iPhone 4S

Bei diesem Kopfhörer (Abb. 11) handelt es sich um einen in-ear Kopfhörer, der mit dem *Apple iPhone 4S* ausgeliefert wird. Der Sweep ist mit -23 dB ausgesteuert. Die Energy Time Curve und der Frequenzgang sind in Abb. 12 gezeigt.

iPhone Kopfhörer (Nachbau)

Bei diesem Kopfhörer (Abb. 13) handelt es sich um einen in-ear Kopfhörer, der dem *Apple Kopfhörer* nachempfunden ist. Der Sweep ist mit -20 dB ausgesteuert. Die Energy Time Curve und der Frequenzgang sind in Abb. 14 gezeigt.

Creative

Bei diesem Kopfhörer (Abb. 15) handelt es sich um einen in-ear Kopfhörer der Marke *Creative*. Der Kopfhörer ist dem MP3 Spieler *Creative Zen* beigelegt. Der Sweep ist mit -23 dB ausgesteuert. Die Energy Time Curve und der Frequenzgang sind in Abb. 16 gezeigt.

Deluxe

Bei diesem Kopfhörer (Abb. 17) handelt es sich um einen in-ear Kopfhörer mit der Bezeichnung *Deluxe Stereo Earphones MX-028*. Der Sweep ist mit -18 dB ausgesteuert. Die Energy Time Curve und der Frequenzgang sind in Abb. 18 gezeigt.

Abbildung 3: Beyerdynamics DT 770 Pro.

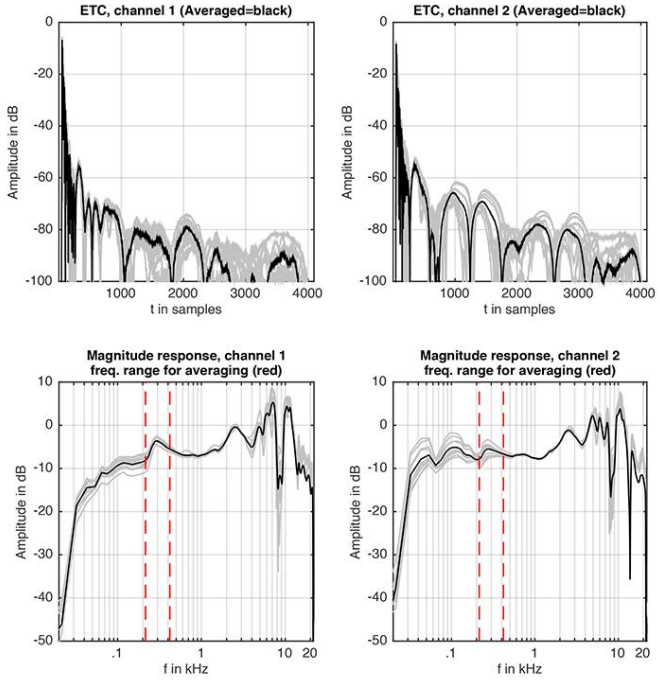


Abbildung 4: Beyerdynamics DT 770 Pro: Energy Time Curve (oben) und Frequency Response (unten) pro Messung (grau) und im Durchschnitt (schwarz).

Abbildung 5: UrbanEars Zinken.

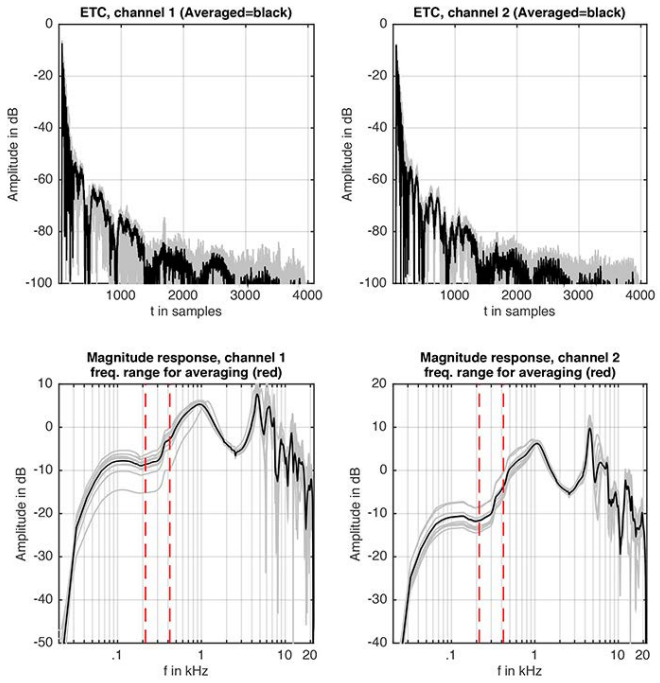


Abbildung 6: UrbanEars Zinken: Energy Time Curve (oben) und Frequency Response (unten) pro Messung (grau) und im Durchschnitt (schwarz).

Abbildung 7: iPhone 6 Kopfhörer

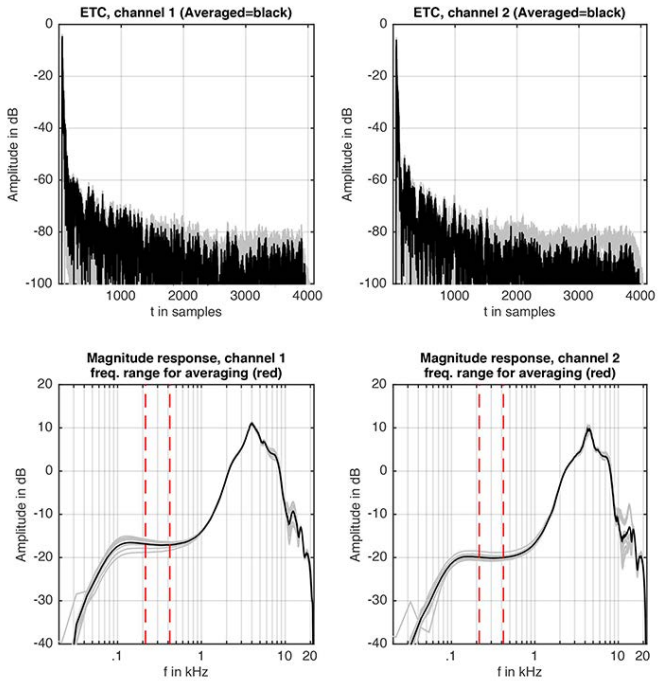


Abbildung 8: iPhone 6 Kopfhörer: Energy Time Curve (oben) und Frequency Response (unten) pro Messung (grau) und im Durchschnitt (schwarz).

Abbildung 9: iPhone 5S Kopfhörer

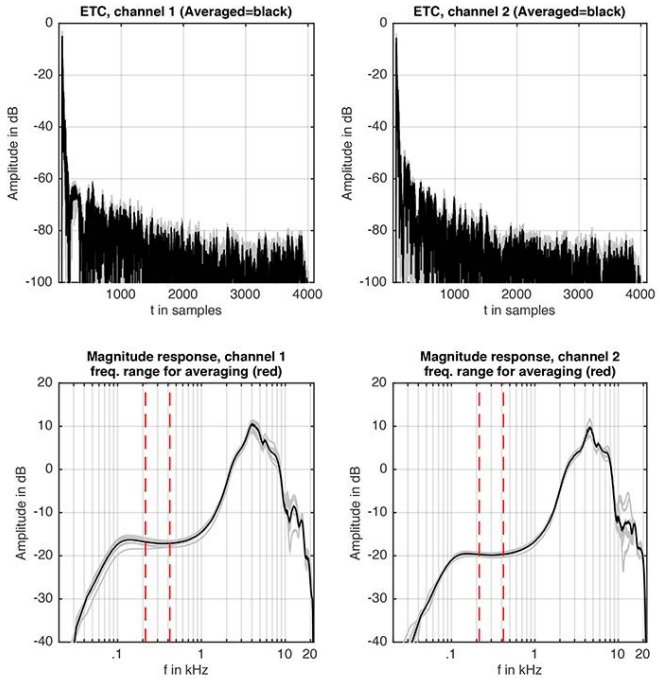


Abbildung 10: iPhone 5S Kopfhörer: Energy Time Curve (oben) und Frequency Response (unten) pro Messung (grau) und im Durchschnitt (schwarz).

Abbildung 11: iPhone 4S Kopfhörer

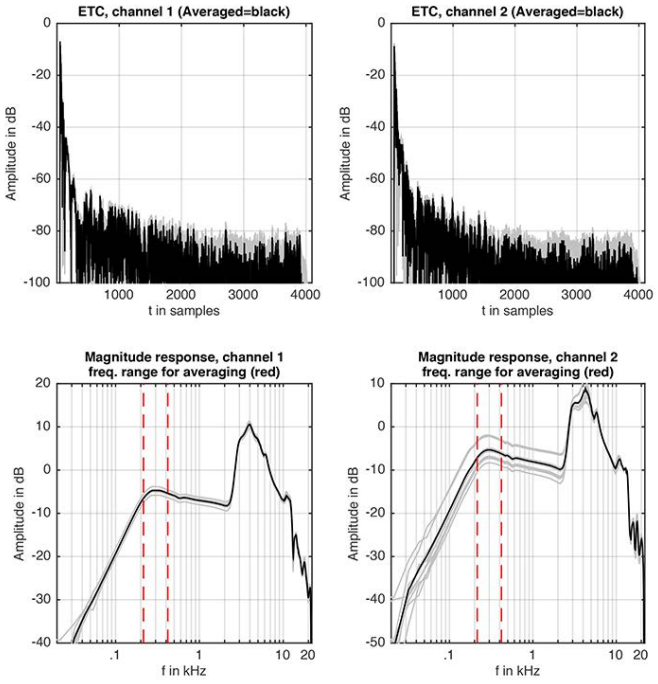


Abbildung 12: iPhone 4S Kopfhörer: Energy Time Curve (oben) und Frequency Response (unten) pro Messung (grau) und im Durchschnitt (schwarz).

Abbildung 13: iPhone Kopfhörer (Nachbau)

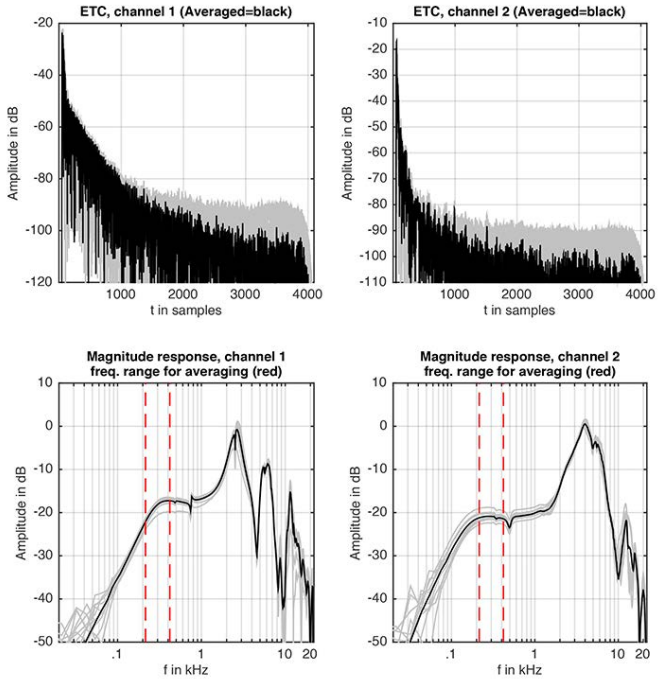


Abbildung 14: iPhone Kopfhörer (Nachbau): Energy Time Curve (oben) und Frequency Response (unten) pro Messung (grau) und im Durchschnitt (schwarz).

Abbildung 15: Creative in-ear Kopfhörer.

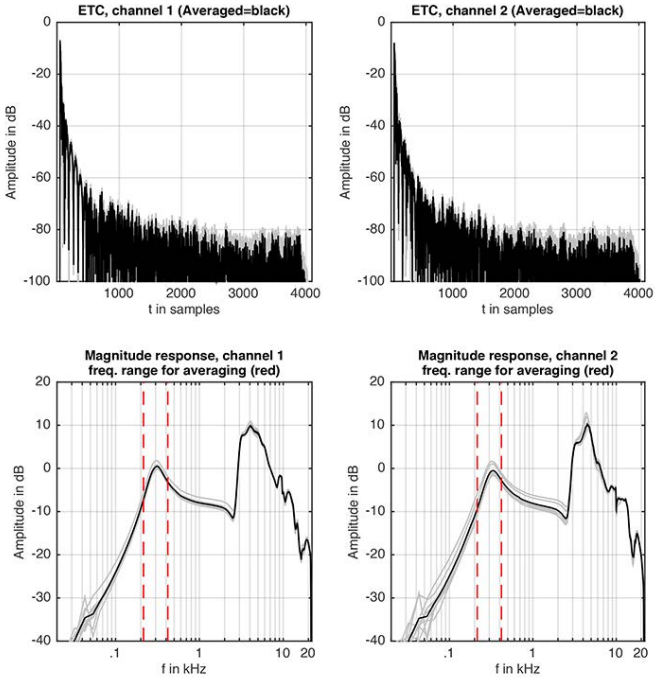


Abbildung 16: Creative in-ear: Energy Time Curve (oben) und Frequency Response (unten) pro Messung (grau) und im Durchschnitt (schwarz).

Abbildung 17: Deluxe Stereo Earphones MX-028.

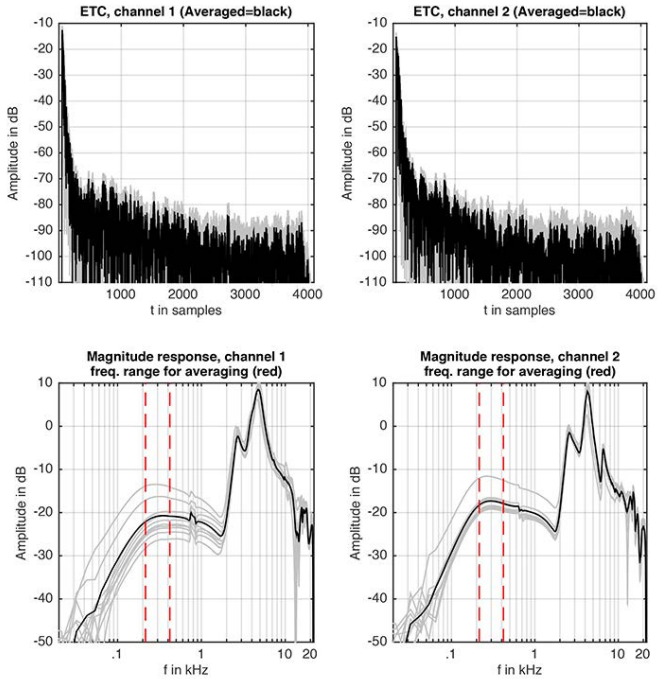


Abbildung 18: Deluxe Stereo Earphones MX-028: Energy Time Curve (oben) und Frequency Response (unten) pro Messung (grau) und im Durchschnitt (schwarz).

LITERATUR

[1] LINDAU, Alexander ; STEFAN, Weinzierl: FABIAN - An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom. In: *24. Tonmeistertagung – VDT International Convention* (2006)

B. Data sheets

Beyerdynamics DT 770 Pro

DT 770 PRO

Dynamic Headphone

Order # 459.046 (250 Ω)
Order # 474.746 (80 Ω)
Order # 483.664 (32 Ω)


32 Ω version


80 Ω version


250 Ω version

Features

- Closed diffuse field studio headphone
- Innovative bass reflex system
- Robust spring steel headband
- Single sided cable
- Soft inner headband
- Gold plated jack plug (3.5 mm) and adapter (6.35 mm)

Technical Specifications

Transducer type	Dynamic
Operating principle	Closed
Nominal frequency response	5 - 35,000 Hz
Nominal impedance	32 Ω / 80 Ω / 250 Ω
Nominal SPL	96 dB SPL
Nominal T.H.D	< 0.2%
Power handling capacity	100 mW
Sound coupling to ear	Circumaural
Ambient noise isolation	
32 Ω version	approx. 20 dBA
80 Ω / 250 Ω version	approx. 18 dBA
Nominal headband pressure	approx. 3.5 N
Weight (without cable)	270 g
Length and type of cable	
32 Ω version	1.6 m / straight cable
80 Ω version	3 m / straight cable
250 Ω version	3 m / coiled cable
Connection	Gold plated stereo jack plug (3.5 mm) and 1/4" adapter (6.35 mm)

(all specifications according to EN 60 268-7)

Applications

The DT 770 PRO is a closed dynamic headphone of exceptional quality suitable for the most demanding professional and audiophile applications. The long term comfort and accurate performance make the DT 770 PRO the perfect monitoring headphone for recording studios, post production or broadcasting situations.
The low mass coil and diaphragm assembly produce a transient performance equalled only by electrostatic earphones, and, in combination with a carefully tailored frequency response offer a natural and balanced sound.
Soft earpads and adjustable, sliding, earpieces together with a single sided connecting cable ensure listening comfort during extended periods of use.
The DT 770 PRO features 32, 80 or 250 ohm drivers and a gold plated 3.5 mm stereo jack with 1/4" inch adapter, and is therefore suitable for use with almost all headphone amplifiers.

Spare Parts

EDT 770 S	Ear pads, soft PVC, circumaural	Order # 904.783
EDT 770 V	Ear pads, velours, circumaural	Order # 926.660
EDT 770 VB	Ear pads, velours, circumaural, black	Order # 906.166
EDT 990 S	Ear pads, soft PVC, especially for 32 Ω version circumaural, black	Order # 904.791
BN 59-53/D	Headband pad	Order # 990.681

beyerdynamic GmbH & Co. KG
Theresienstr. 8 | 74072 Heilbronn – Germany
Tel. +49 (0) 71 31 / 617 - 0 | Fax +49 (0) 71 31 / 617 - 204
info@beyerdynamic.de | www.beyerdynamic.com

Panasonic WM-61a

Omnidirectional Back Electret Condenser Microphone Cartridge

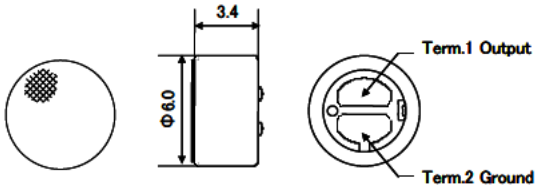
WM-61A

- Small microphone for general use
- Back electret type designed for high resistance to vibrations,
- High sensitivity, high signal to noise ratio
- Solder type for leadwires

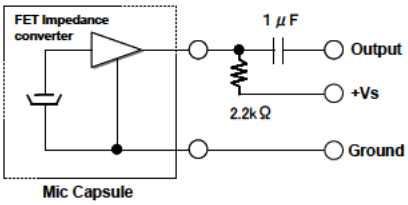
■Appearance

■Dimensional Drawing

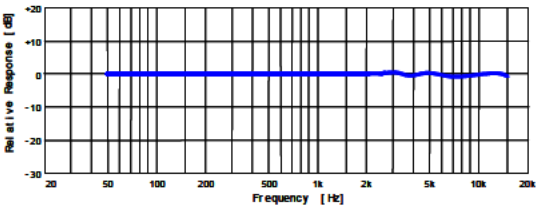
Unit : mm



■Schematic Diagram



■Typical Frequency Response Curve



■Sensitivity

(0dB=1V/Pa, 1kHz)

Sensitivity	Model No.
−35+/−4dB	WM−61A

Vs=2.0V
RL=2.2kΩ

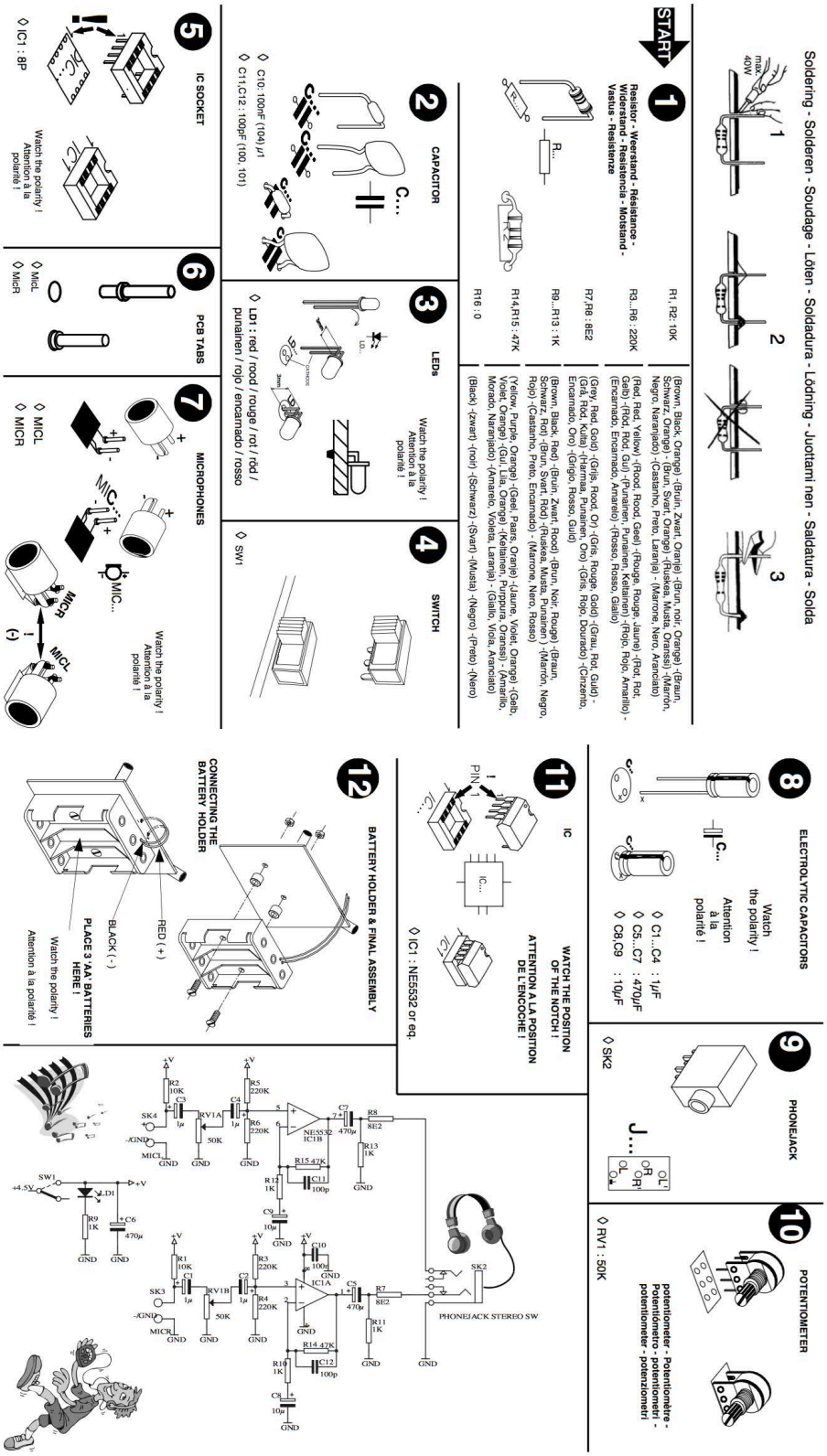
Specifications

Dimensions	6 x 3.4 mm
Impedance	Less than 2.2kΩ
Frequency	Omnidirectional 20−16,000Hz
Max operation voltage	10V
Standard operation voltage	2V
Current consumption	Max 0.5mA
Sensitivity reduction	Within −3dB at 1.5V
S/N ratio	More than 62dB

Specifications are subject to change without notice.

Panasonic Corporation

Velleman Super Stereo Ear



LightBlue Bean

Sensor board

Model	GY-85 (9 DOF break through)
Quantity	1
Color	Blue
Material	PCB
Features	Nine-axis module (three-axis gyroscope + triaxial accelerometer + three-axis magnetic field); Chip: ITG3205 + ADXL345 + HMC5883L; Power supply : 3-5V; Communication : IIC communication protocol (fully compatible with the 3-5v System, circuit contain LLC)
Application	For DIY project

specification
Dimensions: 0.87 in x 0.63 in x 0.12 in (2.2 cm x 1.6 cm x 0.3 cm)
Weight: 1.09 oz (31 g)

C. Matlab scripts

Name	Function
run_regulated_inversion.m	Script to create minimal phase headphone compensation filter written by Fabian Brinkmann, Alexander Lindau and Zora Schaerer.
wav2ir.m	Script to convert wav file containing impulse response to text file containing impulse response.
itd_diameter_error.m	Script to plot ITD error as function of the head diameter.
volume_curves.m	Script to plot volume curves, used for volume adjustment in the demonstrator application.
MVBNAP.m	Script to plot the panning gains of MVBNAP.
upsampling_and_externalisation.m	Prototype script for up-mixing and binaural auralization.
UpsamplingExternalisationScript.m	Script to execute upsampling_and_externalisation.m.

Table C.1.: The *Matlab* scripts used for prototyping, analysing or realising parts of the application.