A real time compensation system for spatial audio

Development of a functional prototype

Nikolaj Villefrance Lerke

Sound and Music Computing, SMC10, 2019-06

Master's Project

Architecture and Media Technology
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
A real time compensation system for spatial audio

Theme:
Headphone Transfer Function, Frequency response compensation, sound level compensation, spatial audio

Project Period:
Spring Semester 2019

Author:
Nikolaj Villefrance Lerke

Supervisor(s):
Michele Geronazzo(mge@create.aau.dk)
Stefania Serafin(sts@create.aau.dk)
Christian Buch Lorentzen(cbl@aiaiai.dk)

Copies: 1

Page Numbers: 67

Date of Completion:
May 28, 2019

Abstract:

A real-time headphone compensation system for spatial audio is presented in this thesis. To achieve optimal spatial audio accuracy it is necessary to know the listener's personal ear response. Ear response has previously been measured using complex setups that are impractical for regular consumer applications. So, The purpose of this project was to develop a simple compensation system using affordable components.
Based on results from the a small ear response study, a prototype system was developed. The hardware component of the system was based on a modified AIAIAI TMA-2 headphone, while the software component was divided into three subsystems implemented in Matlab Simulink. Prototype system evaluation showed promising results, however, the convergence speed of the adaptive subsystems as well as the accuracy of the magnitude estimation in the peak EQ subsystem could be improved.

Contents

Preface

Here is the preface. You should put your signatures at the end of the preface.

Aalborg University, May 28, 2019

Author 1
<username1@XX.aau.dk>

Author 2
<username2@XX.aau.dk>

Author 3
<username3@XX.aau.dk>

Chapter 1

Introduction

1.1 Problem

The listener experience of headphones is generally not consistent. Headphones that sound natural for one listener may not provide a similar experience for another listener. This is a problem when reproducing spatial audio using headphones. An investigation done by Gutierrez-Parera and Lopez conclud that the frequency response provided by a set of headphones was highly correlated with the localisation accuracy [19].

Not only that, but the experience also depends on the exact placement of the headphone on the listener's head, so the sound of the headphones varies slightly each time the listener put on the headphones. Factors affecting the frequency response, which the headphone manufacturers cannot influence include:

- Shape and size of listener's pinna and ear-canal
- Distance from driver to the entrance of the ear-canal
- Cushion leak
- Headphone placement
- Left-right disparity
    - Frequency response
    - Sensitivity

A listener's pinna and ear-canal creates sound reflections influencing the frequency response measured at the ear-drum [9]. The shape and size of the pinna and ear-canal affects the behaviour of the reflections, and therefore also the frequency response measured at the ear-drum. This response depends on the direction of the sound source relative to the listener's head [26]. The shape of the

response is then used by the listener's auditory system as one of the cues for locating the sound source. Using the listener's personal HRTF for spatial audio synthesis therefore enhances the perceived realism of the sound. However, gathering the listener's personal HRTF is a complicated and resource intensive task that is unpractical for regular consumer audio systems [26]. As such a simple HRTF extraction system using a set of headphones could be a great asset.

The distance from the driver unit to the ear-canal entrance is largely the result of three factors; headband clamping force, stiffness of the cushion and the size of the listener's head [4]. This distance affects the size and shape of the room between the ear-canal and driver unit, so the behaviour in a room is also affected.

The base response of closed back over/on ear headphones very much depends on a complete seal between the headphone cushion and the listener's head [13]. A broken seal will result in a drastic decrease in magnitude at the low end of the frequency spectrum. This can be a problem if the listener has a lot of hair or is wearing ear rings or glasses, but is especially problematic for on-ear headphones since they can be difficult to place on the ear.

Frequency response at different positions inside the space between the speaker unit and the ear canal entrance varies due to the previously discussed reflections [4]. Since the placement of the headphones vary slightly each time a listener is wearing them, the resulting frequency response at the ear-drum will also vary. This may result in an inconsistent listener experience where the headphones sound good to the listener some times and not so much other times.

Perceptually deviations between what is heard at one ear and what is heard at the other is used for sound source localisation [13]. So the listener is very sensitive to differences between the two speaker units. As mentioned above the exact placement of the headphones on the head affects the resulting perceived frequency response and the headphones rarely sits exactly the same on both ears. There may furthermore be small deviations in frequency response, phase response and sensitivity from a driver unit to driver unit due to manufacturing tolerances.

Compensating for the deviations in frequency response and sensitivity between two speaker units will therefore improve the spatial reproduction accuracy of the headphones. The perception of localisation cues is very sensitive [13], so it is very important that the left and right signals behave the same.

A real-time compensation system continuously adjusts the compensation filters to achieve a desired frequency response or sound level [19]. Some of the aforementioned factors may vary over time. If the listener adjusts the position of the headphones the frequency response at the ear drum will change. A compensation system running in real-time is able to adjust for the changes, so bringing the response back to the desired shape or correct sound level. Beside being able to update continuously, a real-time system must also be able to calculate the appropriate filter coefficients based on an arbitrary input signal [19]. Frequency response

or sensitivity measurements are usually gathered using dedicated stimulus signals such as logarithmic sine sweeps or pink noise [11]. Whereas an arbitrary stimulus signal is just any audio signal. Compensation using an arbitrary stimulus may be achieved by comparing a measured version of the signal with the original signal. It is then possible to calculate the frequency response at the frequencies that are included in the original arbitrary stimulus signal.

1.2 Scope of the project

In theory the most accurate spatial audio reproduction would require that the exact response is measured at the listener's ear-drum. Personal HRTFs are measured using arrays of loudspeakers and elaborate test setups [26]. The complexity and time consumption of these measurement setups makes personal HRTF compensation for consumer applications completely impractical. Such a system is not within the scope of this project as it would impractical to implement. This thesis focuses on a practical application that is realistic for consumer headphone applications.

For this project it is therefore much more interesting from a consumer standpoint to explore how an accurate compensation can be achieved by using a simple microphone placed inside the ear-cup. It is quite realistic to imagine a microphone placed inside the ear-cups of a headphone. Such microphones are actually already becoming more common with growing popularity of Active Noise Cancelling (ANC) headphones. Feedback ANC systems require a microphone to be placed inside the ear-cup, so it therefore realistic to imagine a compensation system running alongside an ANC system using the same microphone.

The choice of hardware, including microphones, are limited by price and availability to construct a hardware prototype that is realistic for consumer applications. The project is completed in collaboration with AIAIAI[1], so all hardware components is sourced from their existing products. This naturally limits the accuracy of the compensation, but as mentioned earlier a perfect compensation system is not within the scope of this project, which should rather be read as a proof of concept for a more realistic system in terms of spatial audio applications for regular consumers.

Synthesis or reproduction of spatial audio is not within the scope. As this project is limited to headphone compensation and HRTF feature extraction. Utilisation of HRTF features to produce or reproduce realistic spatial audio is an entire field of research in itself thus could not be included in this project given the time and resources available. See section 6.3 for a discussion of how the findings of this project could possibly be used for spatial audio production or reproduction in the future.

[1] AIAIAI.dk

TMA-2 headphones produced by AIAIAI is utilised for the project as the modularity of the headphones makes them flexible to work with. Furthermore, it was decided to limit the project to apply a circumaural type of headphone. Several different types of headphones are characterised by the way they are worn by the listener [4]. Circumaural headphones feature cushions placed around the listener's ears thus creating a closed space, while supraaural headphones utilise smaller cushions and are placed directly on the listener's outer ears. During initial prototype development, see section 4, it was found that circumaural cushions was easier to work with, so it was decided to focus on this type.

1.3 System proposal

The objective of this project is to develope a functional prototype consisting of a circumaural headphone connected to a Laptop running the proposed algorithm using Matlab Simulink[2]. Microphones will be installed in each ear-cup to measure sound levels and frequency responses at each ear. Initial studies showed that a microphone placed inside the ear-cup was able to accurately determine frequency response changes in the range 50Hz-10.000kHz.

The system will run in real-time, which means that the stimulus is an arbitrary audio signal. This signal is compared with the signal measured inside the ear-cup. The parameters of a compensation filter will then be adjusted based on the results of the comparison. The audio signal will be put through a set of compensation filters before being played through the driver-unit. The system will then be able to compensate for:

- Left-Right sound level disparity
- Left-Right frequency response disparity
- Individual outer ear shape

A general overview of the adaptive algorithm components can be seen in figure 1.1.

Based on the audio input and the recorded microphone response, it will be possible to adjust the sound levels of the left and right headphone speaker units to compensate for any sensitivity disparities. The power ratio between the left and right side will be calculated for both the audio input and the microphone input. These two ratios will then compared to determine the appropriate sound level compensation at the left or right side.

Frequency response compensation of the low and high frequencies will be divided into two separate subsystems. Requirements for a frequency response compensation system depends on the frequency range [4]. At low frequencies the

[2] se.mathworks.com/products/simulink

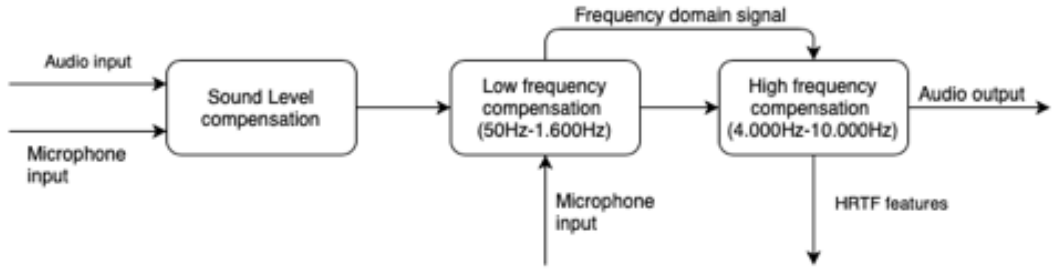


Figure 1.1: Overview of the proposed compensation system

system will compensate for leaks between the cushion and the listener's head and irregular frequency response introduced by the headphone driver. To determine the compensation coefficients the frequency domain of the audio signal and microphone signal will be calculated.

At high frequencies, the system will compensate for resonances caused by the outer ear and ear canal. These resonances will be detectable as peaks in the 4.000Hz-10.000Hz range [15]. The high frequency compensation subsystem therefore needs to be able to detect these peaks and utilise precise filters to compensate for them. The peak characteristics must furthermore be extracted as they provide information that can be used to make a rough estimate of the listener's personal HRTF.

the next sections of this thesis describes the research that formed the basis the prototype, the implementation and an evaluation of the system. Section 2 introduces the current state of the art in the fields of auditory localisation perception, spatial audio reproduction using headphones and presents theoretical solutions for problems discussed earlier in this section. A study of the ear response is then discussed in section 3. Results from the study was used together with research from section 2 to form a set of requirements for the prototype system that is discussed in section 4. Section 4 also presents implementation details for both the hardware and software components of the prototype. An evaluation of the prototype is then presented in section 5, and the results of the evaluation is discussed in section 6. Conclusions drawn from the thesis are presented in section 7.

Chapter 2

Background

2.1 Perception of sound localisation

2.1.1 Auditory coordinate system

When describing the location of an auditory event relative to the user's head a coordination system such as the one shown in figure 2.1 is used. This system proposed by Blauert is utilised by assigning three coordinates to the location of a sound source relative to the listener; azimuth, elevation and distance [3]. The azimuth is the horizontal angle relative the sound source where 0 degrees is right in front of the listener. The elevation denotes the vertical angle. The distance is simply the distance from the listener to the location of the sound source. The location of the auditory event can then be evaluated in relation to three distinct auditory planes; horizontal, median and frontal. The origin of the system is placed halfway between the user's ears.

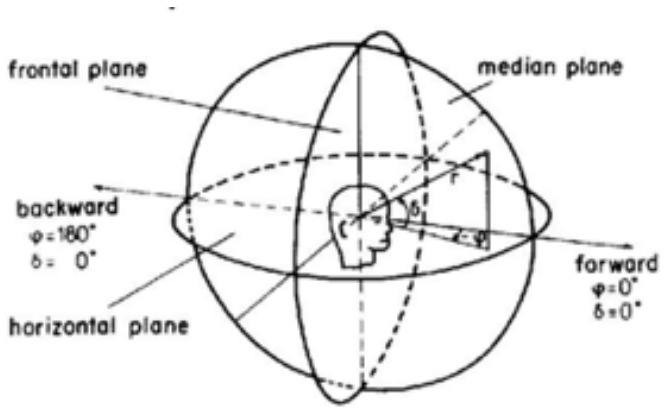


Figure 2.1: Illustration of the auditory localisation system [3]

2.1.2 Localisation cues

The localisation cues used by the human auditory system can be divided into two categories; binaural and monaural cues. Binaural cues are ones that use both ears for localisation and includes Interaural Time Difference (ITD) and Interaural Level Difference (ILD) [3]. The ITD is determined by the delay between the sound arriving at one ear and arriving at the other ear. It is therefore used to approximate the azimuth of the auditory event relative to the listener's head. If the auditory event is located to the right relative to the listener's head the sound will have to travel farther to reach the left ear than the right and will therefore be slightly delayed. If the distance between the listener's ears and the azimuth angle is known the ITD can be calculated using equation 2.1 [21].

$$ITD = \frac{a}{c}(\theta + sin(\theta)) \tag{2.1}$$

Where a is half the distance between the ears, c is the speed of sound and θ is the azimuth angle. The maximum ITD value therefore depends on the distance between the listener's ears, and is achieved when the azimuth angle is either 90° or 270°. Experiments have shown that the most sensitive listeners may detect ITD changes as small as 10 μS [17][31].

ILD, as the name suggests, is the sound pressure level difference at the two ears. Sound pressure level decreases as sound moves further away from the source, so the sound pressure level at the ear drum closer to the sound source will be slightly higher than at the ear drum farther from the sound source. The ILD is furthermore dependant on the acoustic shadow caused by the size and shape of the listener's head [3]. Sensitive listeners have been shown to be able to notice ILD differences as small as 0,5dB [14][16].

Sensitivity to changes in ILD and ITD depends on the spectral content and the azimuth of the auditory event [21]. Broadband noises are generally more difficult to accurately localise than pure sine tones. Perception of ILD and ITD changes is also more sensitive around 0° and 180° than for sound sources placed around 90° or 270°. One of the difficulties of localisation using ILD and ITD is the so called cone of confusion [3]. If the sound source is placed at a certain distance and azimuth relative to the listener's head a cone can be drawn which axis lies along the axis between the two ears. At every point on this cone the ILD and ITD values will be similar. So, the listener will not be able to determine the location of the sound source; it may be in front, behind, above or below relative to the listener. Monaural cues are therefore invaluable for accurate sound source localisation.

Monaural cues, as opposed to binaural cues, may use a single ear and involves evaluation of the auditory event's level and spectral characteristics. These cues are primarily used to determine the elevation and distance of the sound source relative to the listener. The Head Related Transfer Function (HRTF) describes the spectral

changes of the sound caused by the transmission from free-field to the ear drums [22]. These spectral changes are primarily caused by the shape and size of the listener's shoulders, head, pinna and ear canal. The shape of the ear canal and pinna varies hugely from person to person, so predicting the spectral content at the ear drum is no trivial task. When the sound arrives at the ear the pinna causes a series of reflections that in turn results in different resonances at different frequencies. The ear canal and pinna can therefore be said to act like a simple acoustic filter. The spectral content furthermore depends on the direction from which the sound arrives at the pinna. This allows the listener to distinguish whether the auditory event is in front or behind the listener and estimate the elevation of the auditory event. An investigation by Iida et al looked into how the frequency response at the ear drum changed as the location of the sound source moved [15]. They found that the frequency response included three important features that may be used for auditory localisation. The first feature is a pronounced peak somewhere around 4-6kHz, which is then followed by two notches. The peak remains mostly stationary as the sound source is moved, while the gain and centre frequency of the two notches varied wildly depending on the position of the sound source. This indicated according to Iida et al that the hearing system analyses the centre frequency and gain of the two notches in relation to the first peak to form a positional cue. Their investigation furthermore concluded that a high level of localisation accuracy could be achieved by replicating just those three spectre elements.

2.1.3 Distance cues

he distance from the listener to the sound source is also estimated using monaural cues. Four factors contribute to the distance perception[26]:

- Intensity cues
- Familiarity of experience
- Direct to reverberant ratio
- Spectral characteristics of reverb

When propagating through air in a completely anechoic environment the sound intensity will decrease by roughly 6dB each time the distance is doubled. A sound that is closer to the listener will therefore be louder to the listener than the same sound at a longer distance. This cue is efficient when estimating the relative distance of multiple sound sources, but cannot be used to estimate the absolute distance to a single source. If it was the only distance cue it would not be possible to perceive the difference between a soft sound that is close to the listener and a loud sound that is far away from the listener.

Based on previous experience a listener forms an expectation of the distance to the sound source. An example of this is when two sounds reach the listener at the same sound level, one is a high pitched shout and the other a low pitched whisper. The listener will assume that the shout is far away while the whisperer is near by. Psychoacoustic experiments by Gardner concluded that listeners had a tendency to overestimate the distance when the stimuli was a shout, while they underestimated the distance when presented with a whisper[12]. When in a familiar environment the distance perception is furthermore affected by the listener's previous experience of how sound behaves in that environment. If the listener is close to a road and hears a car, it will be assumed that the car is somewhere on that road.

The relative energy of the direct sound compared to that of the reverb is also dependant on the distance between sound source and listener. A smaller distance will result in a greater ratio of direct-to-reverberant energy. At the reverberation distance the pressure of the direct sound is the same as the reverb sound pressure. However, the direct-to-reverberant energy ratio is also dependant on the surrounding environment. If the listening environment is completely anechoic the direct-to-reverberation ration will equal 1 no matter the distance.

The spectral characteristics of the reverb also depends on the distance between sound source and listener. How the sound spectrum changes as the distance increases is frequency dependant. The frequency contents of the reverb will therefore be slightly different than that of the direct sound as it has traveled further [26]. Another factor for changes in the reverb spectrum is the materials that are present in the listening environment. Absorption and diffusion characteristics of different materials are also frequency dependant [4]. Reverberant sound that is reflected off a certain material may therefore have a significantly different energy spectrum than the direct sound. The spectral characteristics does not form the basis for the distance cues as discussed by Durlach and Colburn, however, when combined with a listener's previous experience of an environment it may improve the accuracy of the distance perception [8].

2.2 Inside-the-head locatedness when using headphones

Realistic reproduction of spatial audio is dependant on the ability to reproduce the tonality and localisation of an auditory event as if the sound source was placed in the real world. If a sound event from a real world sound source results in a certain wave field at the listener's ear drum, a hypothetic perfect set of headphones for spatial audio reproduction is one which is able to exactly replicate the audio wave field at the ear drum [6]. Such a perfect set of headphones currently do not exist, so to get as close as possible it is important to consider how the location of a sound source is perceived by the listener. The auditory localisation system is not

easily tricked. Failing to do so accurately will result in inside-the-head locatedness, which is a common occurrence when using headphones [26].

The human auditory system is very sensitive to changes in the ILD, as discussed in section 2.1.2. It is therefore of great importance that the level sensitivity of the left and right headphone driver remains the same. Sensitivity disparity results in the output being louder to one side than the other, thus in turn resulting in an inaccurate perception of the azimuth angle. An investigation by Gutierrez-Parera and Lopez concluded that driver sensitivity disparity started affecting the localisation perception from a level of 1dB [13]. They furthermore found that the perceived spatial accuracy had high correlation with the perceived sound quality of a set of headphones.

The headphone frequency response must not be coloured and must be similar for both the left and right driver. Spectral cues are an important factor for sound localisation as discussed in section 2.1.2. A non-uniform frequency response will result in sound quality that is perceived as being coloured rather than neutral. This may affect the monaural cues thus confusing the localisation perception. The headphone frequency response must therefore not contain any sudden peaks or notches. According to the investigation by Gutierrez-Parera and Lopez a non-uniform frequency response results in a significant decrease in perceived spatial accuracy [13]. Irregularities in the frequency response between 100Hz and 1600Hz resulted in greater front-back confusion. Imperfect frequency response between 4000Hz and 7000Hz, on the other hand, was found to degrade the accuracy of azimuth perception.

Spectral changes to the sound caused by the transmission from electric signal, which is sent to the headphone driver, to the sound at the ear drum is known as the Headphone Transfer Function (HPTF) [23]. The HPTF depends both on the frequency of the headphones and factors that vary for each individual listener.

The sound output of the headphone is transduced by the headphone driver, but is then affected by the size and shape of the ear cup space both behind and in front of the driver unit and the shape and material of the ear cushion [4]. All these elements are designed and specified by the headphone manufacturer to output a desired output response. Engineering tolerances, especially in the driver unit, cause slight differences in the frequency response even when comparing headphones of the same model.

The HPTF furthermore is highly dependent on the following factors that all varies for each individual listener:

- Shape and size of listener's pinna and ear-canal
- Distance from driver to the entrance of the ear-canal
- Cushion leak

- Headphone placement

The listener's pinna and ear-canal creates reflections of the sound which influence the frequency response measured at the ear-drum [9]. The shape and size of the pinna and ear-canal will therefore affect the behaviour of the reflections, and therefore also the frequency response measured at the ear-drum.

Distance from the driver unit to the ear-canal entrance is largely the result of three factors; headband clamping force, stiffness of the cushion and the size of the listener's head [4]. This distance will affect the size and shape of the room between the ear-canal and driver unit, so how the reflections behave in this room is also affected.

Bass response of closed back circumaural headphones very much depends on a complete seal between the headphone cushion and the listener's head [13]. A broken seal will result in a drastic decrease in magnitude at the low end of the frequency spectrum. This can be a problem if the listener has a lot of hair or is wearing glasses, but is especially problematic for on-ear headphones since they can be difficult to place on the ear.

Using the blocked ear canal method discussed by Møller et al it is possible to estimate the HPTF for an individual listener wearing a certain pair of headphones [24]. However, they also found that the HPTF varies slightly each time the listener puts on the headphone. This is because the HPTF is dependant on the precise location of the headphone driver relative to the entrance of the ear canal. This may result in an inconsistent user experience where the headphones will sound good to the listener some times and not so much other times.

2.3 Solutions to avoid inside-the-head locatedness

2.3.1 Sound level compensation

It is expensive for headphone producers to keep sensitivity tolerances below 2dB [4], so sensitivity disparity less than 1dB is usually reserved for expensive high end headphones. Sensitivity measurements are commonly performed using test setups where either a single microphone is utilised to measure the two sides alternately, or two microphones are used to measure both sides simultaneously.

A simple but resource intensive solution to mitigate sensitivity disparity is to measure the driver units and match them pairwise, so that a driver is matched with a driver that exhibits similar sensitivity characteristics [4]. Such a procedure creates a bottleneck in the production line and require extra staff. It is for these reasons only a reasonable solution for high-end headphones with low production numbers.

If the level of sensitivity disparity is known it is alternatively possible to adjust the volume levels at each side to equalise the disparity. This solution may be

appropriate for headphones where a Digital Signal Processor (DSP) and a Digital-to-Analog converter (DAC) is embedded in the headphone. Otherwise the volume adjustment would need to be implemented on all devices the headphone would be connected to.

2.3.2 Frequency response compensation

A set of headphones may be calibrated by the manufacturer to exhibit a natural and uniform frequency response. however, the frequency response that is perceived by the listener depends on multiple individual factors that vary from listener to listener as discussed in section 2.1.2. A system that aims to provide a uniform and consistent frequency response to all users has to compensate for these inconsistencies [19]. Such a system relies on three main subsystems each of which are discussed in this section.

Frequency response analysis

To perform accurate frequency compensation it is naturally important to determine the frequency and magnitude of any inconsistencies. Frequency response analysis is commonly achieved by completing three steps as seen below [28]:

1. Playback of stimulus signal
2. Recording of stimulus response
3. Analysis of recording

Different types of stimuli signals may be used depending on the purpose of the frequency response analysis. For analysis of audio equipment one of the most common approaches is to utilise a logarithmic sine sweep as proposed by Farina [11]. A logarithmic sine sweep is a pure sine tone the frequency of which is increased logarithmically over time. This means that the power spectrum of the signal is reduced by 3dB per octave. In contrast a linear sine sweep exhibits a power spectrum that is completely flat. The benefit of utilising a logarithmic sweep rather than a linear sweep is improved signal-to-noise ratio and improved suppression of pre-ringing [11]. For certain applications it is more appropriate to use noise signals as stimuli for a frequency response analysis [25]. Very high frequency tones at the end of a sine sweep can be annoying if a test participant is wearing headphones. This is not a problem when pink noise is used instead of a sine sweep. The tradeoff, however, is that measurements done using sine sweeps produce results that generally are more accurate. Pink noise exhibits the same 3dB power spectrum reduction per octave as the logarithmic sine sweep, and is therefore preferred rather than white noise. Another option is to utilise an arbitrary stimulus signal. Such signals may

not cover the entire frequency spectrum, however if the recorded stimulus signal is compared to the original signal it is possible to establish a rough estimate of changes to the frequency spectrum. This is useful for real-time applications where it is not possible to use either noise or sine sweep stimuli. An investigation by Liski et al concluded that it was possible to achieve results using arbitrary signals such as rock and pop songs that was comparable with measurements done using logarithmic sine sweeps [19].

While the stimulus signal is being played through the headphone driver units it is simultaneously being recorded by a measurement microphone. However, recording the stimulus in a way that provides accurate results is not a trivial task. Microphones themselves has frequency responses which if not taken into consideration may interfere with the accuracy of the measurement. This problem can be mitigated either by using a calibrated microphone or a microphone with a known frequency response. Measurement microphones such as the MiniDSP UMIK-1[1] are calibrated by the manufacturer to provide a near flat frequency response between 20Hz-20.000Hz. However, they are physically larger and more expensive relative to less precise microphones. A less precise microphone may alternatively be used in cases where the frequency response of the microphone is known[19]. In such cases the measured response can be filtered to equalise the microphone's effect on the measurement.

After the stimulus response has been recorded the next step is to analyse the recording by transforming it to the frequency domain, which is done using the Fourier Transform[11]. Computationally efficient Fast Fourier Transform (FFT) algorithms such as the Cooley-Tukey makes it possible to quickly compute the frequency spectrum of any given audio signal [10]. In broad strokes the Cooley-Tukey algorithms works by dividing the recorded signal into progressively shorter intervals through iterative steps until the length of each interval is just one sample. The time domain value of each sample is then equal to the frequency domain value. The iterative steps are then reversed to reconstruct the signal in the frequency domain.

Filters

Digital filters are fundamental to Digital Signal Processing (DSP). In general a filter can be defined as a system that removes components or in some way modifies the characteristics of a signal [20]. Two types of filters may be used for DSP applications; Finite Impulse Response (FIR) or Infinite Impulse Response (IIR) filters. Both FIR and IIR filters are Linear Time-Invariant systems. A system may be classified as linear if all frequency bands in the input signal are represented in the output of the system. Time-Invariant means that when the input is shifted by n number of

[1] www.minidsp.com/products/acoustic-measurement/umik-1

samples, the output of the system will be shifted by the same number of samples.

A filter may be defined as a FIR system if the output impulse response is finite in length [20]. FIR filters often consist exclusively of feedforward coefficients, but may include feedback coefficients as long as the length of the output is equal to the length of the input. A great benefit of FIR filters is that it is easier to achieve a linear phase response, which means that changes to the phase are independent of the signal frequency.

IIR filters, on the other hand, is a designation given to filters that produces impulse responses with a length that is infinite. These filters always include a nonzero number of feedback coefficients and are therefore often called recursive filters. It is often possible to construct IIR filters that are more accurate or more computationally efficient than comparable FIR filters [20]. However, it is difficult to construct IIR filters that produce linear phase responses. It is furthermore important to take stability into account when constructing IIR filters. The feedback coefficients makes it possible to create filters that has impulse responses that are perpetually rising in amplitude similar to the howling noise that is created when placing a microphone in front of an amplified loudspeaker.

Filters used for audio equalisation applications can either be implemented as a parametric or graphic equaliser [27]. Parametric equalisers are quite flexible as they allow the user to control the centre frequency, bandwidth and gain independently. This makes parametric equalisers well suited for applications where a high degree of precision is required. Graphic equalisers consist of a set of filters the centre frequency of which are placed at pre-determined intervals. The user is only able to control the gain of each filter in the graphic equaliser. Graphic equalisers are therefore simpler to operate and are well suited for wide-band equalisation where a lower level of precision is tolerated.

Adaptive algorithms

The third sub-system required to implement a frequency compensation filter is an adaptive algorithm that adjusts the coefficients of the equaliser based on the measured frequency response. The compensation system proposed by Liski et al utilise a Least Mean Square (LMS) algorithm to continuously update the filter coefficients [19]. The overview of a simple LMS implementation can be seen in figure 2.2. The purpose of the illustrated implementation is to estimate the coefficients of the S(z) filter. By comparing the output of a secondary filter called Sh(z) with that of the S(z) filter an error signal is calculated. Equation 2.2 is then used to calculate a set of updated coefficients for the Sh(z) filter based on the values of the error signal [19]. In this equation the W is the set of coefficients, e is the error signal, x is the noise signal, n is the iteration number and μ is the step size. The sequence illustrated in figure X runs continuously so that the accuracy of the Sh(z) filter increases as the error signal is reduced.

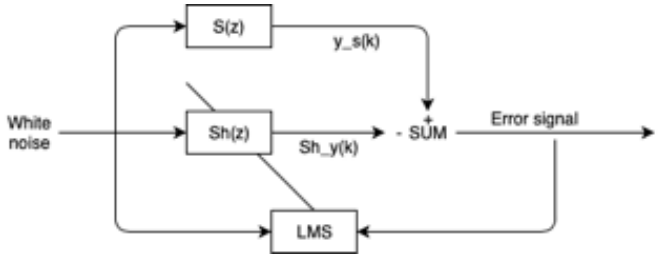


Figure 2.2: Simple implementation of a LMS algorithm

$$eq.X : W(n+1) = W(n) + \mu x(n)e(n) \tag{2.2}$$

The performance of different variations and implementations of LMS algorithms are often evaluated based on measurements of accuracy, stability and convergence time [29]. Over time the error signal will stabilise and will not be reduced significantly any further. The amount of time it takes to reach this error level is known as the convergence speed. Since the updated filter coefficients depends on the input signal a sudden change in the input could upset the system. The step size μ is included in equation 2.2 to mitigate the chance of this happening, however, it is still relevant to test the stability of the algorithm [29].

2.3.3 HRTF replication

When both the HRTF and HPTF is known it is theoretically possible to perfectly replicate the monaural cues of a sound as measured at the ear drum. If a perfect replication of binaural and monaural cues are achieved the localisation perception of the sound will be completely indistinguishable from a real world version of the same sound [23]. However, both HRTF and HPTF depends greatly on the individual listener as discussed in section 2.1.2. HPTF furthermore depends on the precise placement of the headphones, while the HRTF depends on the listening environment and the relative location of the sound source. Great care must therefore be taken to obtain accurate HRTF and HPTF measurements.

Measurements of the HPTF needs to be done while the listener is wearing the headphones. A measurement microphone is therefore placed either as close to the listener's ear drum as possible or at the ear canal entrance. Placing a microphone at the listener's ear drum is impractical for several reasons [23]. It is often uncomfortable for the listener to get things placed so far inside the ear canal. The physical size of the microphone furthermore changes the acoustic properties of the ear canal, thus resulting in inaccurate measurements. The alternative is to place the microphone at the entrance of the ear canal and then block the ear canal completely [9]. The ear canal will then not affect the measurement at all. After the blocked ear

response is measured an estimate of the ear canal response may be added based on measurements of the ear canal properties such as length and diameter.

The blocked ear microphone placement is also utilised for HRTF measurements. However, instead of the stimulus signal being played through a set of headphones it is played through a loudspeaker placed at certain positions around the listener [26]. HRTF measurement standards such as CIPIC and ISVR utilise a rotating loudspeaker setup where the listener remains stationary [1][30]. Other HRTF measurement methods including IRCAM and MARL utilise loudspeakers that remain stationary while the listener is rotated [7][2]. The location of the loudspeakers relative to the listener change position because each personal HRTF consist of a series of measurements. Each measurement is done with the loudspeaker placed at a specified azimuth and elevation. The azimuth and elevation angles are chosen so that they make an even sphere around the listener's head [26]. Most HRTF measurements are done far-field, meaning that the distance between listener and loudspeaker is at least 1 meter.

By inverting the HPTF it is possible to create a filter that will ensure a near flat response at the ear drum. An investigation by Brinkmann and Lindau showed that it was possible to achieve a flat response up to roughly 10kHz at the ear drum [5]. This by itself will no sound natural to the user, however if the appropriate HRTF is then added to the sound it is possible to accurately replicate the monaural cues. However, inaccurate compensation may reduce the spatial accuracy of the system rather than improve it. Since the HPTF changes slightly depending on the precise placement of the headphones, regularisation methods have been implemented to take these slight changes into account [18].

2.4 Personalised audio in headphones

Unfortunately the processes involved in obtaining both the individual HRTF and HPTF makes it unpractical for consumer oriented products. Some manufacturers has included features that personalises the sound of the headphones. The AKG N90Q[2] are able to measure the impulse response inside the ear cup, and will then adjust the frequency response to compensate any non-uniformities. However, they are not able to measure the complete HPTF.

Some other headphone manufacturers such as Beyerdynamic[3] utilise hearing tests similar to the audiometry tests audiologists perform to adjust the sound of the headphone to the individual listener. The benefit of these tests is that it is possible to take the individual listener's hearing ability into account and the headphone frequency response. Although, it requires the user to sit in a quiet environment and spend up to 10 minutes of complete concentration to complete the test.

[2] akg.com/n90q

[3] north-america.beyerdynamic.com/kopfhorer-headsets

Graphic equalisers are another common way to personalise the frequency response of a pair of headphones. Applications such as the Beoplay App[4] present the user with a Graphic User Interface (GUI) where it is possible to change the characteristics of frequency response. The user inputs are then translated to gain values that are used to adjust a graphic equaliser. Users are thereby able to obtain their desired response, however, the user must be willing to spend time on making fine adjustments. The user may furthermore not know what their desired response is and may not know how the controls provided by the GUI can be used to improve the sound quality.

[4] bang-olufsen.com/en/apps/bang-olufsen-app

Chapter 3

Ear frequency response study

3.1 Purpose

Different parts of the human ear affect different parts of the frequency spectrum as discussed in section 2.1.2. A short study of the frequency response of a synthetic ear was therefore performed to separate the frequency response of the pinna from the frequency response of the ear canal. This information was important to evaluate whether the frequency response compensation system was able to compensate for both the pinna and ear canal effects. Synthetic ears such as those used on Head And Torso Simulators (HATS) do not provide a completely realistic frequency response compared to measurements from real persons [23]. However, it may be used to establish a rough estimation of which frequency bands are affected by the pinna, and which are affected by the ear canal resonances.

If the frequency response of the entire ear is known and the pinna response, it is possible to split the response into two separate frequency responses; the one caused by the pinna and one caused by the ear canal [23]. This may be achieved by simply removing the pinna response from the entire ear response as seen in equation 3.1 and 3.2.

$$R_e = R_p + R_c \tag{3.1}$$

$$R_c = R_e - R_p \tag{3.2}$$

where R_e is the response of the entire ear, R_p is the response of the pinna and R_c is the response of the ear canal. The goal of the procedure was therefore to measure the response of the entire ear response including ear canal of the synthetic ear and a blocked ear measurement that only included the response of the pinna.

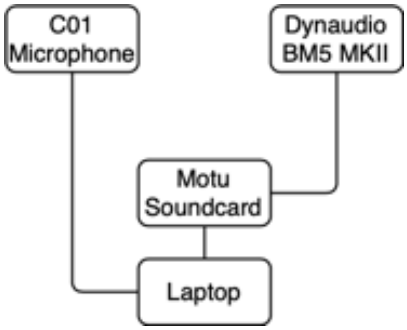


Figure 3.1: Diagram showing the components used for the ear response study

3.2 Setup

The setup for the ear frequency response study consisted of several components as can be seen in figure 3.1. A laptop running a frequency response analysis application called Room EQ Wizard[1] was connected to a Motu UltraLite-MK3 soundcard[2] which in turn was connected to a Dynaudio BM5 MKII loudspeaker[3]. An AIAIAI C01 cable[4] was modified to remove one of the audio connectors and place the microphone in a synthetic ear taken from a B&K type 4128-C HATS system[5]. This microphone was then connected directly to the laptop through the 3,5mm audio jack.

It was decided to perform the measurements of the study in a free field environment. Had headphones been used the HPTF would affect the measured response. To obtain the free field measurements the loudspeaker was placed in an anechoic chamber at Aalborg University Copenhagen. Anechoic characteristics of the chamber meant that the measurements was not affected by room resonances or any other noise interference. The loudspeaker was placed at a distance of 1 meter from the synthetic ear at three different locations as seen in figure 3.2. The three angles that was chosen for the loudspeaker placements were 0 degrees, 90 degrees and 180 degrees relative to the entrance of the ear canal. Such loudspeaker placements made it possible to determine an averaged response that is less dependant on the location of the sound source relative to the ear. A loudspeaker was not placed at 270 degrees because the study was done using a single synthetic ear rather than an entire HATS system.

In terms of frequency response the Dynaudio BM5 MKII provides a flat re-

[1] roomeqwizard.com
[2] www.motu.com/products/motuaudio/ultralite-mk3/hybrid
[3] www.dynaudio.com/professional-audio-discontinued/bm-series/bm5a-mkii
[4] aiaiai.dk/headphones/tma-2/parts/cables/c01
[5] www.bksv.com/en/products/transducers/ear-simulators/head-and-torso/hats-type-4128c

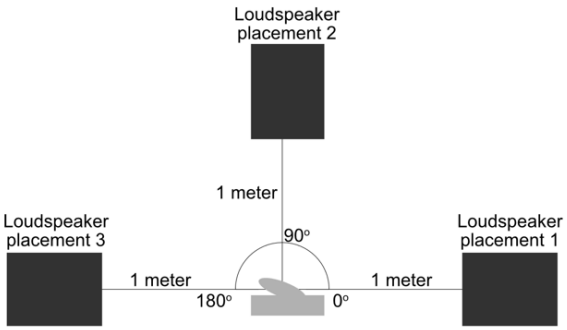


Figure 3.2: Diagram showing the setup of the loudspeakers for the ear response measurements

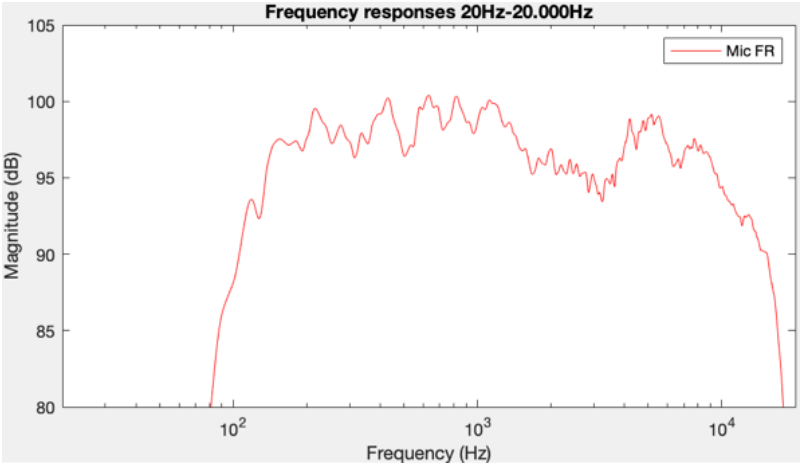


Figure 3.3: Measured microphone frequency response

sponse in the 48Hz-21.000Hz range with a +-3dB tolerance[6]. However, the response of the C01 microphone was not known, so a preliminary frequency response measurement was conducted to record the combined frequency response of the loudspeaker and microphone. The resulting response can be seen in figure 3.3. This frequency response is fairly even in the 100Hz-10.000Hz range, so further analysis within this range may be done without compensating for the loudspeaker and microphone response, which otherwise would add complexity to the analysis procedure.

The synthetic ear setup was based on an ear taken from a B&K type 4128-C HATS system as mentioned earlier. However, a few custom elements were added to the ear to make it easier to use it for frequency response measurements. Images of the assembled ear setup can be seen in figure 3.4. Two plates of acrylic were laser cut and mounted on a metal bracket to keep the ear on an upright position.

[6]www.dynaudio.com/professional-audio-discontinued/bm-series/bm5a-mkii

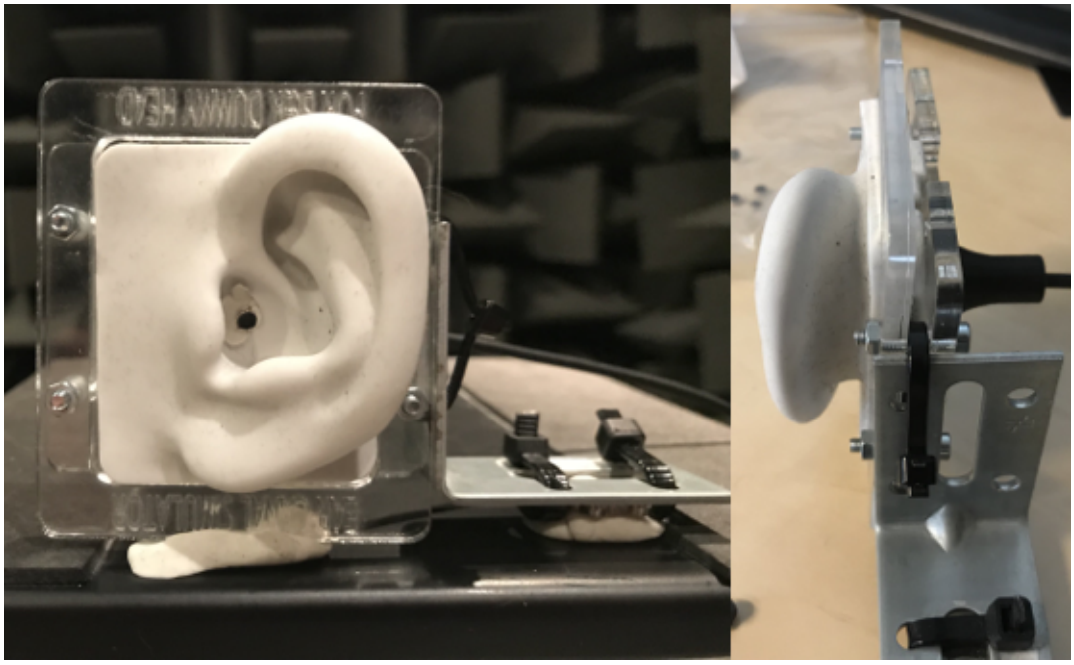


Figure 3.4: Synthetic ear setup including 3D-printed and lasercut elements

A 3D printed extension was furthermore added to the ear canal of the synthetic ear. This made it easier to mount the microphone in the ear canal. It also made it possible to mount the microphone 25mm from the the entrance of the ear canal, which is roughly the length of an average adult ear canal [23]. The STL files for the acrylic plates and ear canal extension can be found in Appendix A.

A logarithmic sine sweep was used for the measurements stimulus since it provides accurate frequency response measurements as discussed in section 2.3.2. This sine sweep was 256k samples long and and covered the 20Hz-22.000Hz range. It was generated on the laptop using Room EQ Wizard, which also was used for recording and analysis of the response from the C01 microphone.

3.3 Procedure

The study procedure consisted of two sets of three measurements each. One frequency response measurement was taken using each of the loudspeaker placements indicated in figure 3.2. After a measurement was taken the stand, on which the ear setup was mounted, was rotated 90 degrees to prepare the setup for the next measurement. The first set of measurements was gathered with the microphone placed at the ear drum position, so it was placed in the ear canal 25mm from the entrance. After all three measurements with this setup was gathered the procedure was repeated with the microphone placed at the blocked ear position.

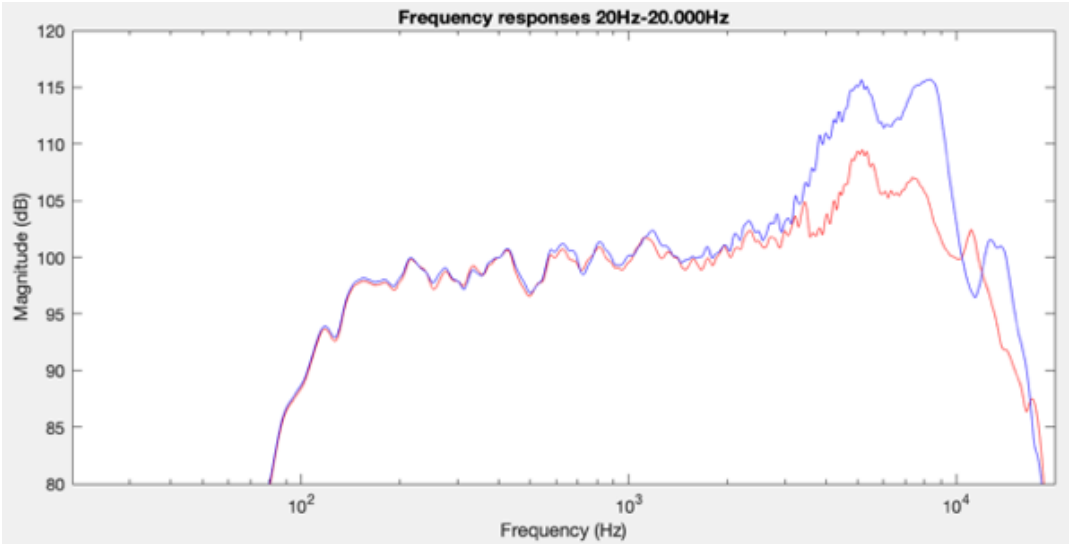


Figure 3.5: Raw measured frequency response of pinna (red) and entire ear (blue)

An average of the thee measurements in each set was then calculated in Room EQ Wizard. The average of the measurements with the ear canal included and the average of the blocked ear measurements were exported as a text file and imported into Matlab for further analysis. The response of the ear canal could then be calculated by removing the pinna response from the ear response.

3.4 Results

The raw frequency responses of the isolated pinna and entire ear can be seen in figure 3.5. The two responses are close to identical outside of the 4.000Hz-10.000Hz range, and both feature pronounced peaks. Three peaks of varying magnitudes can be observed in the pinna response. The total ear response consist of two peaks with a magnitude of roughly 15dB relative to the magnitude at 1.000Hz.

The compensated versions of the frequency responses can be seen in figure 3.6. The same peaks can be observed, but the magnitudes of the peaks are changed slightly. The peaks were amplified generally by roughly 5dB more than the raw versions. The compensated pinna response peaks furthermore exhibit the same magnitude, which was not the case for the raw response.

The ear canal response that was calculated is mostly present in the 4.000Hz-10.000Hz range, where the magnitude of it varies 5dB-10dB. Pronounced peaks in the ear canal response can be observed at roughly 4.000Hz and 10.000Hz.

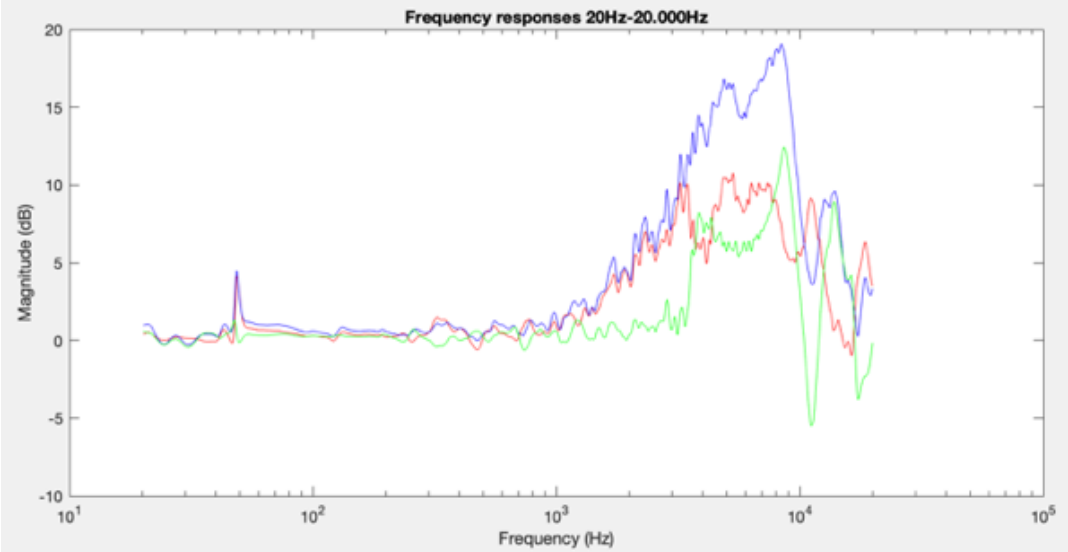


Figure 3.6: Compensated frequency response of pinna (red), entire ear (blue) and calculated ear canal response (green)

Chapter 4

Implementation

4.1 Requirements

The primary purpose of the project discussed in this paper is to test the feasibility of the system proposed in section 1.3. A prototype system consisting of both hardware and software components was therefore developed. Before this prototype system was developed a set of requirements were established to ensure that the system was able to achieve the goals of the system proposal. Three aspects were considered when the requirements were established; the goals and overall purpose of the system proposal, information and data gathered from research discussed in section 2 and data gathered through the ear frequency response study in section 3. As the prototype system consisted of both hardware and software components the requirements were divided into two sets. Both sets of requirements were established before the development of the prototype was initiated, however, the software requirements in particular were updated throughout the development process.

The final set of hardware requirements included four points:

1. Prototype headphones must be circumaural
2. Must include one microphone in each ear-cup
3. Microphones must be placed centrally on the plate that separates the driver and the listener's ear
4. Must include interface able to handle 2 input channels and 2 output channels

The scope of the projected was limited to circumaural headphones as discussed in section 1.2. So, the headphone on which the hardware prototype component was based had to be a circumaural type headphone. A benefit of using circumaural headphones is that they are generally larger than supraaural headphones, and therefore provides more space in the ear-cup behind the driver and between the

driver and listener's ear. This additional space was important for the modifications that was done to the prototype headphones.

To achieve the real time compensation of sound level and frequency response, which is a crucial factor for the system proposed in section 1.3, a microphone must be placed in each ear-cup. Without these microphones nothing is known about the sound pressure and frequency response in the space between headphone driver and ear.

It was furthermore established that the microphones must be placed inside the space between the headphone driver and listener's ear. To get the microphone as close to the entrance of the ear canal as possible the microphone had to be placed centrally on the plastic plate that protects the headphone driver. This placement furthermore ensured that the microphone did not touch the listener's ear, which could introduce unwanted noise and discomfort for the listener.

The additions of the microphones in the ear cups meant that the system required 4 audio channels in total. Two input channels sent audio signals to the headphone drivers for playback, while two output channels sent audio signals from the microphones to the laptop where the algorithm was running. To keep the hardware system as simple as possible it was decided that the audio interface had to be integrated into the headphone prototype. Such an arrangement meant that the prototype was independent of an external audio interface, which otherwise would be cumbersome during the evaluation process.

The final set of software requirements can be seen below:

1. Must compensate for divinations in sound level
2. Must include wide band compensation in 50Hz-1.600Hz range
3. Must include peak compensation in range 4.000Hz-10.000Hz
4. Must be able to compensate sound level and frequency response using arbitrary input signal
5. Wide band frequency compensation and sound level compensation must run in real-time
6. Must output centre frequency and gain of 1st, 2nd and 3rd HRTF peaks

Disparity between the sound level in the left and right side of the headphone may confuse the localisation perception as discussed in section 2.1.2. One of the objectives of the algorithm is to compensate for this disparity, so that the sound pressure is equal on each side given the same input signal.

The study by Gutierrez-Parera and Lopez furthermore found that poor frequency response in the 100Hz-1.600Hz had averse effects on localisation perception accuracy [13]. Software requirement number 2 was therefore established.

Requirement number 3 was established to ensure compensation of the frequency response at the higher end of the frequency spectrum. This requirement was also established based on the Gutierrez-Parera and Lopez study, which found that irregular frequency response between 4.000Hz to 7.000Hz affected the lateral accuracy of localisation perception [13]. This range was widened to compensate for the three ear response peaks discussed by Iida et al [15]. The 4.000Hz-10.000Hz range was determined as this is the range where the peak centre frequencies are found.

To achieve good accuracy each time the listener is wearing the headphones the system must readjust the compensation filters [9]. A dedicated stimulus signal such as logarithmic sine sweep or pink noise would degrade the user experience, so the compensation system must be able to adjust the filters based on an arbitrary signal. This arbitrary signal could be a music track or binaural audio.

Both the wide band frequency compensation (100Hz-1.600Hz) and the sound level compensation must be run in real-time. This requirement was established to achieve a good level of accuracy even if conditions change slightly. An ideal real-time system is completely transparent to the listener. The reason that the high frequency compensation (4.000Hz-10.000Hz) is not included in this requirement is that the high frequency range is vulnerable to noise, so a real-time system in that frequency range was found to be unstable.

The peak detection system must not only be used for frequency response compensation, but also to output the estimated centre frequency values and magnitudes. These values may then be used to estimate the listener's personal HRTF using another system. Such a HRTF estimation system is not within the scope of this project, however, the peak value output makes it a possibility in the future.

4.2 Hardware

The headphone prototype developed for this project was based on a set of TMA-2 headphones made by AIAIAI[1]. The TMA-2 is a modular system which allows for a variety of components to be combines to form a set of headphones. For the prototype a specific combination of components was chosen; H04 headband, S04 speaker units and E05 ear pads. The headband did not affect the functionality of the prototype, so the H04 was chosen as it is sturdy and comfortable to wear while it produces an adequate amount of pressure. The S04 speaker unit produce a dynamic V-shaped frequency response, which makes it suitable for EQ adjustments since it is possible to play great amounts of treble or bass without introducing distortion. Speaker units that produce a more neutral frequency response are not as flexible since boosting the magnitude of the frequency response introduces much more non-linear distortion than attenuating the response. The E05 ear pad was

[1] aiaiai.dk/headphones/tma-2

Figure 4.1: Overview of the prototype headphone

chosen primarily because it is a circumaural type ear pad, and secondly because it provides a good amount of sound isolation. An overview of the assembled prototype can be seen in figure 4.1.

After the headphone components had been chosen and assembled a series of modifications were done to ensure that the prototype headphone fulfilled the hardware requirements. A MEMS microphone was mounted centrally on each speaker plate as seen in figure 4.2. These two MEMS microphones were taken from a set of AIAIAI Pipe 2.0 and Tracks 2.0[2] respectively. Care was taken when gluing the microphones in place so that the ear pads could be attached without creating a leak between the microphone circuit board and ear pad. The audio output from the USB-C connector was then soldered to the S04 driver unit as seen in figure 4.2. The audio connectors of the H04 was furthermore removed from the headband as they were no longer used.

Microphones from Pipes 2.0 and Tracks 2.0 was chosen since both of those headphones use USB-C connectors. Using USB-C connectors it was possible to connect 2 output channels and 2 input channels to a MacBook Pro (2017) laptop[3] simultaneously, thus fulfilling hardware requirement number 4. The Pipes 2.0 and Tracks 2.0 are completely identical from the microphone and down, but they were registered as two different audio interfaces by the laptop when connected. When connecting two Pipes 2.0 or two Tracks 2.0 the laptop only registers a single audio interface.

[2] aiaiai.dk/headphones/made-for-google
[3] apple.com/macbook-pro/

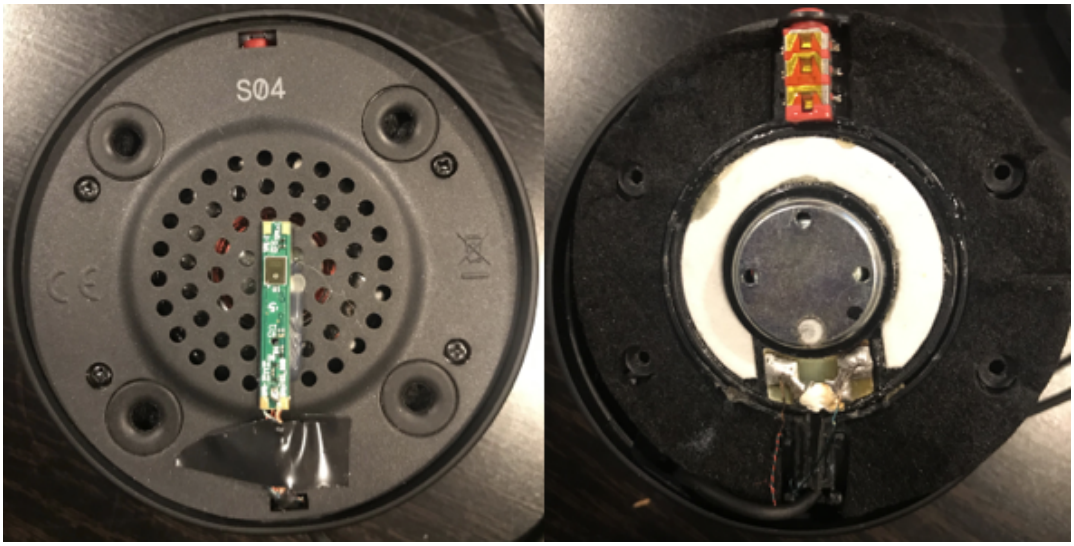


Figure 4.2: Modifications made to the prototype headphone speaker unit

4.3 Software

4.3.1 System overview

The software component of the prototype was entirely developed and implemented in Matlab Simulink. Since the general purpose of the software prototype system was to perform three different tasks, it was decided to divide the system into three subsystems. A complete overview of the three subsystems can be seen in figure 4.3. The content of the green box represents the sound level compensation subsystem, the blue box represents the graphic EQ subsystem and the yellow box the peak EQ subsystem.

Audio signals were routed from left to right, so first step was to adjust the sound levels to compensate for deviations in sensitivity. The graphic EQ subsystem was then used to adjust the frequency response of the audio signal in the 50Hz-1.600Hz range. The third and last subsystem was the peak EQ, which compensated the frequency response in the 4.000Hz-10.000Hz range and determined the parameters of the peaks caused by the listener's ear response. The entire software implementation can be found in appendix B.

4.3.2 Sound level compensation

The first of the subsystems was the sound level compensation subsystem, the purpose of which was to fulfil software requirement number 1. An overview of the final Matlab Simulink implementation can be seen in figure 4.4.

A Matlab function was connected to two audio file inputs, two microphone

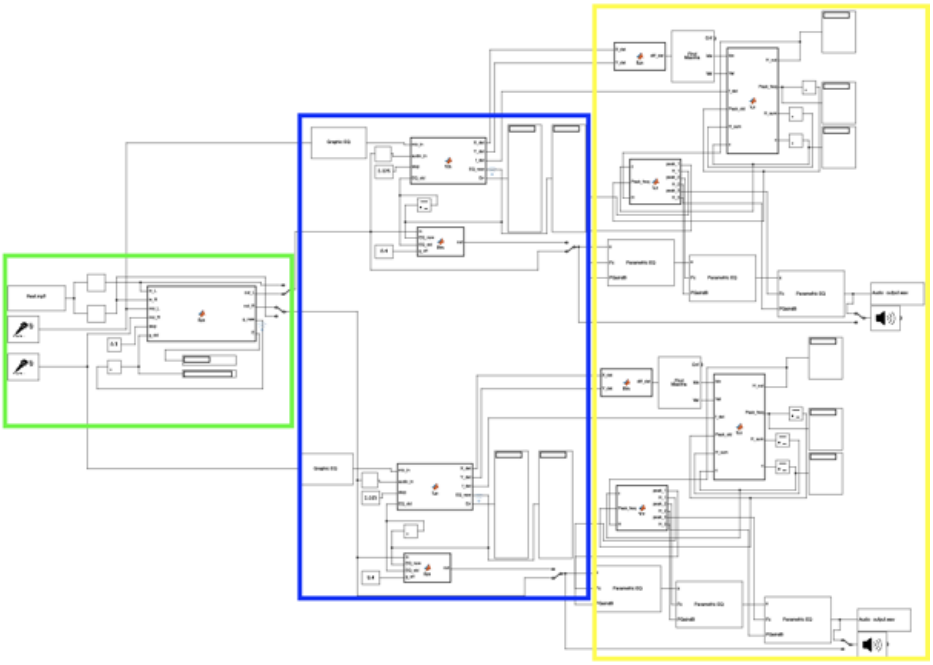


Figure 4.3: Overview of the prototype software implementation in Matlab Simulink

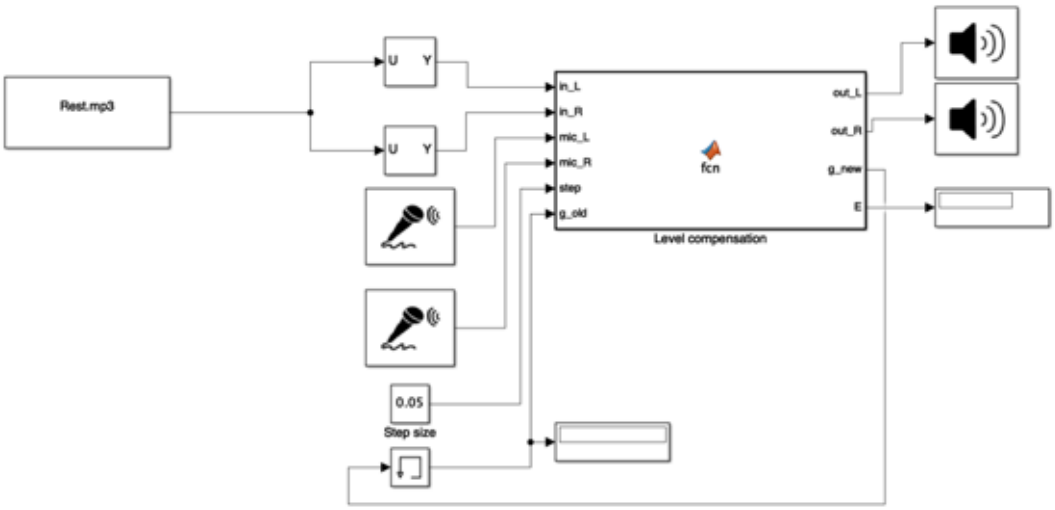


Figure 4.4: Overview of the sound level compensation subsystem in Matlab Simulink

```
function [out_L,out_R,g_new,E] = fcn(in_L,in_R,mic_L,mic_R,step,g_old)

% Make sure g_old is a number value
if isnan(g_old) == 1
    g_old = 1;
end

% Determine length of input signals
N = length(in_L);

% Calculate power of each input channel in dB
in_L_p = 10*log10(sum(in_L(1:N,:).^2)/N);
in_R_p = 10*log10(sum(in_R(1:N,:).^2)/N);
mic_L_p = 10*log10(sum(mic_L(1:N,:).^2)/N);
mic_R_p = 10*log10(sum(mic_R(1:N,:).^2)/N);

% Determine power ratio between left and right input channels
in_ratio = in_L_p/in_R_p;
mic_ratio = mic_L_p/mic_R_p;

% Determine how much gain to apply to left channel to achieve correct power
% ratio
E = mic_ratio-in_ratio;

% account for step size
g_new = g_old + step * E * mic_ratio;

% Determine output
if g_new<1
    out_R = in_R;
    out_L = in_L * g_new;
elseif g_new>1
    g = 1/g_new;
    out_R = in_R*g;
    out_L = in_L;
else
    out_R = in_R;
    out_L = in_L;
end
```

Figure 4.5: Content of the Level compensation function

signal inputs and two audio outputs. The two output signals was sound level compensated versions of the two audio file signals. Adjustments to the gain of the two output channels was dependant on the amount of energy carried by each of the four input signals. The contents of the Matlab function can be seen in figure 4.5.

Since the algorithm is implemented in Matlab Simulink[4] it will run in real-time, meaning that the audio input signals were sent to the function as frames of samples. A frame size of 32768 samples was chosen to make the subsystem compatible with the graphic EQ and peak EQ subsystems.

The first step in the level compensation function was to make sure the g_old input is a number value. The g_old was the gain value calculated during the previous iteration of the function. For the first iteration of the function the g_old equaled 1, so the two output signals would be exactly equal to the two audio file input signals. If the g_old value was not a number it would be reset to equal 1.

After determining the length of the audio file input signal, the ratio of power between the left and right side of the audio signal and the microphone signal was defined. The difference between the two ratios was then calculated and used as an error value. This error value would be equal to zero if the sensitivity of the two

[4] se.mathworks.com/products/simulink

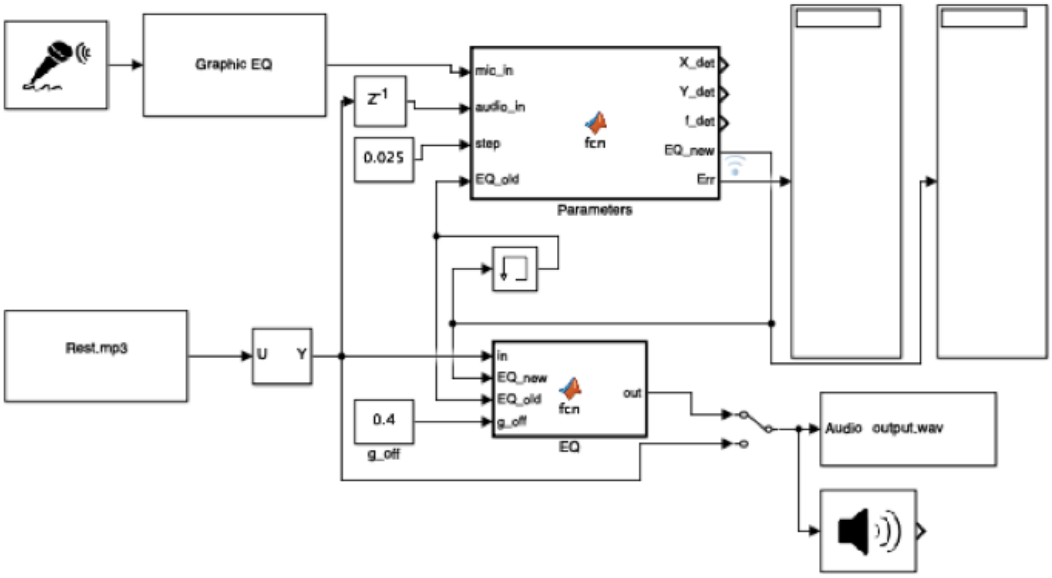


Figure 4.6: Overview of the graphic EQ subsystem in Matlab Simulink

headphone driver units was perfectly matched. A new gain value g_new could then be calculated based on the previous gain value, step size, error value and the power ratio of the microphone input frames.

The last step of the level compensation function was to determine the output signals based on the new gain value. If the g_new value was smaller than 1 it meant that the left side was too loud, while a g_new value greater than 1 indicated that the right side was too loud. The signals was only attenuated to reach the correct power ratio, they were never amplified. Amplifying a digital signal will cause distortion if the sample values are greater than 1.

Outputs of the level compensation function included the two audio output channels, the error value and the new gain value. The error value was displayed for visual inspection by the operator of the system, while the gain value was looped back into the function through a memory block.

4.3.3 Graphic EQ

The two audio output channels of the sound level compensation subsystem were then routed to the Graphic EQ subsystem. Two custom Matlab functions formed the basis of the graphic EQ subsystem. One function called parameters was used to calculate the filter coefficients, while the graphic EQ filter was implemented in the EQ function. The subsystem showed in figure 4.6 was connected to a single microphone and audio file channel, so the subsystem was duplicated for the other side of the headphone.

```
% Calculate mean magnitude for each EQ bin
for i=1:16
    X_EQdiff(i,1) = mean(X(find(f>fl(i) & f<fu(i))));
    Y_EQdiff(i,1) = mean(Y(find(f>fl(i) & f<fu(i))));
end

% Determine Y array for peak detector
Y_det = Y(find(f>fmax));
X_det = X(find(f>fmax));
f_det = f(find(f>fmax));

% Normalize Magnitudes so 1kHz is 0dB
X_EQdiff(:,1) = X_EQdiff(:,1) - X_EQdiff(14,1);
Y_EQdiff(:,1) = Y_EQdiff(:,1) - Y_EQdiff(14,1);

% Determine error in deviation values
E = X_EQdiff-Y_EQdiff;
Err = immse(X_EQdiff,Y_EQdiff);

% Calculate new parameter set based on the old parameter values, step
% size and error
EQ_new = EQ_old + step * E;
```

Figure 4.7: Content of the parameters function

Before being routed to the parameters function the microphone signal was put through a static graphic EQ, the purpose of which was to compensate for the microphone frequency response. Coefficients for the static graphic EQ was determined through measurements done during the initial part of the development process. The audio file input, on the other hand, was delayed by one frame before being sent to the parameters function.

FFT analyses of the audio input signals were carried out during each iteration of the parameters function. Both frequency domain signals were then windowed to only include meaningful frequencies and the magnitude of the two frequency domain signals were converted to decibels. The two resulting signals were then used to calculate new coefficients for the adaptive graphic EQ as seen in figure 4.7.

Both frequency domain signals were divided into 16 intervals using a for loop, and the mean magnitude value was calculated for each interval. These 16 values were then normalised, so that the magnitude of 1kHz equaled 0. This was done so that the frequency compensation was independent of the input sound levels. An error array was then determined by calculating the difference between the two normalised arrays. If the frequency response of the microphone signal was exactly flat then all values of the error array would equal 0. Each of the filter coefficients in the 50Hz-1.600Hz range were then updated based on the previous coefficient value, step size and error value. For the first iteration of the function the previous coefficients were set to equal 0, which would not affect the audio file signal. The error array and the new filter coefficients were then output for visual inspection.

The new filter coefficients were furthermore sent to the EQ function where they were used to adjust the frequency response of the adaptive graphic EQ. A relatively big frame size of 32768 samples were chosen as mentioned earlier. This was done because the computation time of the FFT analyses in the parameters function. Smaller frame sizes resulted in stuttering audio output because the computation

```
% Calculate mean magnitude for each EQ bin
for i=1:16
    X_EQdiff(i,1) = mean(X(find(f>fl(i) & f<fu(i))));
    Y_EQdiff(i,1) = mean(Y(find(f>fl(i) & f<fu(i))));
end

% Determine Y array for peak detector
Y_det = Y(find(f>fmax));
X_det = X(find(f>fmax));
f_det = f(find(f>fmax));

% Normalize Magnitudes so 1kHz is 0dB
X_EQdiff(:,1) = X_EQdiff(:,1) - X_EQdiff(14,1);
Y_EQdiff(:,1) = Y_EQdiff(:,1) - Y_EQdiff(14,1);

% Determine error in deviation values
E = X_EQdiff-Y_EQdiff;
Err = immse(X_EQdiff,Y_EQdiff);

% Calculate new parameter set based on the old parameter values, step
% size and error
EQ_new = EQ_old + step * E;
```

Figure 4.8: Content of the EQ function

time was longer than the frame size. However, the jumps in frequency response were very audible for the listener if the filter coefficients were updated once per frame size. By dividing the transition from the previous set of coefficients to the new it was possible to make the transition more transparent to the listener. This solution was implemented in the EQ function as seen in figure 4.8.

A for loop with 32 iterations formed the core of the EQ function. First a set of indices with equal spacing was calculated, which was then used to divide the audio input signal into 32 segments. The transition from the old EQ coefficients to the new was likewise divided into 32 steps with equal spacing between each step. A graphic EQ function was called before the initiation of the for loop. The coefficients of this EQ function was set to the current coefficient step. The current input segment was then passed through the graphic EQ filter to calculate the output segment, which was then added to the end of the assembled output signal.

The assembled output signal was finally normalised to the same sound level as the input signal after the 32 iterations of the for loop was completed.

4.3.4 Peak EQ

The third and last subsystem was implemented to fulfil software requirements 3 and 6. In general terms the peak EQ subsystem consisted of three custom Matlab functions, a Find Maxima block and three parametric EQ filters. An overview of the subsystem can be seen in figure 4.9.

The two frequency domain signals calculated in the parameters function of the graphic EQ subsystem was sent to a function called FR_diff. The FR_diff function first normalised the two signals so that they would equal 0dB at 1600Hz, then calculated the difference between the two signals. This difference signal was then sent to a Find Maxima block that finds peaks in the input signal and outputs the index and magnitude of these peaks. Peak indices and magnitudes were sent to a

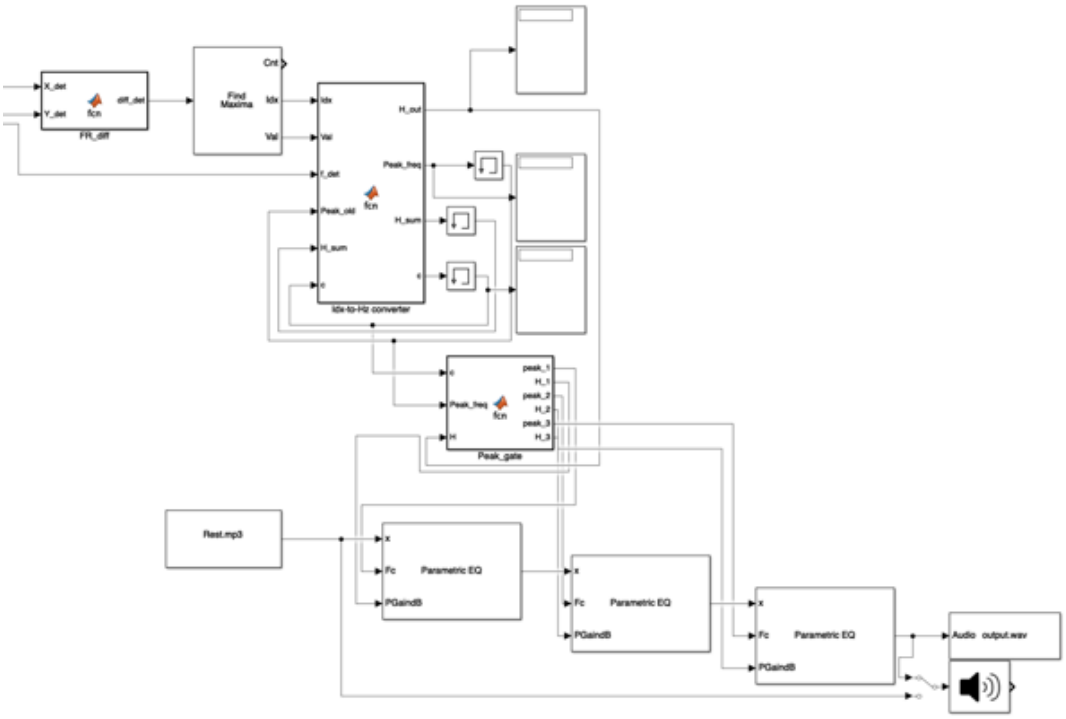


Figure 4.9: Overview of the peak EQ subsystem in Matlab Simulink

```
% Difference between old and new peak
d(1) = peak_current(j,1) - Peak_old(1);

% Calculate average peak magnitude
H_sum(1) = H_sum(1) + Val(j,1);
c(1) = c(1) + 1;
H_avg(1) = H_sum(1) / c(1);

% Calculate new peak frequency
Peak_freq(1) = Peak_old(1) + (peak_current(j,1) * (H_avg(1))/d(1));
```

Figure 4.10: Implementation of the conversion from index values to frequency values

custom Matlab function called Idx-to-Hz converter.

The Idx-to-Hz converter was the core function of the peak EQ subsystem as it was used to convert the indices to values in Hz and calculating the average frequency and magnitude of peaks in the 4.000Hz-10.000Hz range. The conversion from index to frequency can be seen in figure 4.10.

The conversion was based on a f_det array, which represented a linear scale of frequency values. This scale was output from the parameters function of the graphic EQ subsystem and had the same length as the two frequency domain signals. So a frequency value for a certain index value could be found using the frequency scale. The next step of the Idx-to-Hz converter was then to determine an average frequency and magnitude value for each peak in the 4.000Hz-10.000Hz

```
% Difference between old and new peak
d(1) = peak_current(j,1) - Peak_old(1);

% Calculate average peak magnitude
H_sum(1) = H_sum(1) + Val(j,1);
c(1) = c(1) + 1;
H_avg(1) = H_sum(1) / c(1);

% Calculate new peak frequency
Peak_freq(1) = Peak_old(1) + (peak_current(j,1) * (H_avg(1))/d(1));
```

Figure 4.11: Implementation of peak estimation in the Idx-to-Hz converter function

range. Three peaks were established in advance to comply with the research by Iida et al on the peaks caused by the ear response. When a new peak was found by the Find Maxima block, a set of if conditions were used to determine which of the three the new peak belonged to. Only the old peak with the lowest centre frequency would be affected, if the centre frequency of the new peak was lower than the centre frequency of the lowest old peak. If the new peak was located between two old peaks, then both of these old peaks would be affected. And if the centre frequency of the new peak was higher than the old peak with the highest centre frequency only that peak would be affected. How the centre frequency and magnitude was used to affect the old peak can be seen in figure 4.11.

The calculation to determine the centre frequency of the updated peak contains four variables. The centre frequencies of the old and new peak were already known, however, the frequency difference between the old and new peak and the updated average height must be calculated first. A count was furthermore kept for each of the three peaks. When a peak was updated the count was increased by 1. This count value was used first of all to calculate the average height of the peak, and second of all to determine when the peak had been updated 200 times. The frequency difference and average height values were used as regulators, so a great difference value, which indicates that the new peak is far from the old peak, will result in a lesser effect on the updated centre frequency. This implementation was chosen to limit how much the centre frequency of the peak moved from update to update. This process was repeated for any of the old peaks that was affected by the position of the new peak.

An if condition was then implemented before outputting the updated centre frequencies and average height. This if statement checked whether the difference between the old and updated centre frequency was greater than 100Hz or less than -100Hz. If any of these two conditions were true the update would be disregarded. Initial testing showed that even with the aforementioned regularisation the centre frequency could jump wildly, which decreased the accuracy of the result.

The output of the Idx-to-Hz converter function was then sent to a simple custom Matlab function called Peak-gate. The purpose of this function was to send filter coefficients to the parametric EQ blocks when the count for each peak reached 200 updates. The 200 number was chosen as it allowed for a large sample size for

each peak, but was small enough to be reached within 1 minute of operation. The average height of each peak was used to determine the gain the parametric filters, while the centre frequency of the peaks was used for the centre frequency of the filters. When the parametric EQ coefficients were set the filters were static, meaning that they would not be updated further until the system was reset. The peak centre frequencies and magnitudes are furthermore extracted for visual inspection.

Chapter 5

Evaluation

5.1 Level compensation

5.1.1 Purpose

A series of tests were conducted to evaluate the performance of each of the three subsystems. The evaluation of the sound level compensation subsystem was focused on four factors:

- Accuracy
- Consistency
- Stability
- Convergence speed

The accuracy of the level compensation is determined by how balanced the sound level is from the two speaker units. The goal of the entire subsystem is to compensate for any sensitivity deviations so that the sound level in the left ear cup is equal to the sound level in the right ear cup. The greater the difference is between the sound levels the smaller the accuracy is.

An optimal implementation of a sound level compensation system will furthermore produce the same output each time it is given the same input. If the subsystem is run three times with the exact same input signal, it should produce the same compensation gain levels as a result of each of the three runs. However, some variance is to be since the gain levels also depend on the microphone input signals. Noise from outside the headphone may be be picked up by the microphones and skew the results.

The sound levels of the left and right side in the headphone is adjusted continuously. It is therefore important that the system remain stable regardless of the

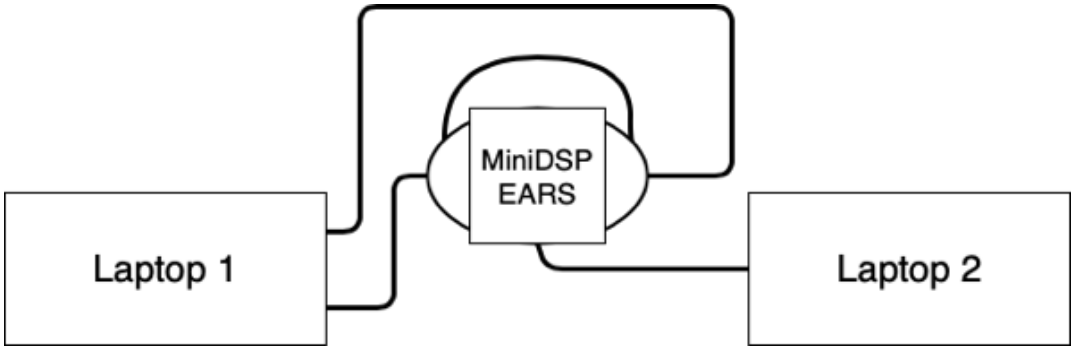


Figure 5.1: Overview of the setup used throughout the evaluation

audio input or microphone input. An unstable level compensation could result in the sound levels getting decreased to a level where the sound is inaudible, or even worse, get increased to a level that could be damaging for the headphone drivers and for the listener's hearing.

While the subsystem must be stable, it must also adapt quickly. The speed at which the subsystem achieves the optimal accuracy is determined by the convergence speed of the LMS implementation. A LMS system with a short convergence speed will adapt quickly to changes and achieve the optimal accuracy quickly, however, it will also be less stable and abrupt changes in sound level may be irritating for the listener.

5.1.2 Setup

The same hardware setup was used for each of the three test sessions. This setup consisted of; 2 laptops, the headphone prototype and a measurement device called the MiniDSP EARS. The MiniDSP EARS[1] is a relatively inexpensive ear simulator that includes a set of synthetic pinna, ear canal and calibrated microphones placed at the ear drum position. An overview of how the four hardware components used for the evaluation were connected can be seen in figure 5.1. Each side of the prototype headphone was connected to laptop 1 via the USB-C connectors. The MiniDSP, on the other hand, was connected to Laptop 2 via a USB-A connector. The headphone prototype was then placed over the synthetic ears of the MiniDSP device. This made it possible to use the MiniDSP EARS to measure the sound level at the ear drum position of both ears.

The software components used for the evaluation naturally depended on the subsystem to be evaluated. For the evaluation of the sound level compensation subsystem the two other subsystems were disconnected, so the sound level compensation could be evaluated in isolation. The sound level compensation subsys-

[1] minidsp.com/products/acoustic-measurement/ears-headphone-jig

tem was running in Matlab Simulink on laptop 1, while laptop 2 ran Audacity[2] to record the two output channels of the MiniDSP EARS. Two different stimuli was used for the sound level compensation evaluation; white noise and a rock song. The white noise signal consisted of a mono signal, so that the exact same audio signal was sent to both sides of the headphone prototype. The 60 seconds long white noise signal was generated in advance using Room EQ Wizard[3] and saved as a WAV file on laptop 1. It was then possible to open the white noise file using Matlab Simulink. The music track was a song called "Rest my chemistry" by the band "Interpol", which was also saved on laptop 1 as an Mp3 file, so that it could be opened using Matlab Simulink. This specific song was chosen because it contains loud and soft sections. These abrupt changes from soft to loud and loud to soft was then used to test the stability of the subsystem.

5.1.3 Setup

The sound level compensation evaluation session consisted of five rounds of measurements to test the system under multiple conditions. During the first round, the purpose of which was to serve as a reference, the sound level compensation subsystem was bypassed. The white noise signal was used as stimulus, which was then routed directly to the output. The two microphone signals from the MiniDSP EARS could therefore be used to determine how great the difference was between the sound level at the left and right side of the headphone without compensation. While bypassed the compensation subsystem still got the audio and microphone inputs, so it was also possible to log the Error signal for the duration of the white noise stimulus.

The level compensation subsystem was then reconnected for the remaining four rounds. Other than that the second round of testing was identical to the first round. The output of the subsystem and the output of the MiniDSP EARS was saved and the error signal was logged.

For the third and fourth rounds the gain of the signals was decreased for the left and right side respectively. This was done to test how quickly the subsystem was able to compensate for more extreme deviations in sound levels. During round three the gain level of the right channel was reduced from 0.55 to 0.27 on laptop 1. While the left channel was reduced from 0.55 to 0.27 for the fourth round. The same white noise stimulus was used for these two rounds.

The fifth round was included to test the stability of the subsystem. The music stimulus was used instead of the white noise signal, and the gain levels was returned to the same standard positions that was used for the first and second rounds. Again the output signals of the subsystem and MiniDSP EARS was saved,

[2] audacityteam.org
[3] roomeqwizard.com

Table 5.1: Final error and gain values for each round of sound level compensation evaluation

Round	1 (off)	2 (Center)	3 (Left)	4 (Right)	5 (Song)
Final error value	0,092	-0,004	0,0009	0,027	-0,071
Final gain value	5,01	1,35	0,376	4,197	1,156

while the error signal was logged.

5.1.4 Results

The final error and gain values for each round can be seen in table 5.1. The gain values vary from round to round, which was expected. Gain values from rounds 2 and 5 are consistent, while the gain from round 3 is less than 1 and greater than 1 for round 4. The final estimated error ratio for rounds 2-5 was within +-0,08.

Figure 5.2 shows the estimated error ratios over time for rounds 1-4. For rounds 2-4, where the sound level compensation subsystem was turned on, the error ratio gradually decreased over time. The convergence time for each round was roughly 19-26 seconds.

The best accuracy was achieved during rounds 3 and 4 since the subsystem over corrected during round 2. This can be seen in the measured error ratios over time in figure 5.3. The sound level disparity was less for all three rounds with the subsystem enabled, than the level disparity for round 1 where the subsystem was disabled.

The measured and estimated error values during round 5 can be seen in figure 5.4. The measured error ratio varied over time, however, this was as expected. Different sound levels at the two sides is often used to separate the instruments of the song. In the estimated error signal it can be seen that the subsystem was generally robust to changes in the song except for a deviation at frame number 218. After that deviation the subsystem quickly recovered.

5.2 Graphic EQ

5.3 Purpose

The graphic EQ subsystem is based on a simple LMS implementation similar to the level compensation subsystem. The same four factors are therefore evaluated to access the performance of the graphic EQ subsystem:

- Accuracy
- Consistency

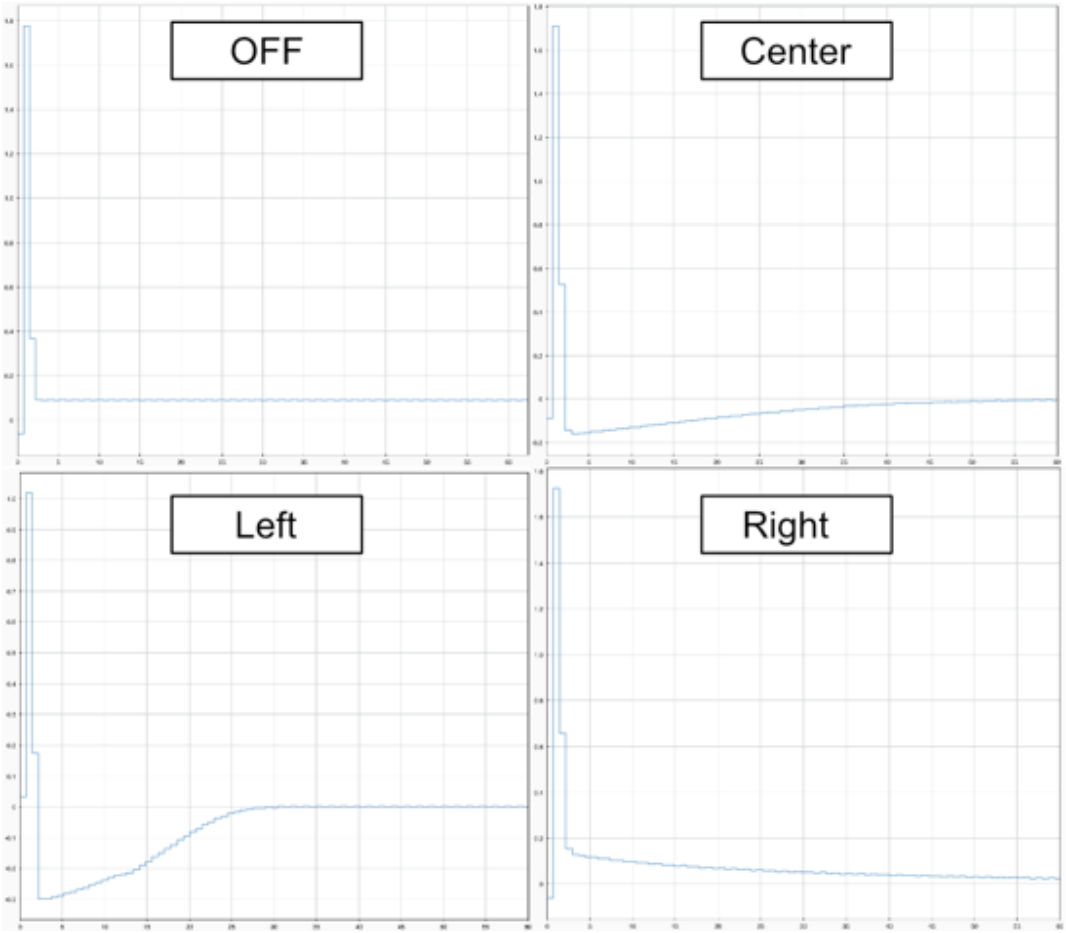


Figure 5.2: Estimated error logs for rounds 1(top-left), 2(top-right), 3(bottom-left) and 4(bottom-right)

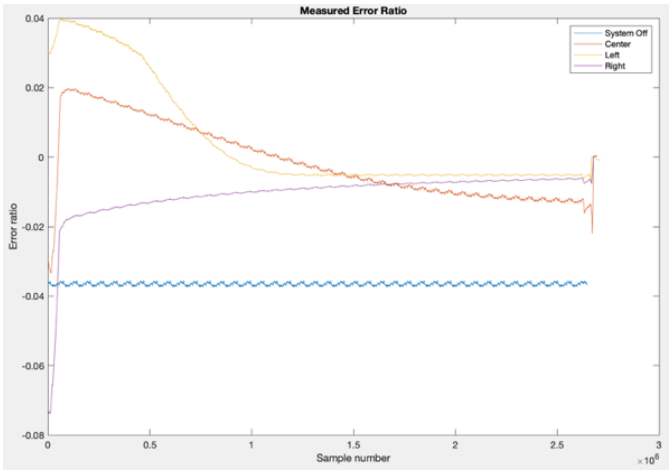


Figure 5.3: Measured error ratio values for rounds 1(blue), 2(orange), 3(yellow) and 4(purple)

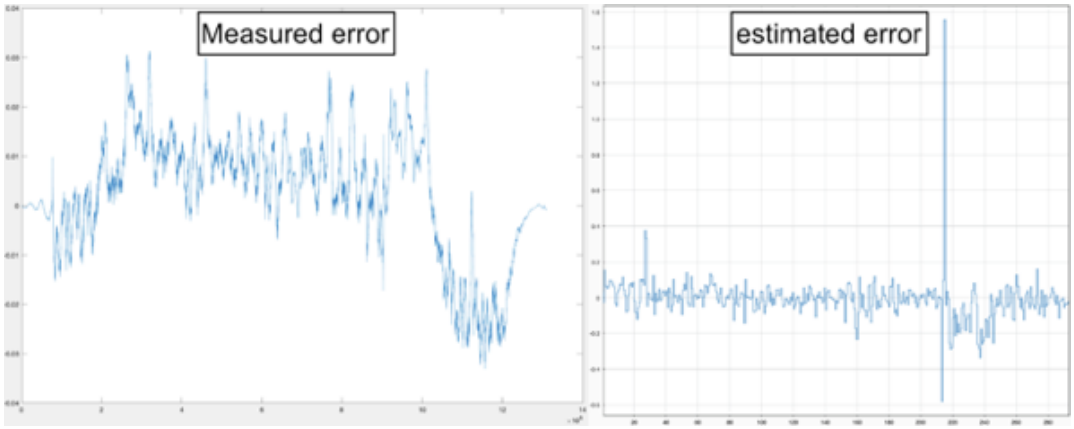


Figure 5.4: Measure error ratio (left) and estimated error ratio (right) using the music track stimulus

- Stability
- Convergence speed

The purpose of the graphic EQ subsystem is to compensate the frequency response between 50Hz-1.600Hz, so that it is as flat as possible at the ear drum. A completely flat frequency response will result in an output signal that is exactly equal to the input signal. Deviations from the flat frequency response will therefore result in a lesser accuracy rating. Like for the sound level compensation subsystem it is important that the graphic EQ output is remains consistent given the same input signal. Varying accuracy will first of all result in an inconsistent user experience, and second of all will second of all reduce the realism of the spatial reproduction. Both the accuracy and consistency of the output may be affected by noise from outside the headphone, so the consistency is also dependant on the subsystem's robustness against noise.

Since the graphic EQ coefficients are computed continuously it is furthermore of utmost importance that the system remains stable regardless of input or microphone signal. If the subsystem is unstable frequencies in the 50Hz-1.600Hz band could get either filtered out entirely or amplified to levels that could be damaging for the headphones and the listener's hearing.

Convergence speed is also an important factor for the graphic EQ subsystem. Like for the sound level compensation subsystem it is important that it adapts quickly so that it achieves the optimal accuracy as quickly as possible. But the steps should also be small enough that the listener does not experience abrupt changes to the tonality of the headphones.

5.3.1 Setup

The hardware setup used for evaluation of the graphic EQ subsystem was identical to the setup discussed in section 5.1. The software setup running on laptop 1, on the other hand, was adjusted to evaluate the graphic EQ subsystem. The graphic EQ subsystem running in Matlab Simulink was isolated, so that the peak EQ and sound level compensation subsystems did not affect the output. Only a single audio channel was utilised to simplify the evaluation procedure and analysis. In terms of stimuli signals the same white noise and music tracks were used as input. A logarithmic sine sweep was furthermore used to evaluate the final graphic EQ coefficients.

5.3.2 Procedure

The evaluation procedure itself was divided into two parts. During the first part a stimulus sound was put through the system while output signals from the subsystem and MiniDSP EARS was saved. The error signal was furthermore logged. After the complete stimulus signal had run through the subsystem the resulting graphic EQ coefficients were logged for use during the second part of the evaluation procedure.

For the second part of the evaluation the EQ coefficients from the first part were extracted and placed in a static Simulink graphic EQ. The logarithmic sine sweep stimulus was then put through the static graphic EQ and the Simulink output and the MiniDSP EARS output was saved. The MiniDSP EARS output could then be analysed to evaluate the frequency response at the simulated ear drum using the final EQ coefficients, and thereby the accuracy of the subsystem.

This entire procedure was repeated six times. The white noise stimulus was used during the first part for the first three repetitions, while the music track was used for the remaining three repetitions.

5.3.3 Results

Final error values and filter gain coefficients for each of the six rounds can be seen in table 5.2. Error values were consistent for rounds 1-3 and 4-6, but there is a big difference when comparing rounds where white noise was used and rounds where the music track was used. Although the final filter gain coefficients are generally consistent for all rounds. Filter coefficients for rounds 1-3 had a tendency to be slightly greater than those for rounds 4-6.

Estimated error values over time can be seen in figure 5.5. Convergence speed for the graphic EQ subsystem was roughly 19-26 seconds, which was similar to the convergence time of the sound level compensation subsysttem. The convergence was smooth for rounds 1-3, where the white noise stimulus was utilised. The

Table 5.2: Final error and filter gain values for each round of graphic EQ evaluation

Round	1 (white noise)	2 (white noise)	3 (white noise)	4 (Song)	5 (Song)	6 (Song)
Final error value	8,722	7,579	10,31	42,43	45,63	47,43
50Hz	8,424	8,272	7,268	8,338	8,625	8,487
63Hz	3,034	3,142	3,051	2,728	2,599	2,926
80Hz	3,144	4,942	3,98	5,561	5,424	5,755
100Hz	2,68	2,694	2,989	3,416	3,015	3,288
125Hz	-1,862	-2,065	-1,714	-2,72	-2,729	-2,838
160Hz	-5,368	-5,587	-5,982	-6,744	-6,767	-6,675
200Hz	-9,063	-8,642	-8,341	-10,01	-10,07	-9,938
250Hz	-8,714	-9,207	-9,374	-10,35	-10,47	-10,52
315Hz	-7,704	-8,038	-7,771	-9,032	-8,935	-8,896
400Hz	-7,221	-6,772	-7,687	-8,031	-8,019	-8,02
500Hz	-5,114	-4,999	-5,128	-6,315	-6,427	-6,275
630Hz	-2,388	-2,46	-2,853	-3,646	-3,702	-3,525
800Hz	-2,107	-2,092	-2,262	-1,69	-1,894	-1,581
1.000Hz	0	0	0	0	0	0
1.600Hz	-3,342	-3,405	-3,675	-4,006	-3,981	-3,889

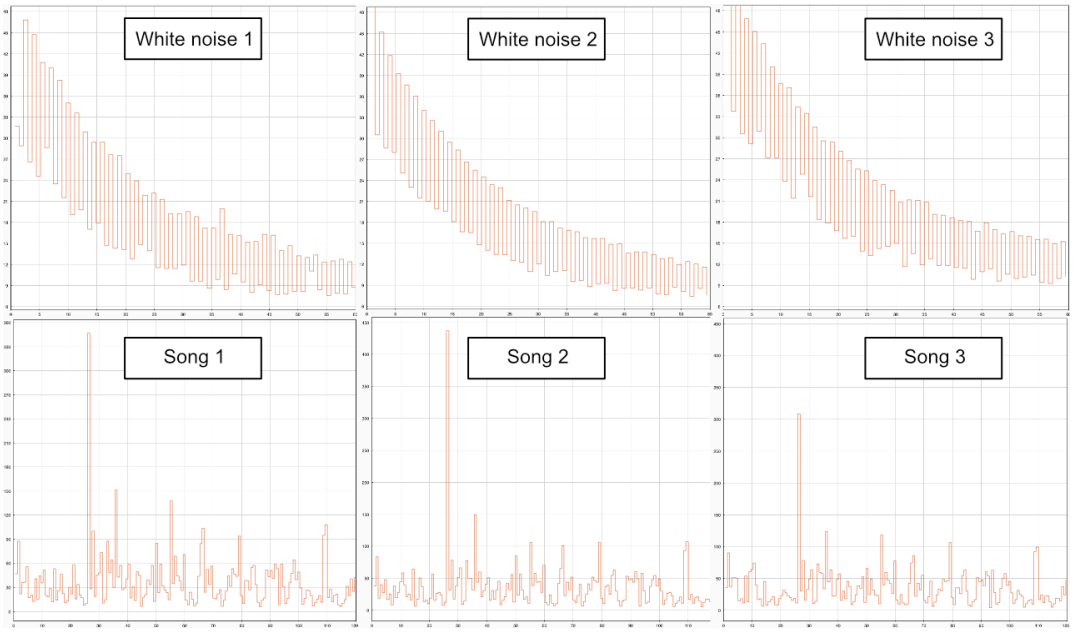


Figure 5.5: Estimated error logs for rounds 1(top-left), 2(top-mid), 3(top-right), 4(bottom-left), 5(bottom-mid), 6(bottom-right)

subsystem was generally robust to rapid changes in the music track, the exception being a peak at frame number 28, after which the subsystem quickly recovered.

Figure 5.6 shows the measured frequency response for rounds 1-3 at the ear drum position of the MiniDSP ears after the final graphic EQ filters had been applied. The frequency response in the 20Hz-1.600Hz was within +-1dB for all three rounds.

the same type of measurements but for rounds 4-6 are found in figure 5.7. Results of rounds 5 and 6 were similar to those of rounds 1-3. The result of Round 4, however, was more inaccurate as it deviated as much as +-2dB.

5.4 Peak EQ

5.5 Purpose

While the two previously discussed subsystems are simple adaptive algorithms that adjust parameters continuously in real-time, the peak EQ subsystem is a discrete system that calculates the appropriate parameters once. So, only two factors are relevant for the evaluation of the system:

- Accuracy
- Consistency

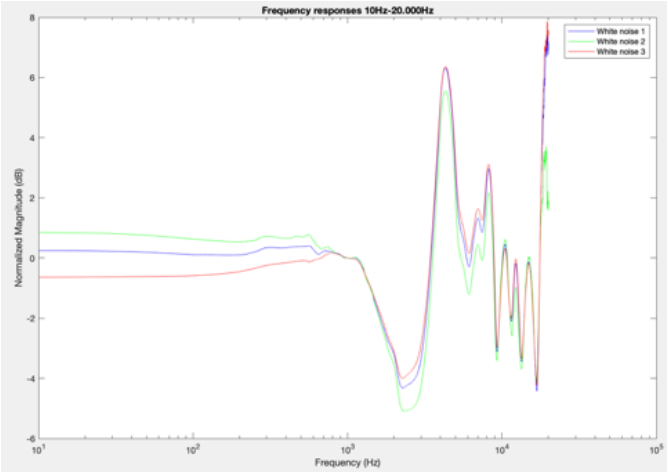


Figure 5.6: Measured frequency response after graphic EQ compensation at eardrum position using white noise stimulus

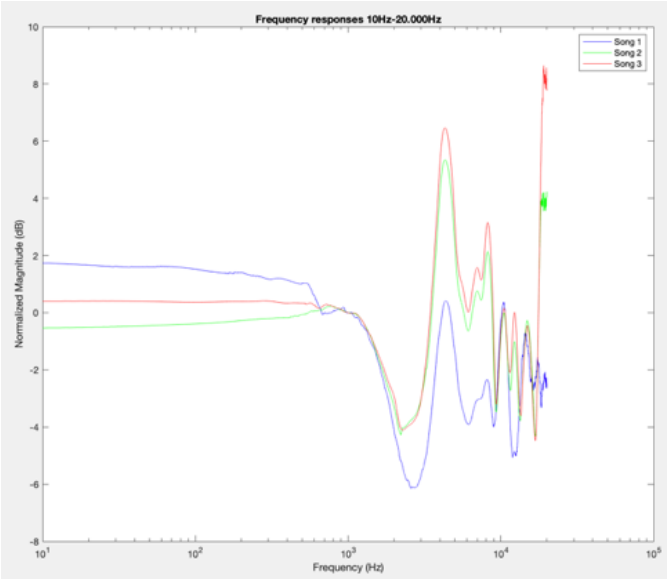


Figure 5.7: Measured frequency response after graphic EQ compensation at eardrum position using music track stimulus

The general purpose of the peak EQ subsystem is similar to that of the graphic EQ subsystem, however, it operates in a difference frequency range. The accuracy of the subsystem is determined by how close to flat the frequency response is in the range 4.000Hz-10.000Hz at the ear drum. Although, the accuracy of the peak EQ subsystem is expected to be less than for the graphic EQ subsystem. Peak EQ is limited to three individual filters and the bandwidth of these three filters is fixed, so achieving a completely flat frequency response is unlikely.

Consistency each time the system is run is naturally also an important factor for this subsystem. Inconsistent results could lead to inconsistent user experiences and inaccuracies similar to inconsistent output from the graphic EQ subsystem.

5.6 Setup

In terms of hardware the setup for the peak EQ evaluation was identical to the sound level compensation and graphic EQ evaluations. On the software side the setup was very similar to the graphic EQ tests, however, the graphic EQ subsystem was bypassed while the peak EQ system was reconnected. The graphic EQ and level compensation subsystem thus had no influence on the output signals. The subsystem was limited to a single channel to limit the scope of the evaluation procedure and analysis.

5.7 Procedure

The procedure for the peak EQ evaluation was very similar to the procedure for the graphic EQ evaluation. The only difference between the two procedures was that three parametric filters were used during the second part of the procedure instead of the single graphic EQ used during evaluation of the graphic EQ subsystem.

5.8 Results

Estimates for peak magnitude and centre frequencies for all six rounds can be found in table 5.3. Magnitude estimates varied +-4dB at most, while centre frequencies exhibited a maximum variance of roughly +-300Hz. Both magnitude and centre frequency estimates had a tendency to be slightly greater for rounds 1-3 than for rounds 4-6.

Figures 5.8 and 5.9 features measurements from the ear drum position of the MiniDSP EARS for rounds 1-3 and 4-6 respectively. All measurements show very pronounced peaks in the 4.000Hz-10.000Hz range.

Table 5.3: Final estimated magnitude and centre frequency values for rounds 1-6

Round	1 (white noise)	2 (white noise)	3 (white noise)	4 (Song)	5 (Song)	6 (Song)
Peak 1 magnitude	11,28dB	11,94dB	11,4dB	10,51dB	10,88dB	9,751dB
Peak 1 centre frequency	5325Hz	5347Hz	5355Hz	5373Hz	5395Hz	5256
Peak 2 magnitude	12,23dB	11,9dB	11,08dB	9,166dB	9,509dB	8,516dB
Peak 2 centre frequency	7053Hz	7292Hz	7430Hz	6981Hz	7279Hz	6831Hz
Peak 3 magnitude	12,61dB	11,86dB	11,03dB	9,142dB	9,469dB	8,318dB
Peak 3 centre frequency	10.000Hz	10.000Hz	10.000Hz	9296Hz	9065	9694

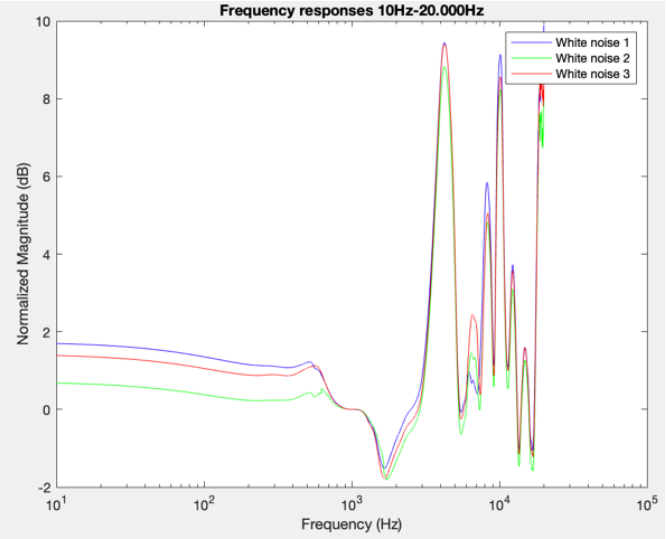


Figure 5.8: Measured frequency response after peak EQ compensation at eardrum position using white noise stimulus

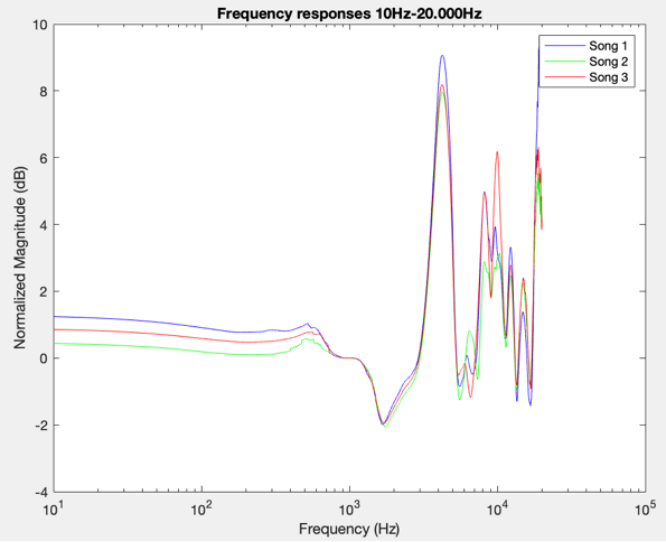


Figure 5.9: Measured frequency response after peak EQ compensation at eardrum position using music track stimulus

Chapter 6

Discussion

6.1 Novelty of prototype system

The project presented in this thesis included three novel features:

1. Sound level and frequency response compensation simultaneously
2. Ear response estimation using affordable and readily available components
3. Real-time implementation

The initial concept of implementing compensation in real-time was inspired by the system presented by Liski et al [19]. However, their system focused solely on frequency response compensation. The compensation system prototype presented in this thesis included a sound level compensation subsystem running alongside the frequency response compensation. Studies such as the one performed by Gutierrez-Parera and Lopez showed that left-right channel sensitivity disparities also had a great impact on the perceived accuracy of the spatial audio localisation [13]. Sensitivity disparities between the two channels are common in affordable headphones, even though small variations are within tolerances during assembly of the driver it can result in significant variations in sensitivity. The sound level at each ear may furthermore be affected if independent frequency response compensation is implemented.

Another novel feature of the system presented in this thesis is the estimation of the listener's ear response. The system proposed by Liski et al estimates the ear canal response to estimate the frequency spectrum at the ear drum [19]. However, since their system is developed for in-ear headphones it is not possible to estimate the response of the pinna. Knowing the listener's entire ear response is a great benefit as it can be used to produce accurate spatial audio [26]. A system proposed by Lindau and Brinkmann was able to estimate the listener's pinna response using

a custom measurement setup where a microphone was placed at the blocked ear canal position [18]. The drawback of the Lindau and Brinkmann system was that it did not run in real time, and secondly was too complex to be realistic for consumer applications. One of the objectives for the system presented in this thesis was an implementation that realistically could be used for consumer applications in the future, as discussed in section 6.3. Therefore all hardware components was picked from different existing products currently being sold by AIAIAI.

6.2 Discussion of evaluation results

6.2.1 sound level compensation

Results from the evaluation of the sound level compensation subsystem was generally promising, especially when using the white noise stimulus. The compensation gain values that was calculated by the algorithm corresponds well with the anticipated results. For round 3 a gain value of 0,376 was calculated, which reduced the sound level in the left channel as the gain value was smaller than 1. The opposite result was calculated when the sound level was reduced in the left channel for round 4. The final gain values calculated in round 2 and 5 were quite similar, indicating that compensation could be done with an arbitrary stimulus signal and a white noise signal.

From the error logs of rounds 2-4 it can be seen that the convergence speed it roughly 25-35 frames, which corresponds to roughly 19-26 seconds. This rather long convergence time was among other things caused by the large frame size, which reduces the frequency of updates. The convergence speed could be improved by either reducing the frame size or increase the step size. Both of these solutions are discussed in section 6.3.

After convergence, the subsystem was able to achieve a good accuracy and consistency. Rounds 2-4 showed error ratios measured at the simulated ear drums that was close to 0. In terms of consistence only round number 2 stood out as it over corrected slightly.

In addition the sound level compensation subsystem was able to handle the arbitrary signal well. The measured error ratio fluctuated slightly, however, it was closer to 0 than the sensitivity disparity measured in round 1. The estimated error ratio had one big deviation at around frame number 218 and a corresponding drop in accuracy in the measured error ratio. This sudden deviation coincided with a sudden drop in volume level in the music track. The algorithm fortunately recovered quickly, so it did not become unstable.

6.2.2 Graphic EQ

Throughout the 6 measurement rounds the resulting filter gains remained largely consistent. Most filter gains varied roughly +-1dB throughout all rounds, which means that the variations are barely noticeable for the listener. The final error values were significantly greater for the rounds using the music track than the rounds using white noise, however, the final filter gain values remained consistent with the earlier results. The AIAIAI S04 speaker units used for the prototype headphone are known to exhibit an elevated response in the bass to low-mid range, which is also reflected in the final filter gain results.

The adaptive algorithm convergence speed proved to be consistent with the convergence speed of the sound level compensation subsystem. The same drawbacks and solution proposals therefore also applies here.

In terms of stability the graphic EQ subsystem generally handled the music track well. The same part of the song resulted in a rather large deviation, but the error logs indicate that the algorithm was quick to recover from the sudden deviation.

The accuracy of the compensation measured at the simulated ear drum is acceptable for the majority of the measurement rounds. Two of the rounds using white noise and two audio track rounds resulted in responses that were well within +-1dB thought the entire 50Hz-1.600Hz range. Further tests are required to determine why two rounds resulted in responses that were flat but elevated until roughly 500Hz. The filter gains for the two problematic rounds do not vary significantly from the other rounds, so it is a possibility that the elevated responses stem from measurement or analysis issues.

6.2.3 Peak EQ

While the results of the peak EQ evaluation were promising in terms of consistency, it was also apparent that improvements have to be made to achieve accurate results. Both the estimated peak magnitudes and centre frequencies exhibited little variation from round to round. Magnitudes estimated using the white noise stimulus had a tendency to be slightly than those estimated using the music track stimulus. This small difference could be the result of the white noise remaining at the same sound level throughout the entire stimulus, while the sound level of the music track fluctuates over time.

Measurements of the accuracy, however, did not provide results as anticipated. The peak EQ subsystem was developed to attenuate the estimated peaks, although it seems the peaks are being amplified instead. Further testing has to be performed to determine the reason for this amplification. Looking at the measurements for the graphic EQ subsystem it can be seen that the centre frequencies of the peaks in the 4.000Hz-10.000Hz range are well estimated.

6.3 Future improvements

Results from the evaluation of the prototype system were promising, however, there are areas where the system could be improved as discussed in the previous section. One of these areas is the long convergence time of the adaptive algorithms used for the sound level compensation and graphic EQ subsystems. Convergence time may be reduced by increasing the step size or reducing the frame size. Reduction of the frame size would require improvement in efficiency as the current frame size was chosen to avoid missing samples because the calculations could not be completed within the time it takes to play one frame. An increased step size is simpler to implement, however, increasing the step size too much could result in an unstable system. A stress test of the system could be performed to determine the greatest step size that still results in a stable and robust compensation system. Alternatively more advanced variations of the LMS algorithm, such as the Robust Variable Step Size Normalised LMS (RVSS-NLMS) used by Liski et al [19], could be explored.

A test should also be conducted to determine why the peak EQ subsystem amplified the peaks rather than attenuating. The peaks seams to be amplified more when using the white noise stimulus than when using the music track stimulus, which corresponds with the greater estimated magnitudes. This indicates that the cause of the problem could stem from the estimation of the peak magnitudes or the filtering process itself. A simple test could be remove the inversion of the filter magnitudes, which is currently implemented in the Peak-gate function. This inversion was implemented to create a notch filter that would cancel out the peak, but in reality it may unintentionally be doing the opposite. if the removal of the inversion does not mitigate the problem, a detailed investigation of the parametric filter blocks will have to be conducted.

An audible click sound is produced when the graphic EQ filter gains are updated. Since they are audible they may be detrimental to the listening experience for the user. For the user they will be perceived as unwanted noise and will likely reduce the perceived realism of the spatial audio. This audible click was one of the reasons a long frame size was chosen. Unfortunately it did not reduce the amplitude of the clicks but only reduced the frequency of them as the frequency of the filter updates were reduced. It is possible that the clicks could be mitigated by implementing a smoother transition from one coefficient to a new coefficient. Further development and tests are required to determine whether that is a viable solution.

A series of user tests should be conducted, when the issues discussed above are handled. Testing the compensation system while a person is wearing the prototype headphones will be crucial to gather information on how a listener perceives the experience. A setup similar to the one used by Gutierrez-Parera and Lopez could

be used to gather objective data on whether the listener's localisation accuracy is improved when the compensation system is enabled. User tests would furthermore provide subjective feedback on how the listener perceived the audio quality and realism.

In the longer term, the compensation system could be improved by taking the delay of the audio signal into account. The auditory perception is very sensitive to time differences at the two ears as discussed in section 2.1.2. Different filter coefficients at each side of the headphone could lead to differing group delays, which in turn could distort the perceived ITD. This problem could likely be mitigated by implementation of a delay compensation subsystem that equalises the delay at the two sides.

The estimation of the ear response peak centre frequencies yielded promising results during evaluation. These centre frequencies could be exported to external applications, that in turn could use them to produce personalised spatial audio. Such a system could in theory produce more realistic audio than systems that utilise generic ear responses. If the ear response is accurately estimated it may be possible to precisely reproduce the sound field at the ear drum, thus providing a sound that is perceptually indistinguishable from a real world version of the same sound.

Chapter 7

Conclusion

This project was conducted in collaboration with AIAIAI to test the feasibility of a headphone compensation system for realistic spatial audio applications. The listener's personal ear response must be known to provide a perfectly realistic auditory experience. However, measuring the ear response has traditionally been an impractical process involving elaborate technical setups. The objective of this project has therefore been to develop a prototype system using affordable and readily available components, that was able estimate and compensate for the listener's personal ear response and imperfections caused by the headphone speaker units.

A frequency response study of a synthetic ear was conducted in order get a general idea of how the pinna and ear canal affects the ear response at the ear drum. By measuring the response of the entire ear and the isolated pinna it was possible to calculate the ear canal response. From this study it could be seen that the pinna response consisted of three prominent peaks in the 4.000Hz-10.000Hz range, while the ear canal resonances generally amplified the magnitude in that same range. These findings correlated well with previous research on the topic.

The findings of the ear response study and an exploration of the current state of the art formed the basis for development of the compensation system prototype. A set of AIAIAI TMA-2 headphones and other AIAIAI components were used to assemble the hardware element of the prototype system. The software element was divided into three subsystems that each fulfilled different sections of the requirement specification.

Results of the evaluation was generally promising. However, the convergence speed of the sound level compensation and graphic EQ subsystems, and the accuracy of the peak EQ subsystem, was not ideal and could be improved through further development. While the accuracy of the peak EQ subsystem was less than ideal the estimates of the peak centre frequencies were quite accurate, which indicates that it is possible to obtain a rough estimate of the ear response using a simple and affordable headphone setup.

Bibliography

[1] V.R. Algazi et al. "The CIPIC HRTF database". In: *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics* (2001), pp. 99–102.

[2] Areti Andreopoulou and Agnieszka Roginska. "Towards the Creation of a Standardized HRTF Repository". In: *Audio Engineering Society Convention* 131 (2011).

[3] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. 1. ed. MIT Press, 1997.

[4] John Borwick. *Loudspeaker and Headphone Handbook*. 3. ed. Focal Press, 2001.

[5] Fabian Brinkmann and Alexander Lindau. "On the effect of individual headphone compensation in binaural synthesis". In: *Fortschritte der Akustik: Tagungsband* 36 (2010), pp. 1055–1056.

[6] Simon Carlile. *Virtual Auditory Space: Generation and Applications*. 1. ed. Springer-Verlag Berlin Heidelberg, 1996.

[7] Thibaut Carpentier, Markus Noisternig, and Olivier Warusfel. "Twenty Years of Ircam Spat: Looking Back, Looking Forward". In: *International Computer Music Conference* (2015), pp. 270–277.

[8] E.C. Carterette and M.P. Friedman. *Handbook of Perception*. 1. ed. Academic Press, 1978.

[9] Anders T. Christensen et al. "Magnitude and Phase Response Measurement of Headphones at the Eardrum". In: *Audio Engineering Society* 51 (2013), pp. 21–24.

[10] James W. Cooley and John W. Tukey. "An Algorithm for the Machine Calculation of Complex Fourier Series". In: *Mathematics of Computation* 19.90 (1965), pp. 297–301.

[11] Angelo Farina. "Advancements in impulse response measurements by sine sweeps". In: *Audio Engineering Society* 112 (2007).

[12] M.B. Gardner. "Distance estimation of 0 degree or apparent 0 degree-oriented speech signals in anechoic space". In: *Journal of the Acoustical Society of America* 45 (1969), 47–53.

[13] Pablo Gutierrez-Parera and Jose J. Lopez. “Influence of the Quality of Consumer Headphones in the Perception of Spatial Audio”. In: *Applied Sciences* 6.117 (2016), pp. 1–18.

[14] R.M. Hershkowitz and N.I. Durtach. “Interaural time and amplitude jnds for a 500-Hz tone”. In: *J Acoust Soc Am* 469 (1969), pp. 1464–1467.

[15] Kazuhiro Iida et al. “Median plane localization using a parametric model of the head-related transfer function based on spectral cues”. In: *Applied Acoustics* 68 (2007), 835–850.

[16] M.L. Hawley H.S. Colburn J. Koehnke C.P. Culotta. “Effects of reference interaural time and intensity differences on binaural performance in listeners with normal and impaired hearing”. In: *Ear Hear* 16 (1995), 331–353.

[17] R.G. Klump and H.R. Eady. “Some measurements of interaural time difference thresholds”. In: *J Acoust Soc Am* 28 (1956), 859–860.

[18] Alexander Lindau and Fabian Brinkmann. “Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings”. In: *Audio Engineering Society* 60.1/2 (2012), 54–62.

[19] Juho Liski et al. “Real-Time Adaptive Equalization for Headphone Listening”. In: *European Signal Processing Conference (EUSIPCO)* 25 (2017), pp. 638–642.

[20] James H. McCellan, Ronald W. Schafer, and Mark A. Yoder. *DSP First*. 2. ed. Pearson, 2015.

[21] A.W. Mills. “Auditory localization”. In: *Foundations of modern auditory theory* 2 (1972), pp. 301–345.

[22] Dorte Hammershøi; Henrik Møller. “Sound transmission to and within the human ear canal”. In: *Journal of the Acoustical Society of America* 100.1 (1996), pp. 408–427.

[23] Henrik Møller. “Fundamentals of Binaural Technology”. In: *Applied Acoustics* 36.3/4 (1992), pp. 171–218.

[24] Henrik Møller et al. “Transfer Characteristics of Headphones Measured on Human Ears”. In: *Audio Engineering Society* 43.4 (1995), pp. 203–217.

[25] Klaus Riederer. “Transfer function measurements in audio”. In: *Signal Process. methods Digit. audio* 41 (2001), 1–24.

[26] Agnieszka Roginska and Paul Geluso. *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*. 1. ed. Routledge, 2018.

[27] Jussi Rämö and Vesa Välimäki. “High-Precision Parallel Graphic Equalizer”. In: *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING* 22.12 (2014), pp. 1894–1904.

[28] S.W. Smith. *Digital Signal Processing: A Practical Guide for Engineers and Scientists*. 1. ed. Newnes, 2002.

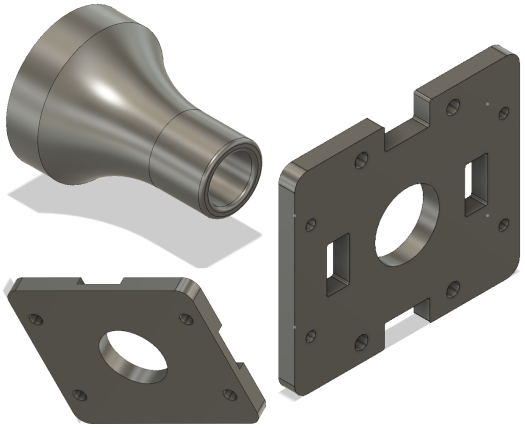
[29] Lizhe Tan and Jean Jiang. *Digital Signal Processing: Fundamentals and Applications*. 3. ed. Academic Press, 2018.

[30] Dmitry N. Zotkin et al. "HRTF PERSONALIZATION USING ANTHROPOMETRIC MEASUREMENTS". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2003), pp. 157–160.

[31] J. Zwislocki and R.S. Feldman. "Just noticeable differences in dichotic phase". In: *J Acoust Soc Am* 28 (1956), 860–864.

Appendix A

Ear response lasercut and 3D-print STL files



Appendix B

Complete software implementation in Matlab Simulink