# Reproduction of Surround Sound in Headphones

AALBORG UNIVERSITY


December 2004
Group 960 – Department of Acoustics

Faculty of Engineering and Science
Aalborg University

AALBORG UNIVERSITET

**Institute of Electronic Systems - Department of Acoustics**

**TITLE:**
Reproduction of Surround Sound in Headphones

**TIME PERIOD:**
9th semester
September 2nd to December 20th, 2004

**GROUP:**
960 - 2004

**GROUP MEMBERS:**
Isabella Biedermann
Ove Obbekjær Jørgensen
Michal Matějka
Sebastian Merchel
Diane Jasmine Poncet
Martin Ordell Sørensen

**SUPERVISORS:**
Dorte Hammershøj
Adrian Celestinos

**COPIES: 10**

**NO OF PAGES:**

**- IN MAIN REPORT: 84**

**- IN APPENDIX: 34**

**TOTAL: 130**

**Abstract:**

When listening to a 5.1 movie soundtrack through headphones, the six channels are normally down-mixed to two-channel stereo, and thus losing the spatial sensations associated with surround sound. The approach in this project was to down-mix the signals into two channels, while additionally including the acoustical filtering that would have occurred, if reproduced through six loudspeakers. The final system should run in real-time based on suitable parameters found through listening tests.
The acoustical filters have been obtained by measuring binaural room impulse responses (BRIRs) in a dedicated multi-channel listening room with the binaural recording head VALDEMAR. The BRIRs were measured in specific positions matching the standard 5.1 surround sound setup. DVD-movies were used as source material from which six discrete audio signals were extracted, convolved with the BRIRs, equalized for measurement and reproduction chain, and down-mixed into two headphone channels.
A listening experiment was conducted with 20 subjects, evaluating different parameters in the binaural synthesis implementation. The implemented system was verified by comparing it to a corresponding binaural recording. It was also concluded that the BRIR can be relatively short without deterioration in sound quality.
A design approach was made for the real-time implementation and preliminary tests were conducted with positive results. However, a complete real-time binaural synthesis has not been achieved within the time frame.

# Preface

This report was written by group ACO-960, $9^{th}$ semester of the international Master of Science programme in Acoustics. It is addressed to students and supervisors at the Institute of Electronic Systems, Aalborg University. The main report is divided into several chapters:

- **Chapter 1** gives an introduction to the subject and presents the problem description.
- **Chapter 2** introduces standards and recommendations for multi-channel setups and analyzes the relevant acoustic areas, such as room acoustics and spatial sound perception. This then leads to the system specification.
- **Chapter 3** assesses the recording and reproduction chain and relevant equalization approaches.
- **Chapter 4** describes the measurements made for obtaining all the needed responses.
- **Chapter 5** analyzes the measurement results in detail and describes the implementation approach.
- **Chapter 6** contains description of listening experiment; goal, method, result, and discussion.
- **Chapter 7** describes the real-time implementation.
- **Chapter 8** gives a final conclusion on the project.

A following appendix contains supplementary documentation. References to literature are in squared brackets, which contain the author's last name and the publication year.

Figure, table, and equation numbering follow the chapter numbering, e.g. second figure of chapter five is called Figure 5.2. References to equations are in parenthesis, e.g. Equation (3.11) for equation number 11 in chapter three.

Attached is a CD-ROM containing the MATLAB source code, the measurement results, movie sequences used in the listening experiment, and a copy of the report in PDF/PS format.

We would like to thank Florian Wickelmaier and Sylvain Choisel for letting us use the multi-channel room and their setup, and Emine Çelik for lending us some of her measurement results.

Note, that the blocked entrance binaural recording technique used in this project is a patented method.

Group ACO-960, Aalborg University, $20^{th}$ December 2004

| | |
|---|---|
| Isabella Biedermann | Ove Obbekjær Jørgensen |
| Michal Matějka | Sebastian Merchel |
| Diane Jasmine Poncet | Martin Ordell Sørensen |

# Contents

Chapter 1

# Introduction

The desire to create a cinema-like experience when watching movies at home, has flourished in a commercial context for numerous years. The first step has obviously been to have a big screen, but in recent years more and more focus has also been given to the sound experience. This was especially influenced by the launch of the Digital Versatile Disc (DVD) in 1996. This new media offered better video quality, while an even greater improvement was the compatibility with discrete multi-channel surround sound. Now consumers could experience cinema sound at home through either Dolby Digital or DTS 5.1 surround sound formats. Today the term "home cinema" has become an everyday expression, and the products to supply this experience range from cheap all-in-one solutions to high-end extravagance for millionaires. The key components in a normal 5.1 surround sound setup include six or more loudspeakers placed at given positions around the listener. However, for one or more reasons, this might not always be a desirable or possible solution. Over the years, manufactures have strived to deliver different solutions that can satisfy a wider consumer range. This includes setups with two or in some cases as little as one loudspeaker that, through different methods, tries to give the listener a sensation of being enveloped in the sound.

The goal of this project, and what is presented in this report, is to reproduce the standard 5.1 surround sound binaurally through a pair of headphones. Normally when listening to a 5.1 movie sound track through headphones, the six channels are down-mixed to two-channel stereo, and thus losing the spatial sensations that are associated with surround sound. The approach in this project is similarly to down-mix the signals into two channels, while additionally including all the spatial information that would have been present, if reproduced through six loudspeakers.

The reproduction of surround sound through headphones has several advantages compared to the normal loudspeaker approach:

- Usually, when having a surround setup with loudspeakers, there will only be one "sweet spot" in which the listener is positioned optimally relative to the loudspeakers. This limitation is eliminated when using headphones.
- The sound reproduced through headphones will not be affected by the environment surrounding the listener, so the user will always have the same listening experience.

- A single pair of headphones will normally be cheaper than a full loudspeaker setup, especially when considering sound quality.
- The use of headphones does not disturb the surroundings, for example neighbours or children sleeping.
- The amplification needed to yield a proper output level from a pair of headphones, is substantially lower than supplying the same sound pressure level from six loudspeakers. This will reduce both power consumption and the amount of needed equipment.

However, some disadvantages by the proposed approach should also be pointed out:

- In general headphones are not that comfortable to wear over a prolonged period of time.
- The user will not be subjected to the "shaking" of the room, that often occurs when low frequencies are reproduced at high levels through loudspeakers. A typical example would be the explosions often associated with action movies.
- The user might tend to have a relative high level when using headphone, which will increase the risk of damaging the hearing with prolonged use.
- The use of headphones can be considered anti-social.

Another important advantage by using headphones is the mobility that it offers. Today most new notebooks come with a DVD-drive and dedicated portable DVD-players are also available. This gives the user the option of watching movies anywhere at anytime. In such situations the user is normally limited to stereo reproduction through either headphones or built in loudspeakers in the given player. Other possible applications would be in public transportation, e.g. in airplanes and buses.

## 1.1 General design approach

The approach for reproducing surround sound through headphones is roughly to create an algorithm that takes a 5.1 soundtrack as input and then outputs a corresponding binaural signal. The principle of this is illustrated in Figure 1.1.

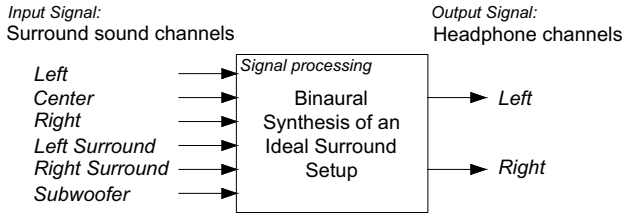



**Figure 1.1:** Simplified model of the binaural system to be created.

The basic idea with such a binaural synthesis is to recreate the exact sound pressure in each ear that would have been present, if the listener was positioned in a normal

surround setup. When the sounds are emitted from the six loudspeakers, they undergo different filtering processes, and are in the end summed into two signals, one at each eardrum. What has to be contained in the *signal processing* block shown in Figure 1.1, is a synthesis of this acoustical filtering. This acoustical filtering can be divided into two parts, the environment in which the normal surround setup would have been situated, and the filtering associated with the physical presence of the listener. Thus, the project will include an assessment of the acoustical properties related to the normal surround sound setup, and the environment in which it is placed. Further, the relevant issues regarding human sound localization and spatial perception will be analyzed. Based on these considerations, a problem description can be formulated.

## 1.2 Problem description

The main goal of the project is to develop a system that can reproduce standard 5.1 surround sound through headphones, by using binaural techniques. Thus, the key elements in the project are as follows:

- Analyse the acoustical filtering that will be lost when moving from a normal surround setup with loudspeakers, to a reproduction through headphones.
- Design a binaural system based on a synthesis of the acoustical filters and use this to convert the 5.1 signal to binaural sound.
- Use formal listening tests to validate the binaural synthesis, and to assess the influence of relevant parameters in order to optimize the system.
- Implement a real-time version of the binaural synthesis according to the results from the listening experiment. The goal is a program that can work as a plug-in for existing PC-based movie players.

Chapter 2

# Analysis

In the introduction it was described how the transition from a normal surround setup to reproduction through headphones, will remove some acoustical filtering of the signals. This analysis will assess what is removed and the consequences of this. The first step is to look into recommendations and standards regarding surround sound formats, and the environment in which the setup is ideally placed. This will give an overview of the subject and convey relevant things for the design phase. The second part is to look into the spatial properties of human sound perception, in order to assess the removed acoustical transmission path between loudspeaker and listener. Based on these two analysis sections, a system specification is presented.

## 2.1 Setup of the surround system

In this section, a general overview of the setup of a surround system, including the history of surround sound, a description of different surround formats, and some standards for both the setup of a surround system and the design of a listening room is provided.

### 2.1.1 History

Since the late $19^{th}$ century when Thomas Edison had first invented a machine able to reproduce sound, improvement of this technique has been a big issue. The first recordings were monophonic and reduced the impression of the room to a one-dimensional experience. The desire for a more natural reproduction of sounds led to the invention of stereophony.

In 1931, Alan Blümlein first patented a stereo technique, and in 1935 he produced the first movie with an optical stereo track. Stereo – derived from the Greek word "stereós" which means spatial, rigid, hard, physical – was from the beginning on not intended to be only a two-channel experience. Though, further electronic knowledge and developement of coding abilities was needed to extend it to a more than two channel technique.

Around 1970, quadrophony – a format in which four original tracks, one for each loudspeaker in the corner of a square, are coded to two – failed to establish itself whereas Dolby achieved to install its surround version as a cinema standard. This surround version

is a matrixing format – two channels carry four signals (Left, Center, Right, Surround) and are played to at least five loudspeakers, the minimum two for Surround channel and occasionally a subwoofer.

The development went on, sound storage became digital, which offered new possibilities. Coding of audio streams made it then possible to use the given storage more extensively. In the 1980s, Dolby introduced its 5.1 system Dolby Digital with five discrete loudspeaker channels and one subwoofer track. Later on the competing company Digital Theatre Systems (DTS) had as well its 5.1 appearance in the cinemas.

The introduction of the Laserdisc (LD) in 1971 gave birth to digital home theaters, first only with stereo sound, since 1982 equipped with Dolby Surround coding, and in 1987 followed Dolby Pro Logic. In 1995, the first Dolby Digital LD was presented. Anyway, this media have not been very spread. The release of the DVD (Digital Versatile Disc) made the LD soon obsolete, and has been very succesfull to date. Until today, surround format development has continued, 6.1 formats have been released. Some of the most significant ones for the project are described in the following section.

### 2.1.2 Different surround formats

Many different sound formats have been developed by now, different surround methods compete with classical mono or stereo techniques.

**Dolby Digital**

Dolby Digital (DD) uses AC-3 coding [ATSC, 2001]. AC-3 coding supports several channels, from 1 to 5 full bandwidth channels (3 Hz to 20.000 Hz) and one optional Low Frequency Effects (LFE) channel (3 Hz to 120 Hz). The use of all six channels is designated as Dolby Digital 5.1, where the LFE channel is the .1 channel. Dolby Digital is part of the DVD-Video standard, but not necessarily with 5.1 channels.

The total bitrate in AC-3 can be set between 32 kbps and 640 kbps, with 448 kbps being the most common on DVDs. A 384 kbps AC-3 stream contains 5 full bandwidth channels of 48000 Hz sample rate in 16 bit resolution which in uncompressed format would require 3.84 Mbps, giving an effective compression ratio of 1:10. The LFE channel is negligible in this calculation, because of its low samplerate (240 Hz). To achieve this high compression, AC-3 utilizes psychoacoustical methods (masking threshold) to remove the sound components that are least audible. It is therefore a so-called lossy method, which means that the sound cannot be reproduced 100% accurately.

**Dolby EX**

The Dolby EX format introduces an additional rear center channel. Stored as a 5.1 stream, the rear center channel is matrixed into both the rear left and the rear right channel and can be re-gained with a matrix-decoder.

**DTS**

DTS is, like DD, a stream capable of 5.1 channels. The main difference between DD and DTS is the used coding technique, which for DTS is called Coherent Acoustics. The typical coding bitrate of DTS is higher than that of the Dolby Digital product and the LFE channel contains the full bandwidth. That means that the overall compression of the DTS stream is generally lower than Dolby Digital.

### DTS Extended Surround

The DTS Extended Surround (DTS-ES) surround format exists in two versions:

- DTS-ES Matrix is actually a 5.1-channel format with the rear surround audio channel matrixed into those of the right and left surround.
- DTS-ES Discrete 6.1 is a true 6.1-channel format, as the rear surround audio channel is discretely encoded into the DTS bitstream. For backwards compatibility, DTS-ES Discrete 6.1 rear surround channel is ignored by DTS 5.1 equipment.

## 2.1.3 Surround setup standards

The DVD-Video should provide the listener with a cinema-like experience in a living room, some standards must be met to receive the optimal reproduced sound. The whole listening environment influences the perception, that is why positioning of the loudspeakers and optimizing size and characteristics of the room (as for example with absorbing materials) is important to make the reproduction of the recorded audio-material as exact as possible.

The loudspeakers in the surround setup should be placed on a circle with the listener in the so called "sweet spot". At this place the artificial environment is supposed to be recreated in the most exact way. If a perfect circle cannot be achieved, the loudspeaker signals should be delayed to shift them virtually to positions on a circle. In Figure 2.1 the loudspeaker setup according to [BS775-1, 1992] and [SMPTE, 1991] standard is shown.

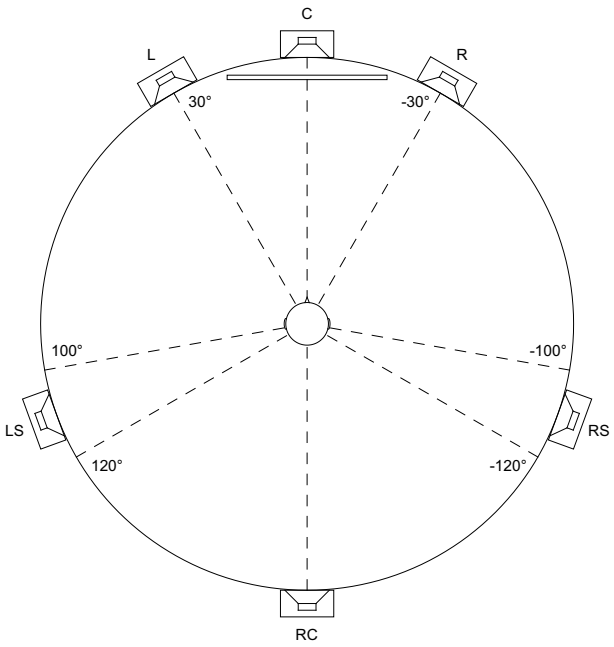



**Figure 2.1:** Standard setup of a 6.1 surround system without subwoofer.

The left front, center and right front loudspeakers are all supposed to be at the same height, around 1.2 m which is assumed to be the average height of the sitting listeners' ears, the rear loudspeakers should at least be at this height. The radius of the circle

depends on the height of the screen and it is expected to be twice that value. As such dimensions would create large radiusses only with wide screens like in cinema application, the diameter of the setup circle may be increased to fit the room better, if this is desirable.

Angles of the loudspeakers should be $\pm 30°$ for left and right, and the center at $0°$ behind the screen if possible. The surround loudspeakers should be placed in the interval of $100°$ to $120°$ according to the circle, their angle to the horizontal plane should not exceed $15°$. The subwoofer may be placed anywhere in the room, according to the preferences of the listener and the room itself. If a rear center loudspeaker is also installed, it should be placed on the circle opposite to the front center loudspeaker.

As the visual aspect of the DVD is the main point in the reproduction, the loudspeakers should be placed to fit the size of the screen. Six loudspeakers are used for left front, center, right front, and the two surround channels, one for the rear center (if the format supports it).

In the playback environment, the LFE channel is amplified with 10 dB according to the bass enhancement in a real cinema environment. A subwoofer is used for reproducing the LFE channel and sometimes for playing the lower frequency parts of the other channels as well, if they cannot handle low frequencies. This can be seen in Figure 2.2.

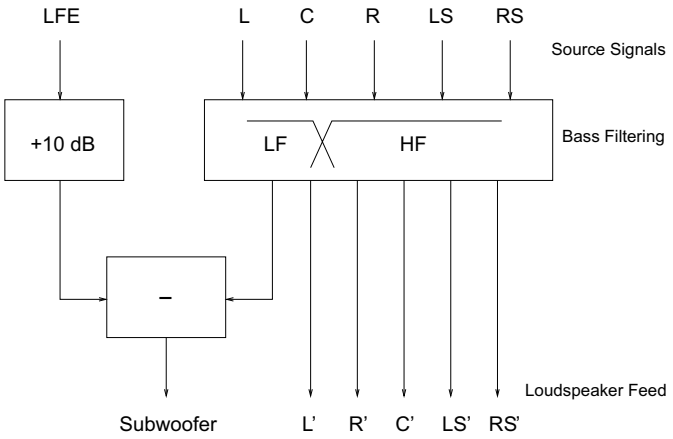



**Figure 2.2:** Derivation of combined subwoofer and LFE signals.

### 2.1.4 Description of the standard listening room

In order to reproduce the surround sound through headphones as if it was played through loudspeakers in a listening room, the room parameters of an typical multi-channel listening room must be taken into account. Some relevant room specifications, according to the AES-recommendation [Rumsey et al., 2001], are listed below.

#### Room dimensions

According to [Rumsey et al., 2001] a volume of $300\,\mathrm{m}^3$ should not be exceeded for studio listening rooms. The suggested room shape is symmetrical around the listening position, especially considering the loudspeakers, equipment, doors and windows. Other properties of a recommended listening room can be seen in Table 2.1.

**Table 2.1:** Suggestions for the reference multi-channel listening room [Rumsey et al., 2001].

| ***Parameter*** | ***Units/Conditions*** | ***Value*** |
|---|---|---|
| **Room size** | S [$m^2$] | $> 40$ |
| **Room proportions** | $L$ = length<br>$W$ = width<br>$H$ = height | $1.1\frac{W}{H} \leq \frac{L}{H} \leq 4.5\frac{W}{H} - 4$ with $\frac{L}{H} < 3$ and $\frac{W}{H} < 3$ (Ratios within $\pm 5\%$ of integer values are considered unsatisfactory.) |
| **Base width** | B[m] | $2.0 - 4.0$ |
| **Basis angle** | [deg] referred to left/right | 60 |
| **Listening distance** | D[m] | 2m to 1.7 B |
| **Listening zone** (radius) | R[m] | 0.8 |
| **Loudspeaker height** (from acoustic center) | h[m] | $\approx 1.2$ |
| **Distance to surrounding reflecting surfaces** | d[m] | $\geq 1$ |

The base width refers to the distance between the front loudspeakers, the basis angle to the angle between them with the model point in the center of the circle. The listening zone specifies the circle around the sweet spot, where the perception of the sound should be optimal.

### Reverberation time

Spaciousness and timbre are two qualities of the sound which are strongly influenced by reflections. Placement of absorbing materials affects the reverberation time. More details on reverberation are given in Section 2.2.3.

The mean reverberation time ($T_m$) is suggested to be between 0.2 s and 0.4 s. Moreover, the reverberation time for different frequencies should be in the tolerance field shown in Figure 2.3. $T_m$ is measured in one-third-octave bands from 200 Hz to 4 kHz.

The reverberation time for adjacent frequencies should not be too different, as that might influence the operational sound level curve. Thus, according to the AES-recommendation [Rumsey et al., 2001], such deviations should not excced $\pm 0.05$ s in the region of 200 Hz to 4 kHz and 25% of the longest reverberation time below 200 Hz.

### Background noise

The noise level in the listening room, which means noise caused by for example air conditioning and other external and internal sound sources, should be as low as possible. The continuous noise level should be measured in one-third-octave band averaged frequencies in the range of 50 Hz to 10 kHz and compared to the ISO noise rating (NR) curves [ISO1996, 1975], the noise level curve should not exceed the NR 10 curve and NR 15 curve is proscribed to be exceeded (the curves are numbered after their sound pressure level at 1 kHz) [Rumsey et al., 2001].

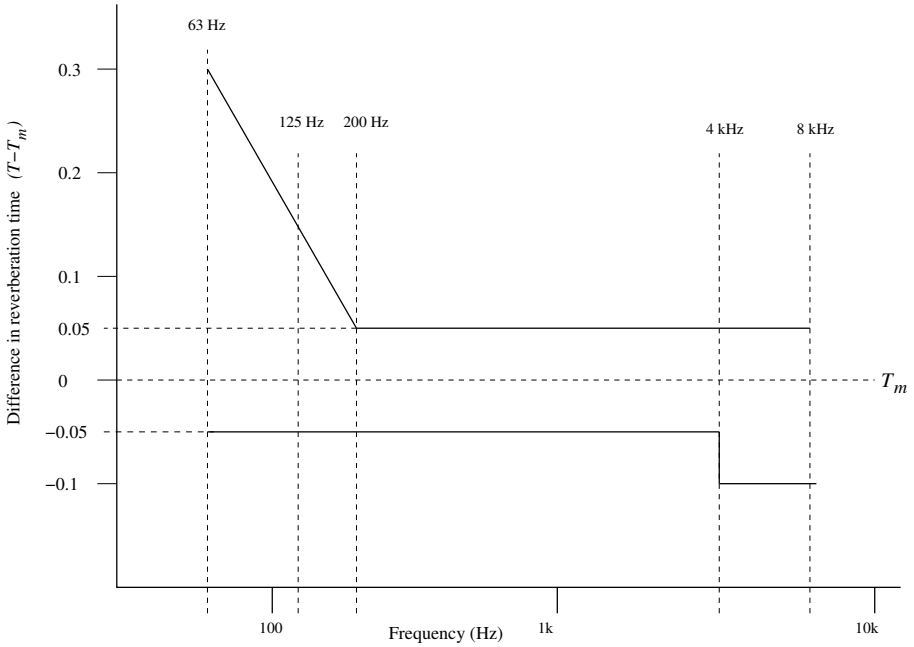



**Figure 2.3:** Tolerance mask for the reverberation time in a multi-channel listening room, relative to the arithmetic average value $T_m$.

## 2.2 Spatial hearing

In order to reproduce a realistic surround sound environment in headphones, it is necessary to know how humans perceive such a spatial environment. Normally a human will pick up sound with two ears which commonly is referred to as binaural hearing. In this case the sounds received at the two ears are compared and different auditory cues, like for example a difference in time of arrival, will make it possible for a subject to localize the sound source. To systemize localization in a spatial environment, a coordinate system is placed with the listener in the center as illustrated in Figure 2.4 [Blauert, 1997, p. 14].

The space is divided into a horizontal, median, and frontal plane which intersect roughly in the center of the head. Two angles are used in order to orient sound sources relative to the head. The azimuth ($\varphi$) describes the horizontal angle and is defined to be positive in the leftward direction while the elevation angle ($\delta$) is positive in the upward direction.

This section describes the principles of spatial hearing and looks into the problems and limitations that occur when a surround sound environment is reproduced through headphones.

### 2.2.1 Auditory cues for localization

When a sound source is not situated on the median plane the signals arriving at the two ears will be dichotic (non identical), in which case two different auditory cues are used for localization:

- Interaural Time Difference (ITD) occurs in response to differences in arrival time of a given sound. Consider a plane wave coming from a location with an azimuth

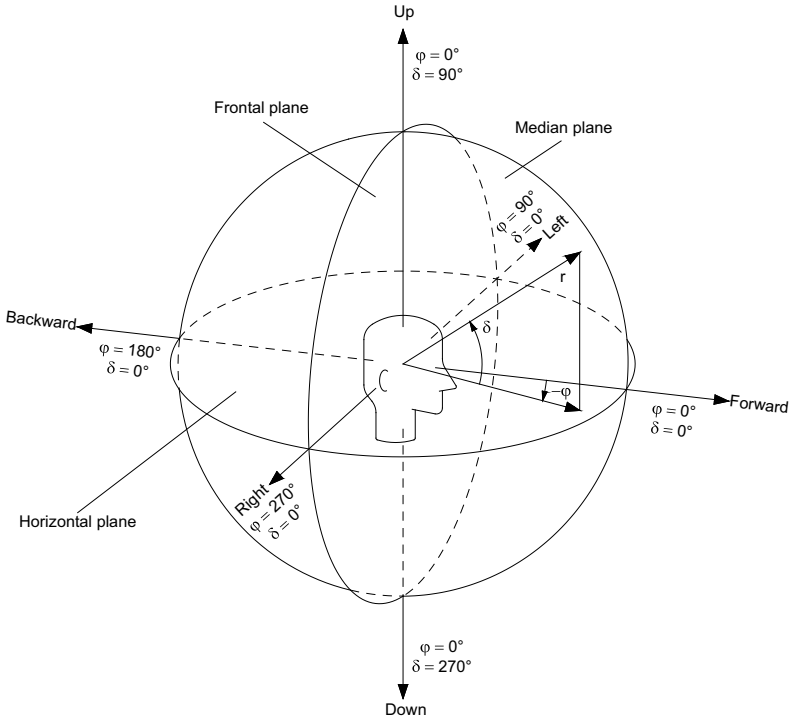



**Figure 2.4:** Coordinate system used in order to position sound sources in the space around a given subject, with $r$ being the distance, $\varphi$ the azimuth, and $\delta$ the elevation.

in the range $0° < \varphi < 180°$. The wave will first arrive at the left ear resulting in a phase shift between the two ears.

- Interaural Level Difference (ILD) occurs as a result of certain "shadow" effects cast by the head so that the sound pressure is larger at the ear closest to the source.

### Interaural time difference

Figure 2.5 illustrates a dichotic condition with a lateral sound source, simplified to a plane wave incidence.

The ITDs range from zero (corresponding to an azimuth of 0° or 180°) to about 690 $\mu$s for an azimuth of ±90°. However, the maximum time difference depends on the size of the head and the frequency of the signal [Moore, 2003, p. 236]. An approximation of the ITD for a given azimuth can be calculated by using the simplified approach illustrated in Figure 2.5. By assuming a spherical head shape and a far field sound source, the difference in path distance $\Delta l$ can be written as:

$$\Delta l = l_1 + l_2 = \sin(\varphi) \cdot r + \frac{\pi \cdot r \cdot \varphi}{180°} \tag{2.1}$$

where $r$ is the radius of the head. The before mentioned maximum ITD of 690 $\mu$s roughly corresponds to a head radius of 9 cm.

When using a sinusoidal signal, the ITD is equivalent to a phase shift between the two ears also referred to as an Interaural Phase Difference (IPD). Because of the limited distance between the ears, ambiguous situations can occur so that the auditory system

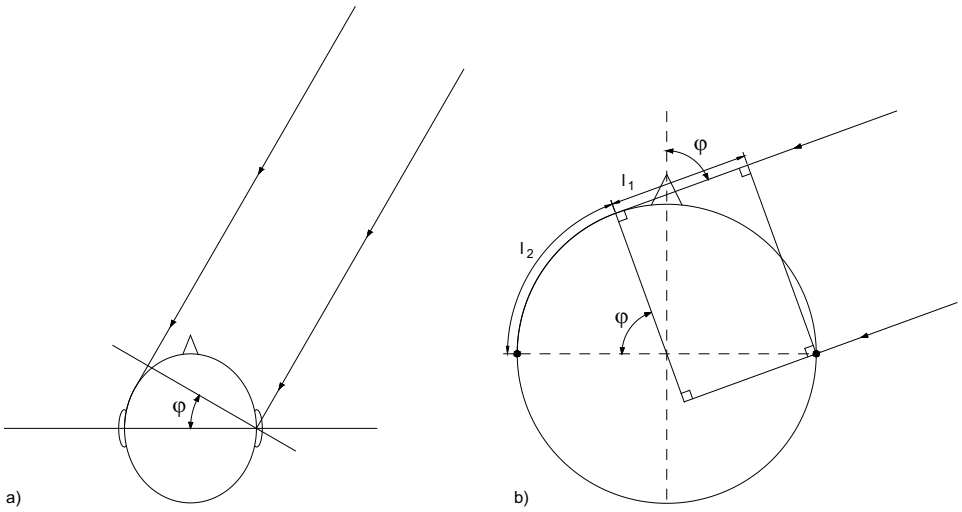



**Figure 2.5:** Lateral displaced sound source which results in non identical ear inputs; **a)** far-field approximation with a plane wave incidence further simplified in **b)** to a spherical shaped head in order to calculate the time of arrival difference between the two ears.

is unable to perform a correct localization through ITDs. This ambiguity is illustrated in Figure 2.6 in which case the auditory system might establish two different auditory events. However, in such a situation the auditory event closest to the median plane will dominate with respect to the other [Blauert, 1997, p. 147].

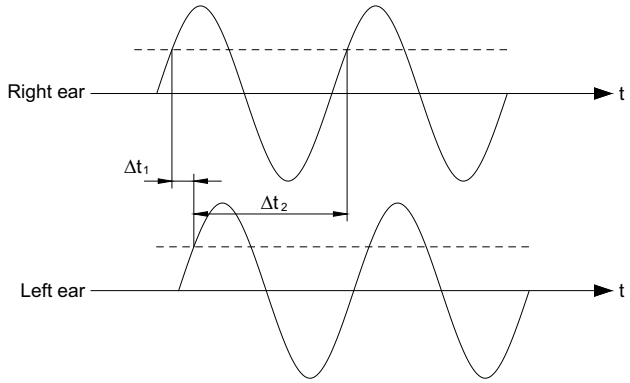



**Figure 2.6:** When working with periodic signals the interaural time difference is ambiguous so that more than one auditory event can occur.

A ITD of 690 $\mu$s corresponds to frequency of about 1450 Hz, which means that a 725 Hz tone at 90° azimuth will be identical to −90° azimuth because of the half-cycle shift. This means that localizing sound sources emitting pure-tone signals above 725 Hz is prone to more errors, while frequencies above 1500 Hz leads to no detectable ITDs [Moore, 2003].

### Interaural level difference

Contrary to ITD, changes in ILD are detectable over the whole audible frequency range. However, for sound sources distant to the listener ILDs are negligible for frequencies below 500 Hz, while sound sources close to the listener can give rise to considerable ILDs even for low frequencies [Moore, 2003, pp. 235-236]. The smallest detectable change in ILD is

about 1 dB when the reference ILD is zero as localization performance is best with sound sources in the front. At high frequencies the interaural level difference can be as much as 20 dB.

Sound source localization is often based on a combination of the two auditory cues ITD and ILD and referred to as the *duplex theory*. The idea is that sound localization is best based on ITDs at low frequencies and ILDs at high frequencies.

### Distance perception

Depending on the situation and the environment different cues are used for judging the distance to a given sound source. For example the intensity changes occurring when a listener is moving toward a sound source will result in a very accurate distance estimation [Moore, 2003, p. 265]. When listening in a room with reflecting surfaces the level and time differences between direct and reflected sound provides important cues for estimating the distance.

In general the precision of estimating the distance to a given sound source is relatively inaccurate, with errors around 20 % for unfamiliar sources. For nearby sources the estimations tend to be overestimated while being underestimated for sources situated far away [Moore, 2003, p. 266].

### Motional theories

In a situation with the head being kept stationary a given ITD caused by a lateral displacement of the source will not be unique. This arises because different source location can result in the same ITD, which is often referred to as the cone of confusion. This is illustrated in Figure 2.7 with the shape of the head simplified to a sphere.

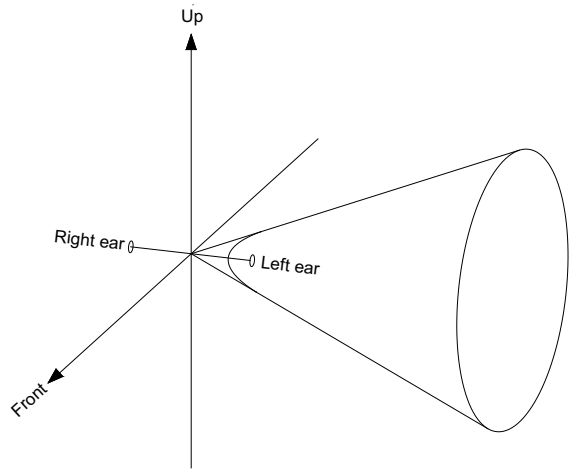



**Figure 2.7:** Example of a cone of confusion for a spherical head with the same ITD occurring for all sources placed on the surface of the cone.

When the head is moved relative to the sound source the interaural changes that occur are used to resolve the exact position of the source. If the head movement is not accompanied by any change in the auditory event the sound source will be perceived as being either directly above or below the listener [Moore, 2003, p. 249].

### Visual cues

Another important cue for correct localization is visual information about the spatial environment around the listener. When sounds are presented to a subject, the position of the auditory event is influenced by what the subject sees. A good example is when watching a person speaking on a television. In this case the auditory event is positioned where the speaker appears on the screen instead of the true loudspeaker position. Such expectation gathered by visual impressions is also the reason for often using curtains in listening experiments, so that the subject is not influenced by his or her expectations.

### Effect of the pinnae, head, and torso

Earlier it was described how lateral displaced sound sources can be localized through cues like ILD and ITD. However, if the head is considered completely symmetrical then neither of these cues will be available when the sound source is positioned on the median plane, in which case the signals at the two ears are diotic. In such cases, head movements can be used to resolve these ambiguities in the vertical direction. However, it is possible to judge the direction to a burst of white noise in the median plane even when the signal duration is too short for head movements to occur [Moore, 2003, p. 250]. A method for explaining such a phenomenon is by the directional filtering performed by the pinna. The main properties of the pinna are listed below:

- The pinna functions as an acoustic linear filter with the transfer function depending on the direction and distance of the sound source.
- The incoming signal is altered by the pinna in such a way that the sound source is perceived as being out in space [Moore, 2003, p. 250] compared with the lateralization inside the head often experienced when using headphones.
- Because of their short wavelength, frequencies above 6 kHz have the strongest interaction with the pinna [Moore, 2003, p. 251].
- The acoustic properties of the pinna are based on physical phenomena such as shadowing, reflection, diffraction, dispersion, resonance, and interference [Blauert, 1997, p. 63].
- The shape of the pinna and the corresponding transfer function is different from person to person, just like fingerprints.

When the sound propagates from a given source to the eardrum of a listener, the signal is not only filtered by the pinna, but also the torso and head can affect the spectrum. This means that modifications in the spectrum are not limited to frequencies above 6 kHz but can instead be found in the frequency range 500 Hz to 16 kHz [Moore, 2003, p. 251]. This is also often referred to as Head Related Transfer Functions (HRTFs), although these do not necessarily need to include the torso, as indicated by the name. In principle an infinite number of HRTFs exists for each ear, one for each possible spatial position of a sound source relative to the head. Because of individual shapes of pinna, head, and torso all these transfer functions will have individual characteristics for a given person.

### 2.2.2 Binaural technology

For reproduction of a spatial environment through headphones, two different techniques are commonly used; binaural recording and binaural synthesis. Binaural recording is performed by placing a microphone in each auditory canal of either a real head or an artificial head. In this way all the spectrum filtering done by the torso, head, and pinna will be included in the recordings and thus making it possible to recreate the spatial environment in a pair of headphones. Binaural synthesis refers to the use of HRTFs either measured with a real head or an artificial head. An example could be to measure the Head Related Impulse Response (HRIR) for each ear for two sound sources placed at $\pm 30^\circ$ azimuth, and afterwards convolving the left and right signal from a CD with the two HRIRs corresponding to $30^\circ$ and $-30^\circ$ respectively. By summing the signals and playing them back through headphones, a sensation of two loudspeakers standing in front of the listener is achieved instead of the usual in-head experience. It needs to be mentioned that such a process also requires proper equalization of the headphones and other relevant elements in the processing line, so that the signals presented at the ears are not coloured by unwanted characteristics. Further discussion and considerations on these matters are presented in Chapter 3.

#### Effect of non individual HRTFs

In both the recording and synthesis technique there is the choice of using either personal HRTFs or non individual HRTFs, which can originate from either another person or an artificial head. The use of personal HRTFs are often impractical or even more often quite impossible, for example in most commercial products. Because the ability to reproduce surround sound through headphones has a certain commercial value, it is important to evaluate the impact of using non individual HRTFs. In order to create believable virtual sound sources around the listener, it is important that the transfer functions obtained with an artificial head, to a certain degree, match the transfer functions of any possible listener. A large variety of artificial heads are available on the market. The one used for this project is VALDEMAR, which was developed at the Department of Acoustics at Aalborg University and is not a commercial product. The pinnae on VALDEMAR are casts of a human pinna, while head and torso are designed from acoustical measurements done on 40 subjects and from anatomical data [Minnaar et al., 2001].

In order to achieve the best possible spatial perception via binaural techniques, personal HRTFs should be used, in which case the localization precision rivals a *real life* situation [Møller et al., 1996b]. Deterioration occurs when non-individual recordings are used, especially in the median plane in which case the sounds received at the two ears are more or less identical. The amount of deterioration however, depends on the head used for a given recording which indicates that some ears are better for localization than others [Møller et al., 1996a]. Ideally an artificial head should represent such a *typical* head, but as shown by [Møller et al., 1999] and [Minnaar et al., 2001] this is not the case. Both investigations showed a significant deterioration in median plane localization with all the artificial heads used, included VALDEMAR although this particular head in general performed better [Minnaar et al., 2001]. No significant change was found for off median plane localization.

### 2.2.3 Considerations on room acoustics

In most real-life situations, sounds are perceived under reverberant conditions caused by reflecting surfaces surrounding the listener. Thus, it is necessary to evaluate how sound fields change under such conditions in order to assess its influence on a given subject's spatial perception.

When considering the sound transmission from source to receiver under non free-field conditions, the sound obtained at the receiving position will be a combination of direct sound and reflections. The direct sound will match the sound occurring under anechoic conditions, while the reflections are correlated to the surrounding environment. The arrival time of the reflections relative to the direct path depends on room dimensions and distance between source and receiver, which are also correlated with the signal amplitude as this is attenuated over distance. Further attenuation occurs by absorption properties in the reflecting surfaces, which in most cases are frequency dependent. Time of arrival and amplitude attenuation is illustrated in Figure 2.8, which shows the direct sound and early reflections for an impulse.

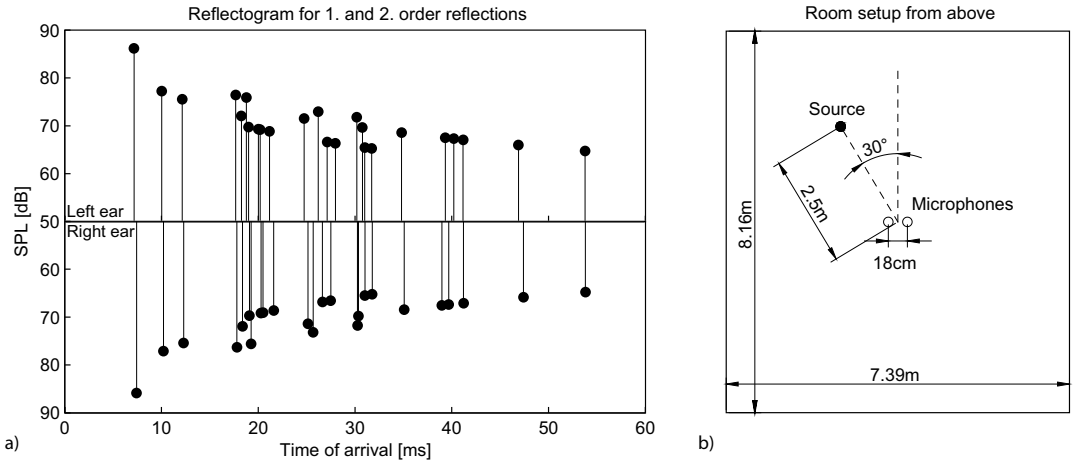



**Figure 2.8:** Room impulse responses with direct sound followed by first and second order reflections picked up by two microphones with the sound source positioned at azimuth $30^\circ$ and distance $2.5\,\mathrm{m}$. Both microphones and source are placed at a height of $1.2\,\mathrm{m}$. a) Reflectogram. b) Illustration of simulated room seen from above.

The illustration is created by simulating a source and two microphones in a room with dimensions $7.39\,\mathrm{m} \cdot 8.16\,\mathrm{m} \cdot 2.88\,\mathrm{m}$ $(w \times l \times h)$. The microphones are positioned in the center at a height of 1.2 m and distanced 18 cm apart to simulate two ears. The source is placed 2.5 m from the microphones at an azimuth of 30° relative perpendicular to a line between the microphones and at a height of 1.2 m. Reflection coefficients for the four walls are set to 0.8 while floor and ceiling is set to 0.5. Both source and receivers are considered omni directional while attenuation in amplitude is calculated according to the wave equation:

$$p = \frac{A}{r} e^{j(\omega t - kr)} \tag{2.2}$$

where $A$ is the amplitude of the source, and $r$ is the propagation distance of the given reflection.

As would be expected, the direct sound reaches the left ear first in which case the ITD will move the auditory event to the left. If the incoming reflections are considered in pairs (left and right ear) many of these will arrive at the right ear first. However, because of the so called precedence effect (or law of first wavefront), the auditory event will remain fixed towards the source [Blauert, 1997, p. 224]. Figure 2.8 only shows first and second order reflections while in reality the number of reflections will increase exponential with time as stated below [Blauert, 1997]:

$$\text{No. of reflections/s} = \frac{4\pi c^3}{V} \cdot t^2 \tag{2.3}$$

where $V$ is the volume of the enclosure, $c$ is the speed of sound, and $t$ is the increasing time.

Higher reflections reflections will overlap so much that the time function can only be described by statistical signal theory, which is the part of the reflectogram usually referred to as the reverberation of the room [Blauert, 1997, p. 277]. This means that the impulse response of a given room can be divided into three main parts, direct sound, early reflections, and reverberation. The influence of the two latter parts on localization is mainly limited to distance perception which cannot be resolved by the direct part alone. However, as the reflections come from all possible directions, the listener will be more enveloped in the sound, which will yield a sensation of spaciousness.

Another way to characterize a room with regard to its acoustic properties is to look at the reverberation time $T_{60}$. If a constant sound source is placed in a room, a steady state situation will arise after some time in which case the emitted power will equal the absorbed. The reverberation time is then defined to be the time from when such a sound source is turned off until the sound pressure level has decreased 60 dB. If unable to raise the sound pressure 60 dB above a given noise floor, $T_{20}$ or $T_{30}$ can be used, which measures either 20 dB or 30 dB decrease on the decay curve, and then scaling this time value to give an approximation of $T_{60}$. With proper knowledge about the acoustic properties of a given room an approximation of the corresponding reverberation time can be calculated by Sabine's formula [Kinsler et al., 2000]:

$$T_{60} = \frac{55.3V}{c \cdot A} \tag{2.4}$$

where $A$ is the equivalent absorption area.

Depending on the intended use of a room the ideal reverberation time will vary. For example with speech, a higher articulation can be achieved by lowering the reverberation time. However, this means that the room must be very absorbent, which will decrease the speech level and thus lowering the articulation [Maekawa and Lord, 1993]. As described earlier, the recommended reverberation time for a multi-channel listening room is between 0.2 s and 0.4 s, which can be considered a rather "dry" environment. This will ensure that the sound perceived by a listener is not radically changed compared to that intended by the sound engineer.

### 2.2.4 Spatial perception in a surround sound environment

So far the important concepts of spatial perceptions has been analyzed and presented. However, this theory must be considered in relation to the application of interest in order to assess possible limitations and error sources.

- As described before, deterioration in localization occurs when VALDEMAR is used as a substitute for personal HRTFs. However, as this is only significant for the median plane, the error is limited to confusion between the front and rear center loudspeakers. The sound originating from the front center is highly correlated with the particular image on the screen, which means that visual cues will help to maintain a frontal auditory event. Sound from the rear center channel is mainly limited to sound effects and often correlated with the left and right surround channels, for example when having the effect of an air plane flying from right to left behind the listener. Another factor that limits the change of having the rear sound perceived from the front is the fact that rear to front mix-up is less likely than front to rear mix-up, 7% against 30% respectively [Minnaar et al., 2001]. However, this is only truly valid for the specific setup and corresponding experimental procedure, and might not reflect the general probabilities for front-rear mix-up.

- An important cue for correct localization was the ability to correlate a head movement with a change in the perceived auditory event. However, these cues are not available if the surround environment is created with fixed virtual loudspeakers relative to the head. A method to compensate for this would be to position the loudspeakers relative to the screen, and then applying some kind of head tracking in order to move the loudspeakers relative to the listener. This requires a substantial amount of HRTFs, a seamless transition between these, and everything done in real time. Such a feature will essentially yield the most realistic environment, but the necessity can be argued. In most cases the listener will be in a more or less fixed position facing the screen, thus the realism will not be compromised by fixed loudspeakers. Another argument is that exact sound localization is not required, as the surround sound environment is meant to give a diffuse spatial feeling and stable sound events mainly occur from the front center channel. However, a head tracking solution might help the listener to become familiar with the virtual environment which is discussed next.

- Other factors that could improve the listening experience are learning and familiarity. It was shown by [Minnaar et al., 2001] that errors caused by non-individual binaural recordings decreases as a function of time, which indicates that subjects adapts to the given environment and HRTFs. When surround sound is reproduced through headphones, the listener will know the position of the virtual loudspeakers and thus know what to expect. Familiarization can be helped along by playing for example white noise in one loudspeaker at a time, which necessarily would be improved if a head tracking solution was applied.

- The HRTFs refer to the pressure difference between measurements done in a given position with and without a head. Ideally no other information should be included in such transfer functions, which also means that the HRTFs must be made under free-field conditions. However, as described in Section 2.1, the surround setup

should be situated in a room, which gives a more diffuse sound field and enables a sense of distance to the loudspeakers which is non-existent in an anechoic environment. The solution will be to either add a virtual room to the HRTFs or measure the transfer functions in a listening room so that these contain both head and room characteristics. The latter is referred to as Binaural Room Impulse Response (BRIR) measurements and is the most straightforward solution although the binaural synthesis will increase significantly in computational complexity as the length of the impulse response increases.

This leads to considerations on how long such BRIRs needs to be, as long BRIRs will describe the room better while shorter ones will be preferable with regard to computational cost. As described earlier, an impulse response of a room can be divided into three parts: direct sound, early reflections, and reverberation tail. If a room has a short reverberation time, the BRIR can correspondingly be truncated as significant information can be considered absent after this time interval. The fact that the recommendation for multi-channel rooms specifies a "dry" environment indicates that the room should add minimum room sensation to the movie soundtrack. Instead such things should be added by the sound engineer in charge of the soundtrack, as it will always be easier to add a room than remove it.

In summary it can be concluded that a realistic reproduction of a surround sound environment through binaural synthesis should be possible from a spatial hearing point of view. Other factors such as the characteristics of headphones and measurement equipment can have an undesirable effect on the outcome, but compensation for this can be made through signal processing.

## 2.3 System specifications and limitations

Based on the previous analysis and according to the main problem description, a more detailed system specification can be made, which is then to be used as a guideline in the following design phase. For better overview, the specifications are divided into blocks of which the first one specifies the system itself that has to perform all the binaural syntheses. This is then followed by a short description of how the final system is to be validated through several formal listening tests. Some considerations and choices for the different measurements needed are made in general terms while a more detailed description is given later in Chapter 4. Finally some specifications are made on the offline and real-time implementation methods.

**System design**

- The surround sound environment reproduced in the headphones must conform with the recommendations set in [Rumsey et al., 2001]. This can then either be done by obtaining BRIRs in a room fulfilling these requirements or using HRTFs and then simulating a virtual room surrounding the listener. It is chosen to use the BRIR method as a multi channel room is available in the facilities and because room simulation is not the focus of the project.

- Ideally the system should support the newest standards such as Dolby Digital 5.1EX and DTS-ES 6.1, but the decoding of such surround formats requires licenses which are not available. However, it is possible to find open source decoding of standard Dolby Digital 5.1, so the project is limited to using this format.
- It is commonly agreed that it is not possible or at least difficult to localize low frequencies, so that the exact position of the subwoofer is often not very crucial. For this reason, it is chosen to add the LFE channel directly to left and right ear in the binaural synthesis, instead of first convolving it with the corresponding BRIRs. This will give a reduction in the overall computational cost and thus making it more suitable for real time implementation.
- The global sampling frequency to be used throughout the processing chain is set to 48 kHz, as this corresponds to the one used for audio on DVDs.

**System validation**

- In order to assess the overall quality of the binaural synthesis it can be compared to some kind of reference situation. This could for example be a real 5.1 surround setup corresponding to the one used for obtaining the BRIRs. However, this can induce several errors/bias, as the subject will be aware of which method is used and might have some prejudice on sounds presented over headphones. Further bias can come from the fact that headphones are incapable of shaking the floor as the loudspeakers in a normal 5.1 setup will do. Instead, binaural recordings of different movie sequences will be made in the multi-channel room with exactly the same setup used when obtaining the BRIRs. If the binaural synthesis is done correctly, there should be no difference between this and the recordings.
- Because it is desirable to minimize the BRIR lengths as much as possible, in order to lower the computational cost, a listening test assessing differences between such lengths will be conducted. Ideally, variations in BRIR lengths should be done in small steps and thus finding the best compromise between sound quality and computational cost. However, as listening tests are highly time consuming it will be limited to three different lengths, selected according to relevant room and BRIR analysis.
- Finally, the consequence of using different equalization approaches for the headphones will be assessed. The goal here is to evaluate differences between equalizing the headphones for VALDEMAR, average of human subjects, or some general equalization curve that is not matched directly to the used headphones. The last case is in relation to having a commercial product that cannot be pre-equalized for a specific headphone type. More details on concrete reasons for doing this compensation of headphone characteristics and considerations on possible general equalization curves will be presented in the design phase (Chapter 3).

### Pre-considerations for measurements

- As described in the previous chapter, the best approach for creating binaural recordings or syntheses is to use individual HRTFs. However, this is not a practical solution when working with commercial products, and it would lead to undesirable complications when doing the listening experiment. Thus, the artificial head referred to as VALDEMAR, which was evaluated earlier, will be used in this project for all head related measurements and recordings.
- Although the system is limited to 5.1 instead of 6.1, and the LFE channel is passed directly to the output, all measurements required in a more comprehensive system will be made. This then includes measurements for a rear center and subwoofer channel. Furthermore, the measurements will not only be conducted in the multichannel room, but also in the anechoic chamber, and thus obtaining all the HRTFs as well.
- As discussed above, headphone measurements on both VALDEMAR and humans are required. However, headphone measurements on human subjects have recently been conducted by Emine Çelik, and these will be borrowed for this project, as such measurements are not within the project objectives.

### Offline implementation

- When doing all the initial system tests, and in the process of creating the needed samples for the listening experiments, it is desirable to have a flexible program in which relevant parameters can be changed quickly and reliably. All simulations not required to run in real-time are done in MATLAB and a graphical user interface (GUI) will be created to handle all the different parameters. The goal is reduce the possibility of creating a binaural synthesis with wrong parameter.

### Real-time implementation

- As specified in the problem description, a real-time version of the binaural synthesis system is to be implemented as a plug-in for PC-based movie players. It is not within the project objectives to implement a decoder for the AC-3 stream, which consequently narrows the options for such a real-time realization. The decoding of Dolby Digital and DTS requires licenses if used in commercial products, and only a few freeware solutions exists, of which most are meant for offline processing. However, the “AC3Filter” is an open-source DirectX filter, which enables any DirectSound-compatible player to decode AC-3 streams. The goal is to insert a binaural synthesis algorithm into this program that can perform the necessary processing just before “AC3Filter” sends the decoded signals to DirectSound.

Chapter 3

# Equalization of the Recording and Reproduction Chain

In the analysis in the previous chapter, it was explained how the pinna, head, and torso acoustically filter the incoming sound. To record and afterwards reproduce the correct filtered sound through headphones, it is necessary to calibrate the entire recording and reproduction chain to remove any unwanted influences. In this chapter, the two chains will be analyzed and filters will be designed to obtain the goal of correct reproduction.

## 3.1 Modelling the ear

The frequency response of a pair of headphones cannot be measured in the same way as it is done with loudspeakers, i.e. placed in an anechoic room and measured in the far-field. Headphones are meant to be used in the near-field, so the measurement should be designed for this.

In [Møller et al., 1995a], a method for calibrating the binaural recording and reproduction chain has been introduced. The sound transmission path of the ear is divided into two parts; one that affects the spatial properties of the incoming sound and one that does not. It is shown that the transfer function of the ear canal is not dependent on the direction of the incoming sound wave, so the complete spatial information is contained in the sound pressure anywhere in the ear canal, even when measured at the entrance of the ear canal. This also applies to blocked entrance ear canals, enabling the use of larger, more noise-free and practical microphones. VALDEMAR is made to utilize this method, and is constructed with microphones at the blocked entrance ear canals.

From this it follows that with VALDEMAR, the goal for binaural reproduction is to reproduce the correct sound pressure at the entrance of the ear canal. Some considerations are needed to ensure that a given headphone can provide this.

In [Møller et al., 1995a], it is assumed that the human ear can be approximated by an electrical analogue Thevenin model, as shown in Figure 3.1.

$P_{open}$ is the sound pressure at the entrance of the non-blocked ear canal, $P_{open}$ is the sound pressure at the entrance of the blocked ear canal, $P_{eardrum}$ is the sound pressure at

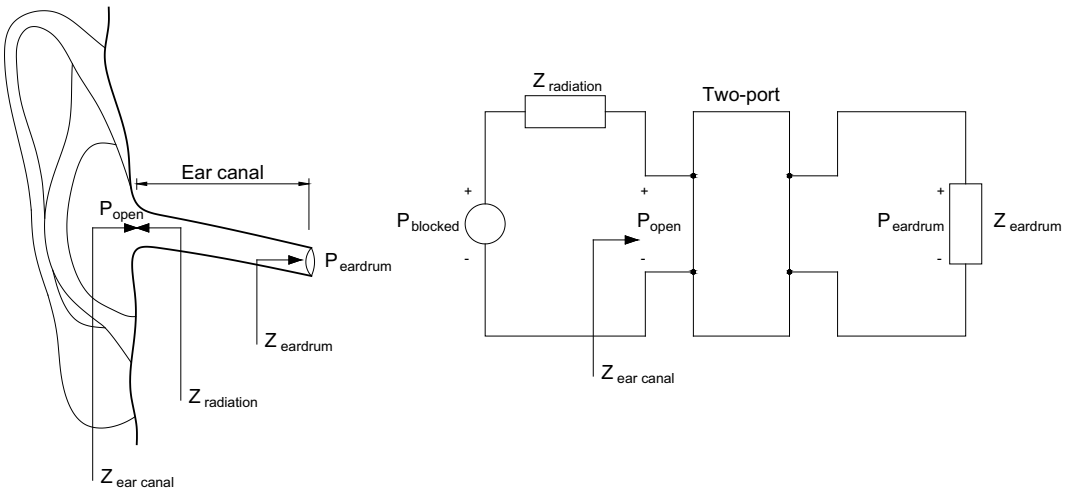



**Figure 3.1:** A simple model of the ear and the corresponding electrical equivalent circuit (from [Møller et al., 1995a]).

the eardrum, $Z_{\text{ear canal}}$ is the impedance of the ear canal when looking into the open ear canal, $Z_{\text{radiation}}$ is the impedance in the same point looking outwards into free air, and $Z_{\text{eardrum}}$ is the impedance of the eardrum from the ear canal.

At this point it is worth to notice that the impedance $Z_{\text{radiation}}$ will change when a headphone is placed over the ear, as the ear canal is no longer looking out into free air. If the positioned headphones are circumaural and form an airtight seal around the ear, the ear canal will look out into a small air cavity and the headphone diaphragm.

Using the electrical equivalent circuit, the relationship between the open and the blocked ear canal pressures can be expressed as a pressure division between $Z_{\text{radiation}}$ and $Z_{\text{ear canal}}$:

$$\frac{P_{\text{open}}}{P_{\text{blocked}}} = \frac{Z_{\text{ear canal}}}{Z_{\text{radiation}} + Z_{\text{ear canal}}} \tag{3.1}$$

Here it is seen that dependent on the value of $Z_{\text{radiation}}$, a recording made with a blocked ear canal might have to be altered in order to provide the correct sound pressure when it is no longer blocked. It is also seen that when $Z_{\text{radiation}}$ change due to a headphone being placed over the ears, the relationship between $P_{\text{open}}$ and $P_{\text{blocked}}$ will change. These considerations will be dealt with later in this section, while they will be considered insignificant at this point.

When considering a complete recording and reproduction chain, several things have to be accounted for. The recording chain includes the ear canal microphone, amplifier and storage device, and the playback chain includes playback equipment, amplifier and headphones. Most of the equipment (cables, amplifiers, storage and playback equipment) is considered to have a flat frequency response within the audible range, and can be neglected as having any influence to the properties of the reproduced binaural sound. Two key elements are not included in this consideration, and that is the ear canal microphones and the headphones used in the reproduction. The transfer functions of the microphone and headphone has to be compensated in some way, and to do this, the complete transfer function of the recording/playback chain should be calibrated to become 1, that is, perfect reproduction of the recorded sound. To get an overview of how this is achieved, the

transfer function is considered from the sound pressure at the entrance of the blocked ear canal, through the microphone, through the electrical equipment, through the headphone and finally back as a sound pressure at the entrance of the blocked ear canal. This is written in Equation (3.2), and illustrated in Figure 3.2.

$$\frac{P^*_{blocked}}{P_{blocked}} = \frac{E_{microphone}}{P_{blocked}} \cdot H_{electrical} \cdot H_{calibration} \cdot \frac{P^*_{blocked}}{E_{headphone}} \tag{3.2}$$

where $P_{blocked}$ is the sound pressure at the blocked entrance ear canal in the recording situation, $P^*_{blocked}$ is the sound pressure at the blocked entrance ear canal in the reproduction situation, $E_{microphone}$ is the output voltage from the microphone, $E_{headphone}$ is the voltage on the headphone input terminals, $H_{electrical}$ is the transfer function of the electrical chain, and $H_{calibration}$ is the transfer function of a filter that will be designed to compensate for the headphone and microphone transfer functions.

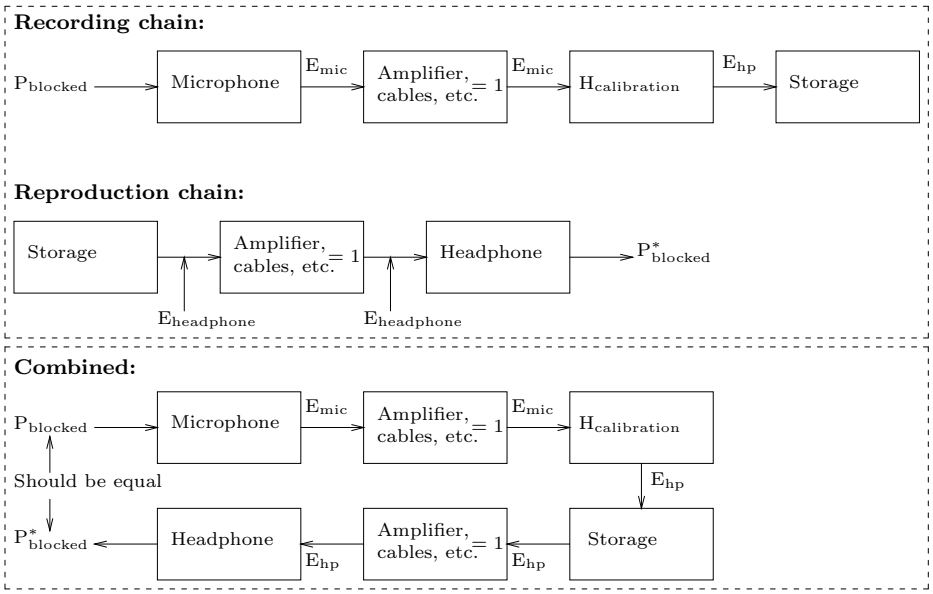



**Figure 3.2:** Illustration of the transfer function derivation of the recording and reproduction chain.

The desired perfect response is obtained when $P^*_{blocked}$ equals $P_{blocked}$, that is:

$$\frac{P^*_{blocked}}{P_{blocked}} = 1$$

As mentioned previously, $H_{electrical}$ is considered to have a flat frequency response within the used frequency range, which means that it can be represented as a scalar, representing the gain of the electrical chain. As the playback volume will be adjustable for the user, the value of this scalar is not currently important, and can be set to 1 for simplicity.

Now Equation (3.2) can be simplified to:

$$\frac{E_{microphone}}{E_{headphone}} = \frac{1}{H_{calibration}}$$

Solved for $H_{calibration}$:

$$H_{\text{calibration}} = \frac{E_{\text{headphone}}}{E_{\text{microphone}}} \tag{3.3}$$

This result indicates that the best recording and reproduction (using blocked ear canals) is obtained when setting $H_{\text{calibration}}$ to the inverse of the transfer function measured from the headphone input terminals to the microphone output. The actual transfer function can be measured by positioning the headphones on the recording head with the microphone in the blocked ear canal, and then measure the transfer function from the headphone input terminals to the microphone output terminals, using for example the MLS measurement technique.

The measured transfer function will most likely not have a flat frequency response due to the characteristics of the headphone and microphone. The problem is now to design an inverse filter that will have the opposite frequency response, so that when it is inserted into the signal chain, it will result in a flat frequency response of the whole chain. This problem is handled in Section 3.4.

### Adaptation for open ear canal

It is necessary to consider what will happen when the headphone is no longer playing into a blocked ear canal as it was calibrated for, but into an open ear canal. This will be the normal listening situation.

Equation (3.1) is now written in two versions, one with the headphone loading the ear canal, and one without. $Z_{\text{headphone}}$ will be used as the radiation impedance when the ear canal is looking out into the headphone, and $Z_{\text{free air}}$ will be used for the impedance of free air:

$$\frac{P^*_{\text{open}}}{P^*_{\text{blocked}}} = \frac{Z_{\text{ear canal}}}{Z_{\text{headphone}} + Z_{\text{ear canal}}} \qquad \text{(with headphone)} \tag{3.4}$$

$$\frac{P_{\text{open}}}{P_{\text{blocked}}} = \frac{Z_{\text{ear canal}}}{Z_{\text{free air}} + Z_{\text{ear canal}}} \qquad \text{(without headphone)} \tag{3.5}$$

Equation (3.5) is valid during recording and Equation (3.4) is valid during playback. The ratio between these two pressure divisions is known as the Pressure Division Ratio (PDR), and is calculated as:

$$\text{PDR} = \frac{P_{\text{open}}}{P^*_{\text{open}}} \frac{P^*_{\text{blocked}}}{P_{\text{blocked}}} = \frac{Z_{\text{ear canal}} + Z_{\text{headphone}}}{Z_{\text{ear canal}} + Z_{\text{free air}}} \tag{3.6}$$

Now it can be seen, that the PDR serves as a correction factor to the system transfer function, when the recording is made without a headphone loading the ear canal, relative to the playback situation, when the ear canal is loaded with the headphone.

The effect of the PDR varies with the type of headphone used. If the headphone impedance $Z_{\text{headphone}}$ is close to the free air impedance $Z_{\text{free air}}$, the PDR reduces to unity and can be neglected. In [Møller et al., 1995a], this headphone property is called FEC (Free air Equivalent Coupling), and it is generally valid for open-type headphones.

A headphone with the FEC property is preferred, and an example of a headphone with this property is the beyerdynamic DT990pro [Møller et al., 1995a]. The beyerdynamic DT990pro headphone is a high quality headphone that is comfortable to wear and has low distortion, making it suitable for listening tests. For this reason, the DT990pro is chosen as the main headphone model that will be used throughout the project.

## 3.2 General frequency response of headphones

Headphones are typically designed for reproduction of music signals originally intended for loudspeaker reproduction. This acoustic transmission path includes the transfer function of the head and ear, and for this reason, headphones are often made to simulate the transfer function from a loudspeaker to the ear canal. This should maintain the timbre of the music when listening to stereo-recordings through headphones. For correct reproduction of binaural signals, this loudspeaker simulation is unwanted and removed as shown in Equation (3.3).

Several methods exist to design headphone transfer functions, but the free-field equalization and the diffuse-field equalization are believed to give the best compromise. Design goals for the two methods are derived in [Møller et al., 1995b], and their results are shown in Figure 3.3.

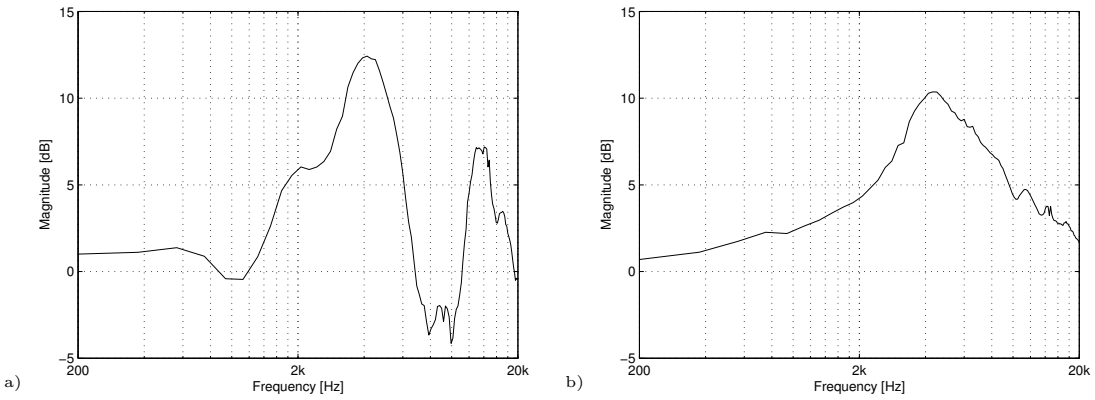



**Figure 3.3:** From [Møller et al., 1995b], typical design goals for headphone calibration; **a)** free-field, **b)** diffuse-field.

The free-field equalization is based on frontal incidence of sound in a free-field, e.g. an anechoic room. It serves to simulate the frequency response of a loudspeaker when placed in the middle in front of the listener. The free-field design target is similar to the HRTF for frontal incidence.

The diffuse-field equalization is based on an average of the HRTFs for all possible incidence angles. This will to some extend simulate the frequency response of a loudspeaker in a reverberant room.

The calculation of the diffuse-field transfer function can be written as [Møller et al., 1995b]:

$$\text{Diffuse-field transfer function} = \sqrt{\frac{1}{4\pi}\int\int |\text{HRTF}(\varphi,\theta)|^2 d\varphi d\theta}$$

As the diffuse-field equalization is equally based on all directions, it does not imply a specific source location, but is generalized as a sound source outside the head.

According to the specifications, the DT990pro is supposed to diffuse-field calibrated.

## 3.3 Obtaining correct HRTFs

In an anechoic environment, the HRTF is defined as [Hammershøi and Møller, 1996]:

$$\text{HRTF} = \frac{\text{P}_{\text{ear canal}}}{\text{P}_{\text{head}}}$$

where $\text{P}_{\text{ear canal}}$ is the sound pressure at a point in the ear canal, and $\text{P}_{\text{head}}$ is the sound pressure at the center of the head (with the head absent). Throughout this project, $\text{P}_{\text{ear canal}}$ will be defined as the sound pressure at the entrance of the ear canal. The measurement of an HRTF should be done in an anechoic room, as the HRTF only includes the head, and not a room response or any of the measurement equipment.

Measurement of $\text{P}_{\text{ear canal}}$ is done by placing a recording head, with a microphone at the blocked entrance ear canal, in an anechoic room and a loudspeaker in the far-field of the head, and then measure the transfer function from the loudspeaker input to the microphone output. Measurement of $\text{P}_{\text{head}}$ is done in a similar way, by using a free-field microphone in place of the recording head. The two measurement chains can be seen in Figure 3.4.

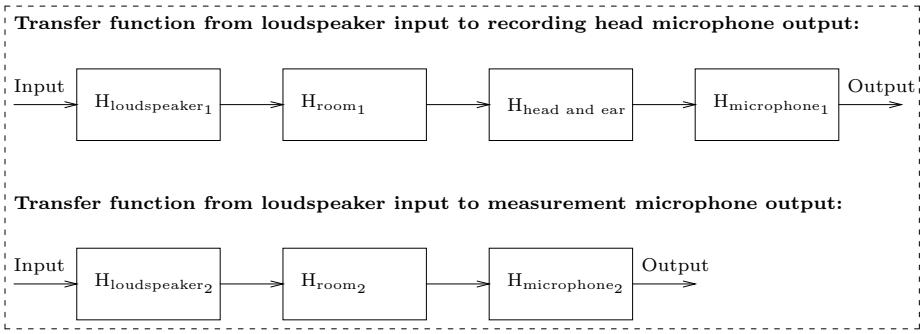



**Figure 3.4:** Block-diagram of the transfer functions in the measurement chains used for obtaining the HRTF.

The resulting measurement can be mathematically described as:

$$\text{HRTF} = \frac{\text{H}_{\text{loudspeaker}_1} \cdot \text{H}_{\text{room}_1} \cdot \text{H}_{\text{head and ear}} \cdot \text{H}_{\text{microphone}_1}}{\text{H}_{\text{loudspeaker}_2} \cdot \text{H}_{\text{room}_2} \cdot \text{H}_{\text{microphone}_2}} \tag{3.7}$$

where $\text{H}_{\text{loudspeaker}_n}$ is the loudspeaker transfer function in room $n$, $\text{H}_{\text{room}_n}$ is the transfer function of room $n$, $\text{H}_{\text{microphone}_n}$ is the transfer function of the microphone used in room $n$, and $\text{H}_{\text{head and ear}}$ is the transfer function of the head and ear.

$\text{H}_{\text{loudspeaker}_1}$ and $\text{H}_{\text{loudspeaker}_2}$ are equal (same loudspeaker), so they will cancel out. The two rooms, $\text{H}_{\text{room}_1}$ and $\text{H}_{\text{room}_2}$ are physically the same anechoic room, but the distance between the loudspeaker and microphone is slightly different in the two measurements so they do not cancel out completely, but correspond to a small delay, which will be denoted as:

$$\mathrm{H}_{\text{delay-diff}} = \frac{\mathrm{H}_{\text{room}_1}}{\mathrm{H}_{\text{room}_2}}$$

$\mathrm{H}_{\text{delay-diff}}$ is a scaled impulse at a position corresponding to the time delay difference between the two measurement positions (entrance of the ear canal relative the center of the head). The consequence of this, is that 50% of the measurements will be non-causal, as this scaled impulse will be positioned before $t = 0$ in the situations where the sound reaches the center of the head before it reaches the ear. In practical situations, the HRTFs are delayed to ensure causality.

$\mathrm{H}_{\text{microphone}_1}$ and $\mathrm{H}_{\text{microphone}_2}$ will not be completely identical, but both microphones can be assumed to have a flat frequency response in the audible range, and thus neglected.

Equation (3.7) can now be simplified to:

$$\mathrm{HRTF} = \mathrm{H}_{\text{delay-diff}} \cdot \mathrm{H}_{\text{head and ear}} \tag{3.8}$$

This result shows that the definition of the HRTF includes not only information about the filtering of the head and torso, but also information about the time delay difference between the entrance of the ear canal and the center of the head.

#### The Binaural Room Impulse Response

The BRIR is related to the HRTF, with the addition that the BRIR also includes a binaural representation of a room. The BRIR includes information about all acoustical properties of the measured room, including the position and orientation of source and receiver, the room dimensions, reflecting surfaces etc.

Measurement of a BRIR is done by placing a binaural recording head in a room together with a source, and then measuring the impulse response from source to receiver. The frequency response of the source should then be compensated to obtain the response of the room itself. When measured in an anechoic room the BRIR will be almost similar to the HRTF.

## 3.4 Creating inverse filters

The previous section concluded that for perfect reproduction, it is necessary to apply an inverse filter to compensate for the headphone and microphone transfer functions. This section will discuss methods to accomplish this goal.

A standard digital filter is described by the difference equation:

$$\begin{aligned} a_1 y[n] &= b_1 x[n] + b_2 x[n-1] + \cdots + b_{n_b+1} x[n-n_b] \\ &\quad -a_2 y[n-1] - \cdots - a_{n_a+1} y[n-n_a] \end{aligned}$$

where **a** and **b** are polynomials, containing the coefficients $a_1 \ldots a_{n_a}$ and $b_1 \ldots b_{n_b}$ respectively, $x$ is the input signal, and $y$ is the filtered output.

If the filter is a Finite Impulse Response (FIR) filter, it contains no **a** polynomial coefficients, except for $a_1 = 1$. If both the **a** and **b** polynomias exist, the filter is an Infinite Impulse Response (IIR) filter, due to the feedback coefficients in the **a** polynomial. The zeros of the filter are the roots of the **b** polynomial and the poles are the roots of the **a** polynomial.

### 3.4.1 Stability criterions

The main criterion for the derived inverse filter is the need to be stable. For a causal filter, this is ensured if all poles of the filter are situated inside the unit-circle in the z-plane ($|z| < 1$). More specific, for a system to be stable, the region of convergence (ROC) must include the unit-circle ($|z| = 1$), and if the system is causal, the ROC will extend from the outermost pole and outwards [Oppenheim et al., 1998, p. 247]. From this it follows, that the outermost pole must lie inside the unit-circle. To illustrate this, an example of a pole-zero plot of a Chebychev 10th order low pass filter is plotted in Figure 3.5a. Poles are illustrated with '×' and zeros with '○'.

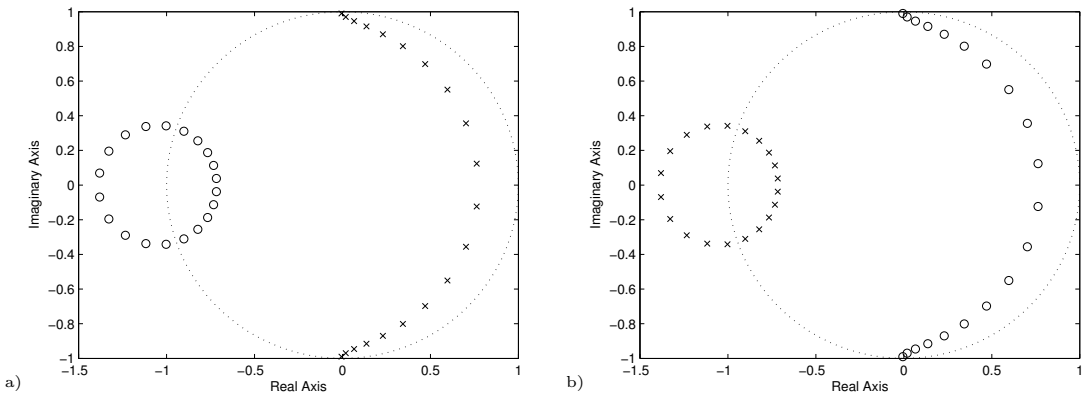



**Figure 3.5:** Pole-zero plot of **a)** Chebychev 10th order low pass filter, **b)** the direct, but unstable, inverse.

The direct inverse filter is made by changing the zeros into poles and vice versa. All poles on Figure 3.5a lie within the unit-circle, which makes the filter stable. All zeros, however, are not located within the unit-circle, so when the direct inverse filter is made (Figure 3.5b), the zeros outside the unit-circle becomes poles, thereby creating an unstable filter.

Filters with all zeros and poles within the unit-circle are called minimum-phase filters. Minimum-phase filters and their inverse are always stable and causal, so this behaviour can be desirable. Any filter can be divided into two sections, a minimum-phase section with a magnitude response equal to the original filter and an all-pass section with a magnitude response of 1, which only modifies the phase. The Chebychev filter in Figure 3.5 is divided into a minimum-phase section and an all-pass section, and the resulting pole-zero plot is shown in Figure 3.6. If the minimum-phase section and all-pass section are multiplied together, it will result in the original Chebychev filter as a pole and a zero will cancel out, if their positions on the pole-zero plot are the same.

From this, it follows that if the phase response is not important, the minimum-phase section can be used as a representative for the original filter.

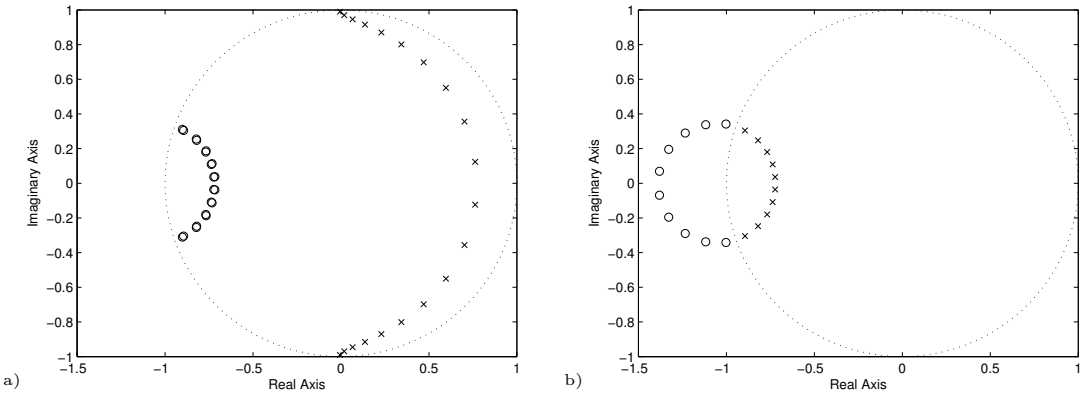



**Figure 3.6:** Chebychev filter from Figure 3.5 divided into **a)** minimum-phase section and **b)** all-pass section. Notice that the zeros in **a)** are double, i.e. there are two zeros almost at the same position.

### 3.4.2 Considerations for headphone equalization

When equalizing headphones, the theoretically ideal inverse function may not be the best choice. Figure 3.7a shows the frequency response of a beyerdynamic DT990pro headphone and Figure 3.7b the target frequency response for the ideal inverse filter.

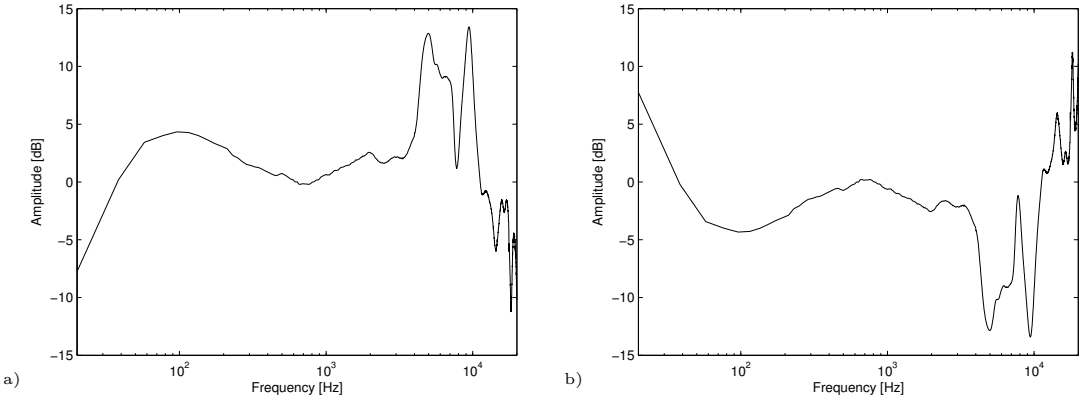



**Figure 3.7:** **a)** Frequency response of a typical headphone, **b)** the frequency response of the ideal inverse filter.

Theoretically, the two filters will cancel out perfectly to a flat frequency response, but due to changes in the temperature, humidity, wear, placement etc., the frequency response changes slightly every time it is measured. Figure 3.8a shows six frequency responses, all measured successively on the same headphone, but with repositioning of the headphone between each measurement. As it is seen, the responses are not completely identical, thus a very accurate inverse filter derived from a single measurement is not the ideal solution, and some kind of averaging is needed to give a general solution. The average amplitude response is shown in Figure 3.8b. Across different headphones, even similar models, the deviation is much larger as described in Section 5.2.1.

When creating the inverse filter, it is generally a good idea to avoid high peaks in the frequency response, as the ear is very sensitive to boosting narrow bands or single fre-

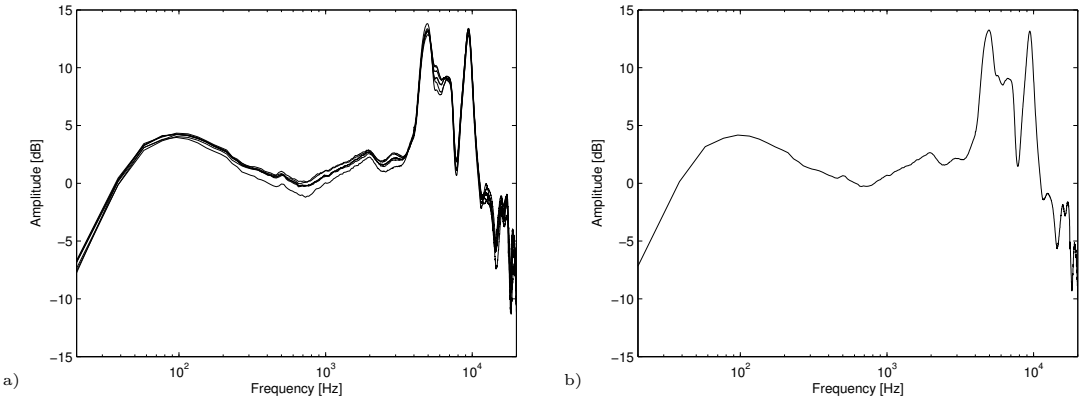



**Figure 3.8:** **a)** Six successively measured frequency responses of the same headphone, **b)** the mean of the frequency responses plotted in **a)**.

quencies. It could generate a "ringing" effect if the headphone response does not have a corresponding dip at the exact same frequency. Narrow band dips, on the other hand, are not as audible. The suppression of peaks and dips will happen naturally if the frequency response is calculated as an average across many measurements and many different headphones (of the same model).

The chosen headphone model will have a natural roll-off at high and low frequencies. The cutoff frequencies will vary with the headphone type, but beyond these frequencies, it is not desirable to let the inverse filter boost the amplitude too much. If a headphone with poor low-frequency response has the low frequencies boosted too much, it could damage the headphone due to high cone excursion and excessive heating of the loudspeaker coils because of the low efficiency.

When the target frequency response is chosen, there exist several methods to generate filters that will approximate the chosen frequency response. Two methods will be discussed, mainly from a practical point of view. They are both using a least squares method to approximate the frequency response. The first is the MATLAB *yulewalk* function, which uses the modified Yule-Walker equations and then uses an iterative algorithm to optimize the filter in the time domain. The second method is the MATLAB *invfreqz* method that creates an estimate by solving (in the frequency domain) a linear system and then minimizes the squared error using an iterative Gauss-Newton algorithm.

The process of limiting peaks in the resulting filter will be done in the shaping of the target function, so the goal of the filter will be to follow the target function as best as possible. A typical headphone transfer function is used as a basis for the target function that is used to test both methods.

As the filtering is performed offline directly on the BRIRs (stated in Chapter 5.3), the goal is simply to get the smallest error when compared to the target function, with low regard to filter order.

### 3.4.3 Shaping of the target function

As mentioned in the previous section, the headphone frequency response of high and low frequencies should not be corrected by the designed inverse filter, if the result can lead

to too high amplification. To limit these extremes, a lower and upper cutoff frequency is selected. By experimentation, it was found useful to create a linear frequency roll-off from the magnitude at the limit-frequencies towards 0 dB at $\frac{Fs}{2}$ and 0 Hz respectively. The method works best if the magnitude response is normalized to lie around 0dB. The normalization is also important to get the desired filters, which should have a gain as close to 0 dB as possible.

To remove the small variations and narrow peaks and dips in the frequency response of the target function, the shape is low pass filtered through a moving-average filter using the MATLAB *filtfilt* function. The result is a smoothed curve without any phase shift. For example, the Yule-Walker algorithm fails to create a filter if there is too abrupt magnitude or frequency changes in the vectors that define the target function. The smoothing of the magnitude response can help to avoid such problems.

An example of a result target function is shown in Figure 3.9 together with the original, non-processed inverse filter magnitude response.

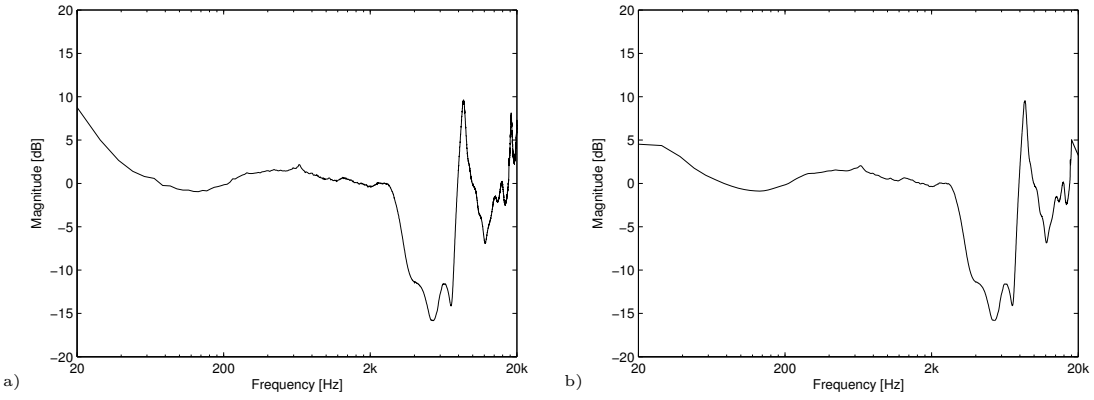



**Figure 3.9:** Two versions of the target function for the equalization filter with the limit frequencies 30 Hz and tunit18kHz. **a)** The original, **b)** modified version with smoothing and high/low end frequency response limited.

The target function in Figure 3.9b is used to test the *yulewalk* and *invfreqz* methods in MATLAB. Plots are shown for visual inspection of the frequency responses, and the mean-squared error between target and filter is calculated from the following equation:

$$\text{error} = \frac{1}{N}\sum_{n=1}^{N}\left(20\log\frac{T(n)}{F(n)}\right)^2$$

where $T(n)$ and $F(n)$ is the frequency response of target and filter respectively, N is the length of $T(n)$ and $F(n)$, and $n$ is the frequency bin.

### 3.4.4 Evaluation of equalization methods

The *yulewalk* function takes three arguments; the order of the resulting filter, and two vectors describing the desired magnitude and the corresponding frequencies.

The target function shown in Figure 3.9b is passed to the Yule-Walker algorithm. Results

for 15th, 30th and 70th order IIR filters can be seen in Figures 3.10, 3.11 and 3.12 respectively.

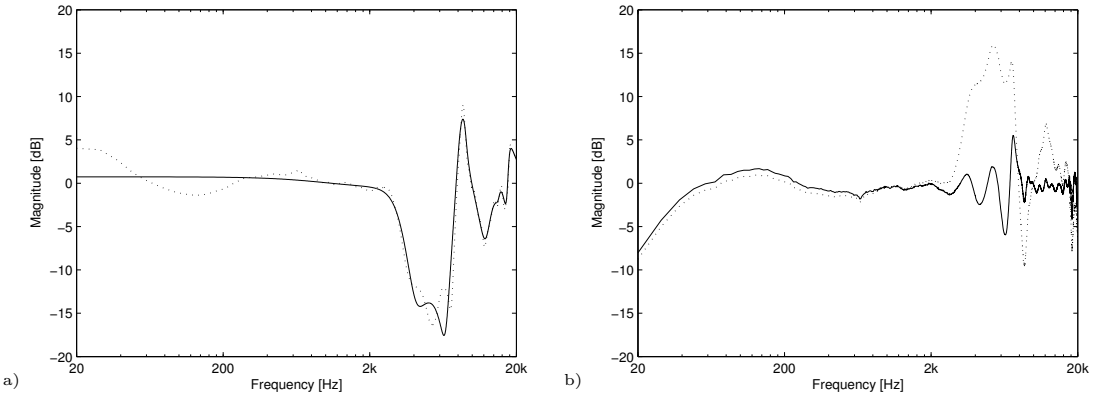



**Figure 3.10:** Example with using the Yule-Walker algorithm to fit a 15th order IIR filter to the target function. The mean-squared error is calculated to $1.891\,\mathrm{dB}$. **a)** Target function (dashed) vs. filter response (solid). **b)** Filtered headphone response (solid) vs. original headphone response (dashed).

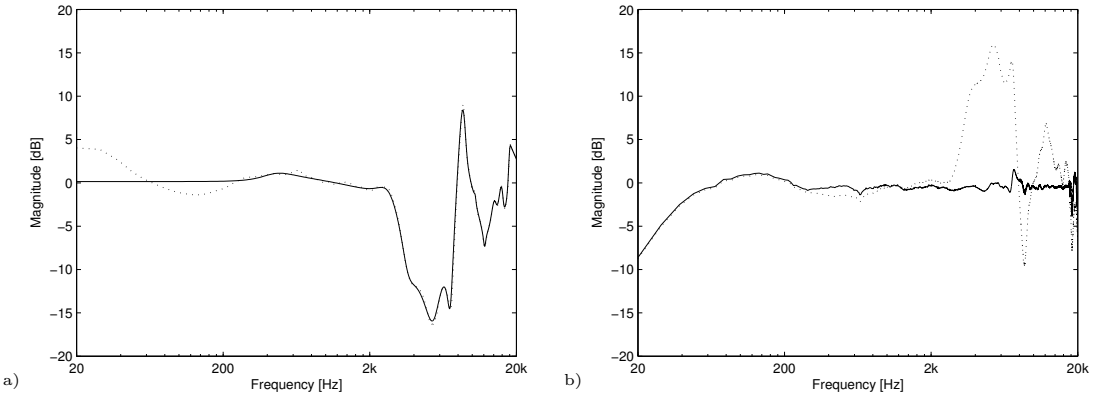



**Figure 3.11:** Example with using the Yule-Walker algorithm to fit a 30th order IIR filter to the target function. The mean-squared error is calculated to $0.144\,\mathrm{dB}$. **a)** Target function (dashed) vs. filter response (solid). **b)** Filtered headphone response (solid) vs. original headphone response (dashed).

As both the plots and the calculated mean squared error indicates, there is a big improvement going from a 15th to a 30th order filter, while the improvement is not so big when stepping up from 30th to 70th order. The main improvement from 30th to 70th order filter can be seen in the low frequencies, where the filter gets closer to the target function.

The error for low frequencies ($<400\,\mathrm{Hz}$) is still considered too high, as there is practically no equalization for those frequencies. With the chosen target function, the Yule-Walker algorithm also has difficulties working above 70th order, as the matrix it tries to solve becomes rank deficient (infinite number of solutions), thereby giving no result. Other than changing the target function, this Yule-Walker method cannot give better results.

The *invfreqz* method provides more direct control of the design of the final filter than the Yule-Walker method, so it might be able to provide better results. The length of

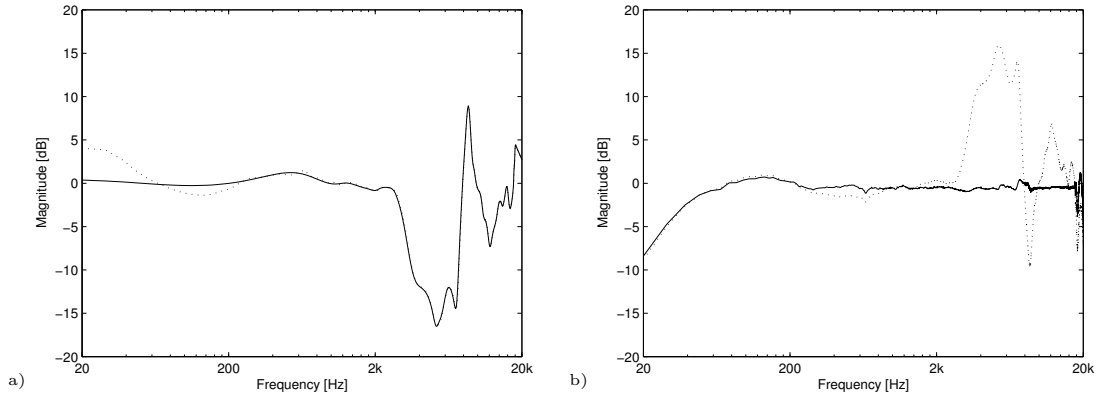



**Figure 3.12:** Example with using the Yule-Walker algorithm to fit a 70th order IIR filter to the target function. The mean-squared error is calculated to $0.055\,\mathrm{dB}$. **a)** Target function (dashed) vs. filter response (solid). **b)** Filtered headphone response (solid) vs. original headphone response (dashed).

the **a** and **b** polynomials can be set individually, and the method allows weighting of the individual frequencies, so it is possible to prioritize certain frequencies. This gives the possibility to tune the filter manually.

By experimentation, improvements could be made when weighting the low frequencies highest, so this weighting is used throughout the tests.

The target function is the same as the one used for the Yule-Walker algorithm. Results for 15th, 30th and 50th order IIR filters created by the *invfreqz* method is shown in Figures 3.13, 3.14 and 3.15 respectively. For the used target function, the method has numerical problems when the order of the **a**-polynomial gets above 50, so a 70th order IIR filter could not be designed with this method. However, the order of the **b**-polynomia can still be increased without numerical problems, so a filter is created with 50 **a**-coefficients and 90 **b**-coefficients. This filter is shown in Figure 3.16.

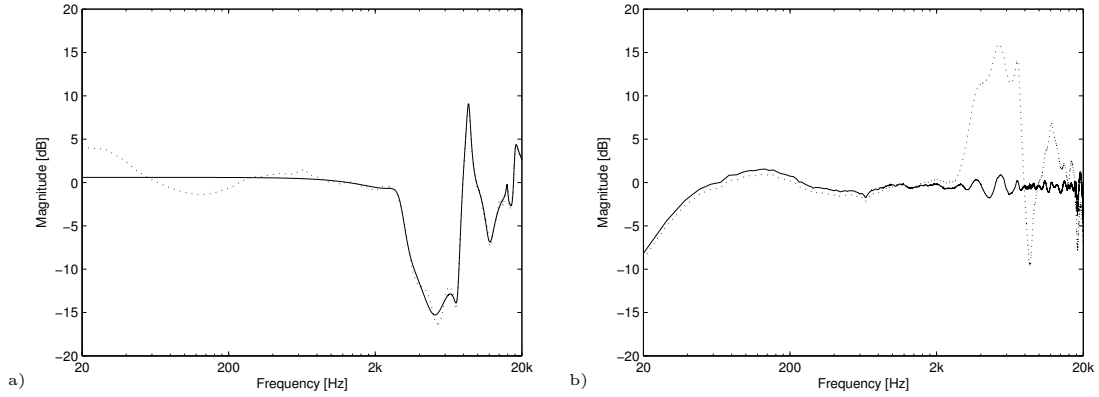



**Figure 3.13:** Example with using the *invfreqz* algorithm to fit a 15th order IIR filter to the target function. The mean-squared error is calculated to $0.184\,\mathrm{dB}$. **a)** Target function (dashed) vs. filter response (solid). **b)** Filtered headphone response (solid) vs. original headphone response (dashed).

The *invfreqz* generally performs better than *yulewalk*, not only through visual inspection, but also on the calculated mean-squared error. *invfreqz* seems to be more accurate in the

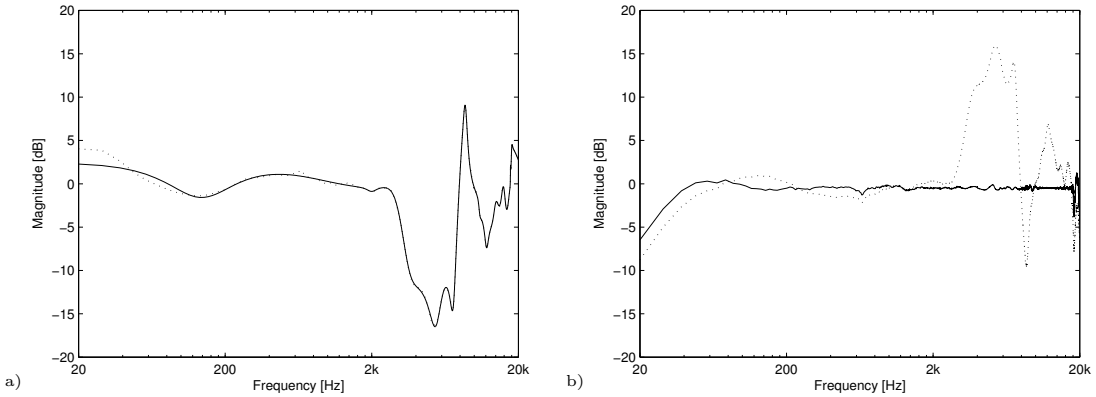



**Figure 3.14:** Example with using the *invfreqz* algorithm to fit a 30th order IIR filter to the target function. The mean-squared error is calculated to $0.016\,\mathrm{dB}$. **a)** Target function (dashed) vs. filter response (solid). **b)** Filtered headphone response (solid) vs. original headphone response (dashed).

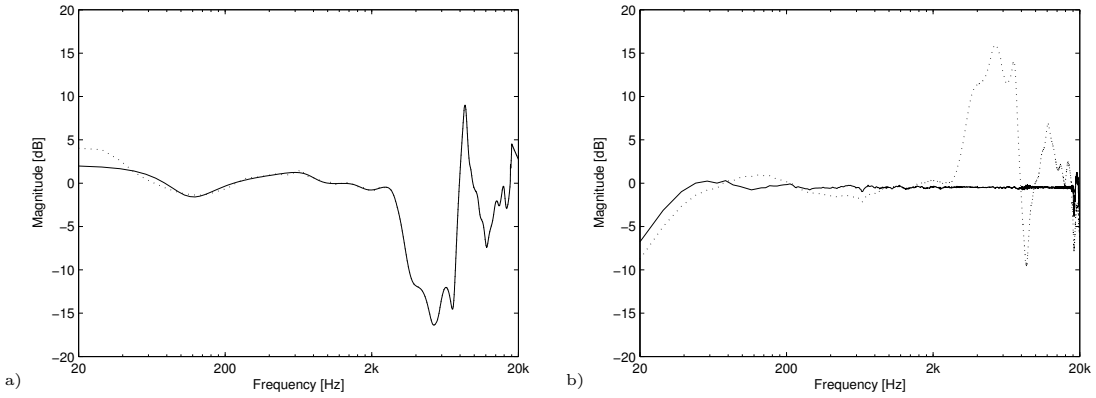



**Figure 3.15:** Example with using the *invfreqz* algorithm to fit a 50th order IIR filter to the target function. The mean-squared error is calculated to $0.009\,\mathrm{dB}$. **a)** Target function (dashed) vs. filter response (solid). **b)** Filtered headphone response (solid) vs. original headphone response (dashed).

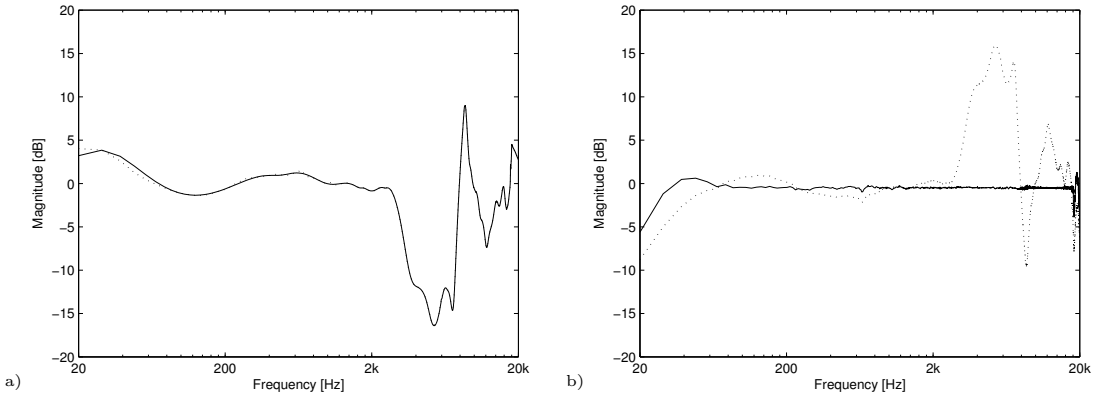



**Figure 3.16:** Example with using the *invfreqz* algorithm to fit a filter to the target function, where the $\mathbf{a}$ and $\mathbf{b}$ polynomias have different length. In this case, the order of $\mathbf{a}$ is 50 and the order of $\mathbf{b}$ is 90. The mean-squared error is calculated to $0.004\,\mathrm{dB}$. **a)** Target function (dashed) vs. filter response (solid). **b)** Filtered headphone response (solid) vs. original headphone response (dashed).

low frequencies, and the 30th order IIR filter created with *invfreqz* performs even better than the 70th order IIR filter created by *yulewalk* with a mean-squared error of 0.016 dB and 0.055 dB respectively. This is probably due to the chosen high weighting of the low frequencies in the implemented *invfreqz* method.

Not surprisingly, a higher order generally gives a lower error, but the *invfreqz* also performs much better than *yulewalk* with the same order, so the *invfreqz* method is chosen for the creation of equalization filters.

Chapter 4

# Measurements

It was found in the analysis that the acoustical filters to be synthesized, can be described by head related transfer functions. In the previous signal chain assessment, it was described how these can be obtained. The first step is to measure the HRTFs and the loudspeaker responses in an anechoic environment. To simulate a standard listening environment, the properties of such must be determined. This is done through BRIR measurements to describe the desired transmission paths between loudspeaker and head in the room. The measurement of the loudspeaker in the anechoic environment makes it possible to subtract the loudspeaker from the measurements in the multi-channel room, leaving only the pure binaural room response. The specifications of the multi-channel room are verified with one-third-octave band reverberation time measurements, to comply with the recommendations described in Section 2.1.3.

To ensure a flat frequency response of the headphones, the individual frequency responses of the left and right channels of two DT990pro headphones are measured for the purpose of designing inverse filters.

The complete measurement report describing the procedures with drawings of the setups and lists of the used equipment is included in Appendix A.

## 4.1 Obtaining the HRTFs

As described earlier in Section 3.3 an HRTF is defined as the pressure at some point in the ear canal related to the pressure at the center of the head when the head is not present. The latter is effectively a loudspeaker measurement with a measurement microphone at the center position of the head.

It has been decided to use a single active Genelec 1031A loudspeaker for all the measurements under anechoic conditions. To use the same loudspeaker for all surround positions has the advantage of reducing any variations that might otherwise exist between the used loudspeakers. But it is still necessary to compensate for the characteristic of the used loudspeaker. The loudspeaker measurements are done under anechoic conditions with 3 m between loudspeaker and microphone. The distance between microphone and

loudspeaker is not crucial as long as the measurement is done in the far-field of the loudspeaker. The frequency dependent absorption coefficient of the air can be neglected for such short distances. More details about the measurement are included in Appendix A.2.1.

The magnitude response of the used loudspeaker can be seen in Figure 4.1. For lower frequencies, the loudspeaker response drops at around 50 Hz which corresponds to the specifications of the loudspeaker.

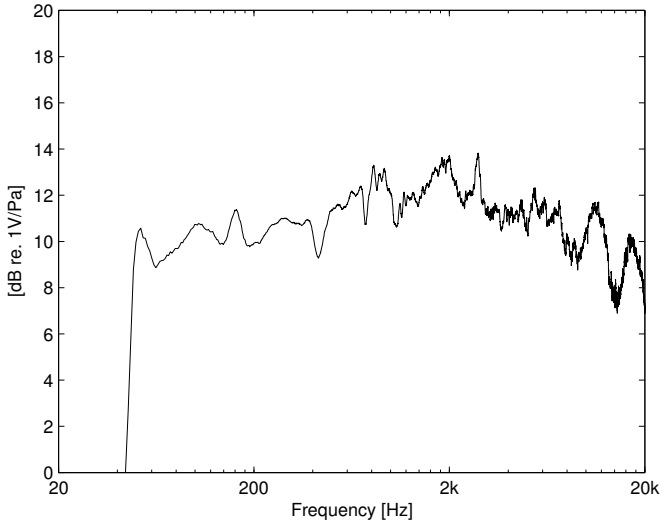



**Figure 4.1:** Magnitude response of the measured Genelec 1031A loudspeaker.

The second part of the HRTF measurement is the measurement of the sound pressure in the ear canal. It is explained earlier in Section 3.3 that the sound pressure at the blocked entrance of the ear canal is adequate. The artificial head VALDEMAR with blocked ear canals was used for the measurements, thus the transfer function of the ear canal is not included in the plotted HRTFs in Figure 4.2, which shows the relation between the sound pressure at the blocked ear canal and the sound pressure in the center of the head. All plots are corrected for the response of the loudspeaker. As expected, the overall amplitude of the HRTFs from loudspeakers at the same side as the measured ear are higher than the amplitude of the HRTFs from the loudspeaker on the opposite side of the head. This counts especially for middle and high frequencies. At low frequencies, the human head is small compared to the wavelength and thus has only a small effect on the sound field. In all plots, characteristical peaks and dips can be seen that contain the spatial information added by reflections on ear and torso. The HRTFs from the front center and rear center loudspeakers to the left and right ear are plotted in the bottom of Figure 4.2. The HRTF measurements for left and right ear are expected to be identical for both loudspeaker positions in an ideal symmetrical setup and they are close to that. Some variations occur in the high frequencies which can result from small deviations from the ideal setup.

In total, three different microphones have been used, one for the loudspeaker measurement and one for each ear in the artificial head. They are all assumed to be calibrated to a flat frequency response within the audible range, so no compensation is necessary. More details about the measurement of the HRTFs are included in Appendix A.2.2.

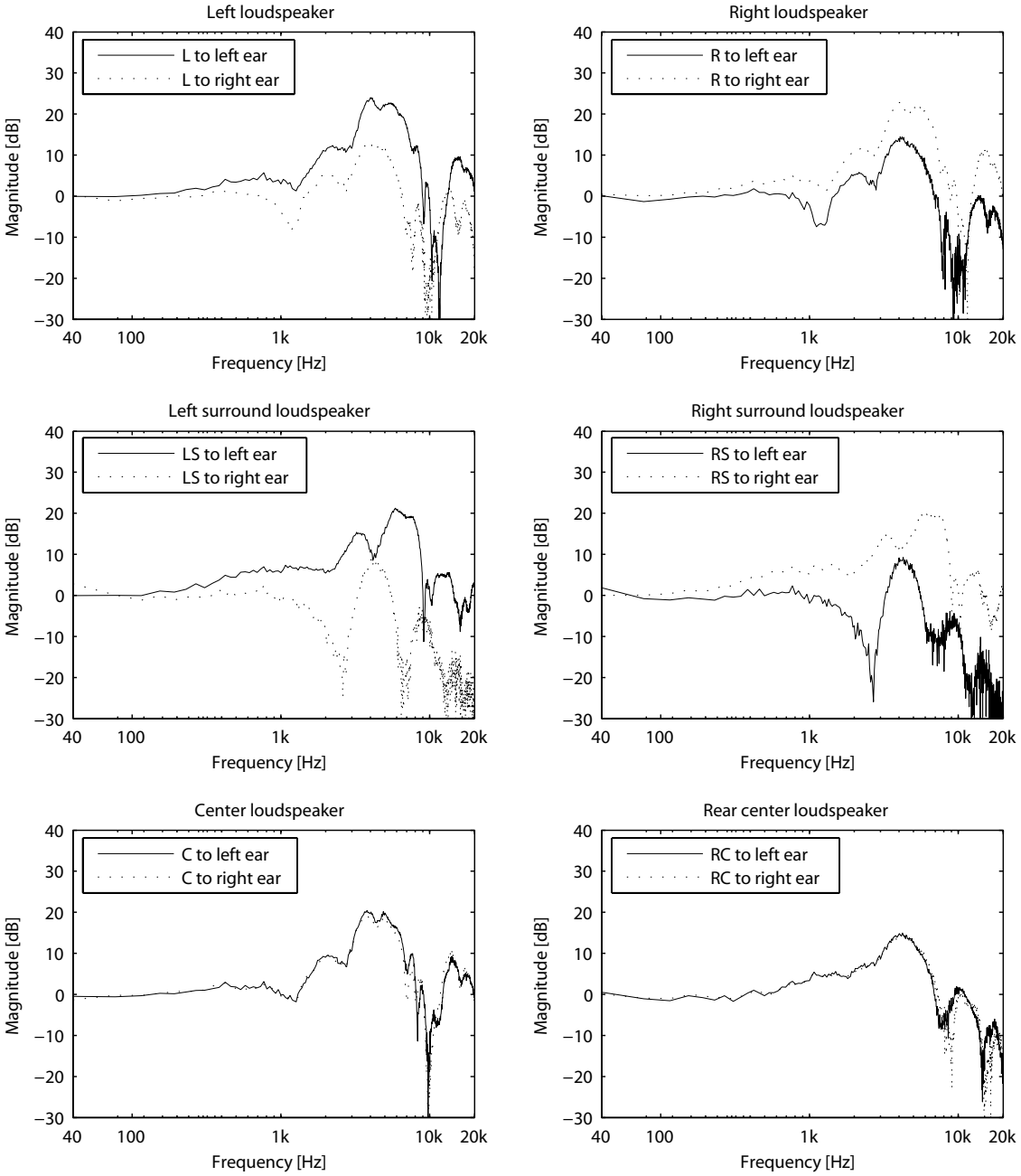



**Figure 4.2:** Measured HRTFs from different loudspeaker positions to the ears. All measurements have been corrected for the response of the used loudspeaker.

The measured HRTFs contain a delay corresponding to the distance from the source to the ear. Because of this, the first 355 samples have been removed in the plots in Figure 4.3, which show the first ten milliseconds of the remaining HRTFs to one ear, including the characteristic of the loudspeaker. Small reflections from the grid are expected, but cannot be seen in these pictures after the direct sound.

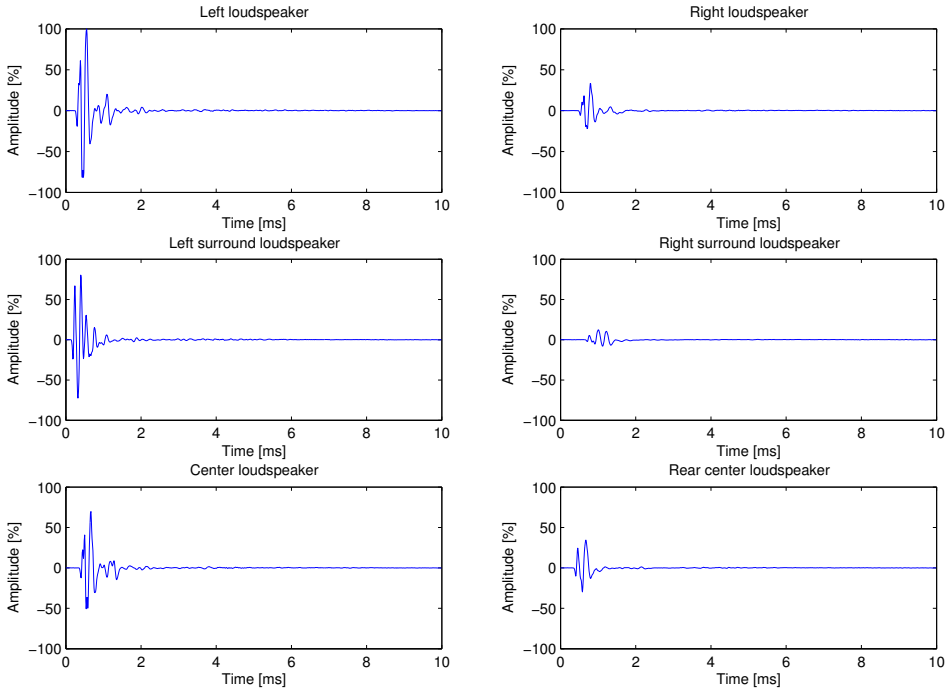



**Figure 4.3:** Ten milliseconds of the truncated HRTFs from different loudspeaker positions to the left ear, including the characteristic of the loudspeaker.

As specified in the system specifications (Section 2.3), a measurement with a subwoofer was also made. More details on this can be found in Appendix A.2.2.

## 4.2 Reverberation time of the multi-channel room

A suitable room, that satisfies the specifications presented in Section 2.1, had to be selected in which the BRIRs could be measured. The multi-channel listening room in the laboratory of the acoustics department at Aalborg University was chosen. This room is designed according to the ITU recommendation for multi-channel reproduction [BS775-1, 1992], and was designed with the intention to be completely symmetrical. To control the specifications in terms of reverberation time, a measurement was carried out as described in Appendix A.3. The results are shown in Figure 4.4. According to the AES specifications for multi-channel surround sound systems [Rumsey et al., 2001] the reverberation time in a reference listening room should comply with certain limits. One is the arithmetic average $T_m$ which is calculated from the measurements in one-third-octave bands between 200 Hz and 4 kHz. It should be between 0.2 s and 0.4 s depending on the

size of the room and an approximate value can be calculated with:

$$T_m \approx 0.25 \sqrt[3]{\frac{V}{V_0}}$$

where $V = 172.66\,\mathrm{m}^3$ is the listening room volume and $V_0 = 100\,\mathrm{m}^3$ equals the reference room volume. The measured value for $T_m$ equals approximately 0.16 s which is much lower than the calculated 0.3 s. This means that the room is too dry, according to the recommendations. This can be changed by adding some reflectors to the room, but to keep the impulse responses as short as possible, it was decided not to do that.

The variations in reverberation time between 200 Hz and 8 kHz should lie within $\pm 0.05$ s relative to the nominal value of $T_m$. Below 200 Hz, the difference between neighbouring frequency bands should not exceed 25% of the longest reverberation time. This tolerance mask is drawn together with the measured values in Figure 4.4. The plot shows that the variations in the reverberation time just lies within the acceptable tolerance. Thus relating to this parameter, the room is suitable for multi-channel reproduction.

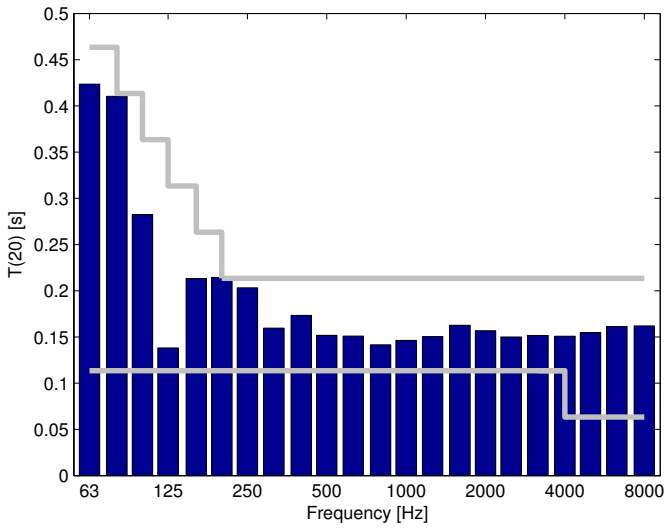



**Figure 4.4:** The reverberation time in the multi-channel room measured in one-third-octave bands. Drawn with solid lines on top is the tolerance mask according to [Rumsey et al., 2001] with $T_m = 0.16\,\mathrm{s}$.

## 4.3 Obtaining the BRIRs

A traditional room impulse response includes all information about the sound transmission in a room from one point to another, that is travel distance, reflections and absorption. It does not include any transfer function of the source and receiver that may be used when measuring the room impulse response. The binaural room impulse response (BRIR) is similar, but adds the influence of the human head and torso to simulate the listening experience of a human in a real room. As defined in Section 3.3 a BRIR relates the pressure in the ear canal to the pressure in the middle of the head without the head and the room being present. The BRIRs can be considered as a mixture of the traditional room impulse response and the HRTFs for a head.

To simulate a human listening to a 5.1 channel surround setup in a real room, it is necessary to know the BRIRs from all loudspeakers to both ears, which is 10 measurements excluding the subwoofer.

For the measurement of the BRIRs, it was decided to use a setup with different loudspeakers for each position because an existing setup was used, which could not be moved. Thus it is necessary to compensate for each loudspeaker seperately. Measurements of the anechoic impulse responses for all used loudspeakers were available and have been used. The loudspeakers in the used surround setup in the multi-channel room had a distance of 2.5 m to the listening position. This distance was used in order to move the loudspeakers far enough away from the wall. According to [BS775-1, 1992] the distance from each source to a reflecting surface should be at least 1 m, which was fulfilled in the used setup.

The level produced by the different loudspeakers at the listening position was calibrated to an equal level. The two microphone channels have been aligned to the same sensitivity in order to maximize the signal-to-noise level for both channels. This was done by playing back white noise from the center loudspeaker. After that the sensitivity of each recording channel was measured using an acoustical calibrator. The calculated sensitivities of the two channels were used to correct the measured BRIRs to eliminate any level differences in the recording chain. A detailed description of the calibration procedure and the measurement of the BRIRs can be seen in Appendix A.3.2.

The same artificial head as before was used with blocked ear canals. It is explained earlier in Section 3.3 that it is possible to do the measurements with a blocked ear canal without losing spatial information. The plots of the BRIRs, which can be seen in Figure 4.5, show the relation between the sound pressure at the blocked ear canal and the sound pressure the loudspeaker would have produced in the middle of the head, without head and room being present. The anechoic loudspeaker measurements were used as the pressure in the middle of the head.

The measured BRIRs for all surround positions can be seen in Figure 4.5. All plots are corrected for the response of the individual loudspeakers. The overall magnitude for these BRIRs is expected to be several dBs higher than the one of the HRTFs. This is because the SPL in a room is expected to be higher than the one in an anechoic environment with the input to the source and the source remaining the same. But the magnitude in the shown plots is much higher than expected. It is concluded that there must have been an unknown factor in the measurement chain, which scaled the signal. As the gain of all other parts of the measurement chain have been measured and compensated for, this gain must appear somewhere in WinMLS, which is the measurement system used for obtaining the BRIRs. Still all needed information is contained in the variations in and between the BRIRs and they will be normalized for further calculation. Thus it is not crucial to find the exact error.

The plots in Figure 4.6 show the first ten milliseconds of the remaining BRIRs to one ear, after their beginning has been removed like explained earlier for the HRTFs. They include as well the characteristic of the loudspeaker. It can be seen that the direct sound is followed by early reflections. The first reflections come from the floor, followed by the ones from ceiling. In some plots reflections from the nearest walls can be seen in the end of the plot. As the room is relatively dry, the amplitude of the reflections diminish quickly. A calculated reflectogram for the left loudspeaker can be seen in Section 2.2.3 on page 16, where room acoustics are considered. Some more detailed evaluation of the BRIRs will

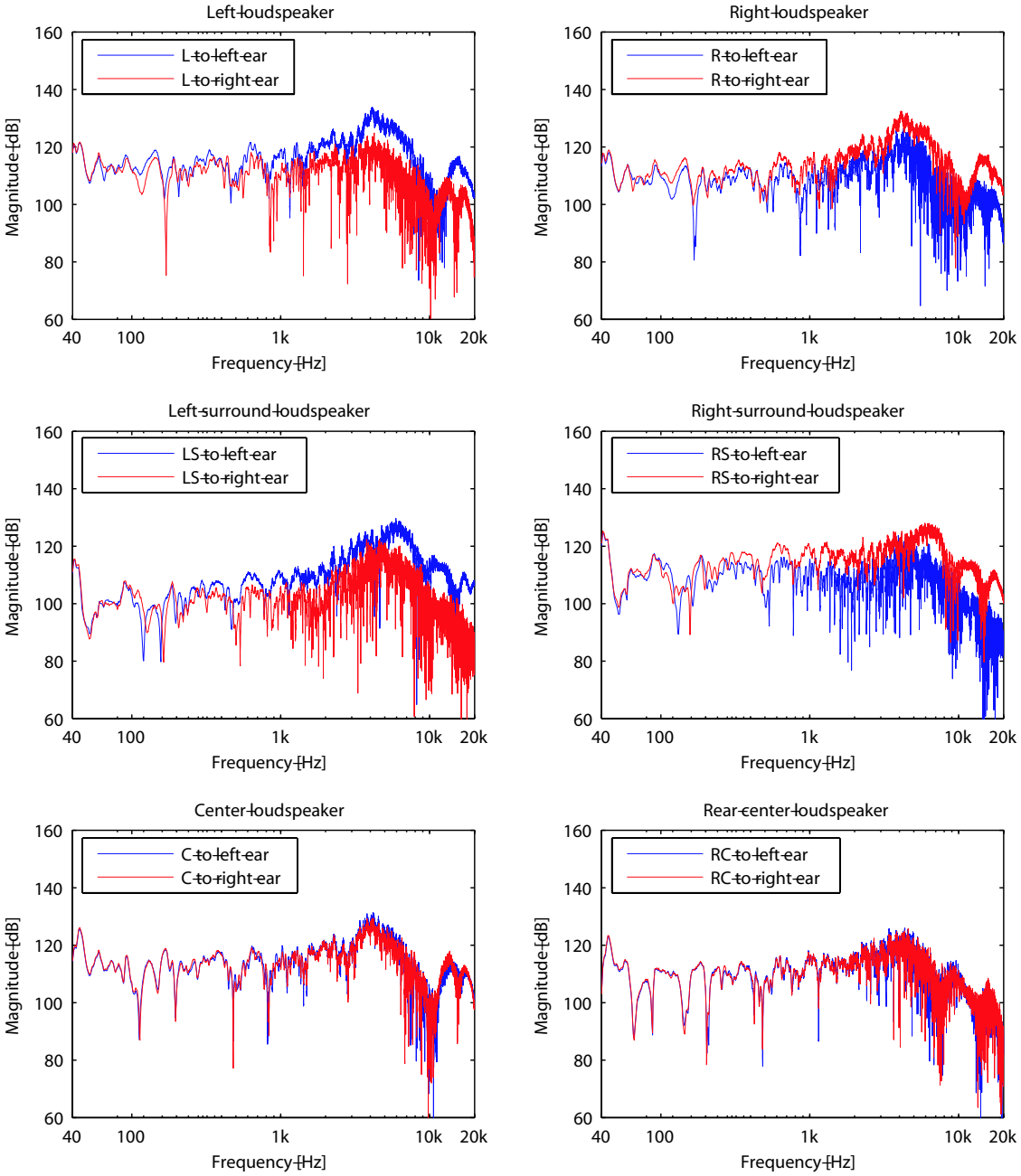



**Figure 4.5:** Measured BRIRs from different loudspeaker positions to the ears. All measurements have been corrected for the response of the used loudspeaker.

be done in Section 5.1.

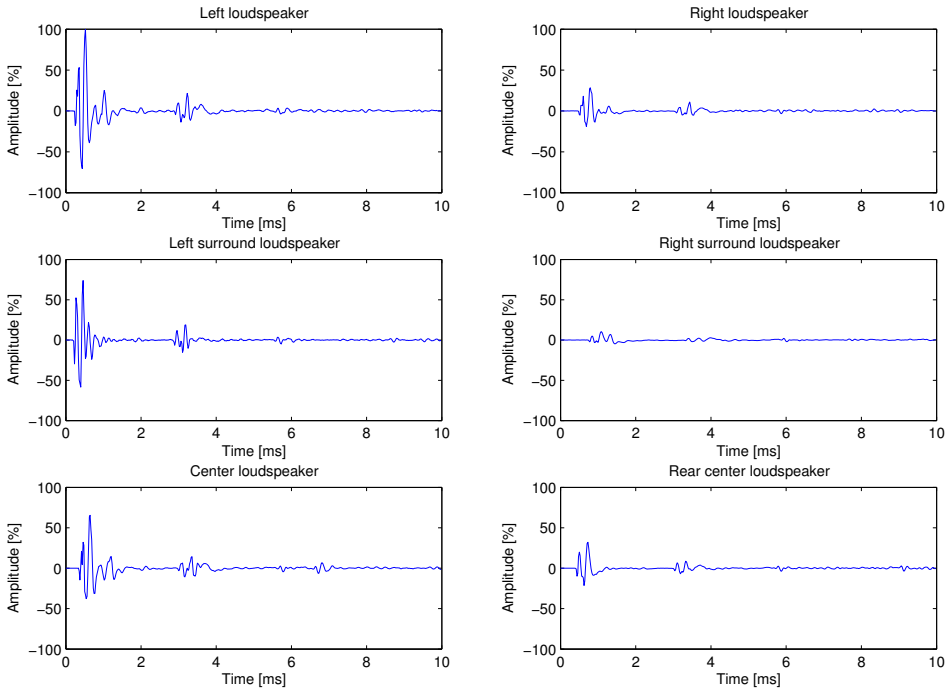



**Figure 4.6:** Ten milliseconds of the truncated BRIRs from different loudspeaker positions to the left ear, including the characteristics of the loudspeakers.

## 4.4 Binaural recordings

According to the system specifications (Section 2.3), a reference for the planned listening test was needed. Thus it was decided to make a binaural recording in the multi-channel room setup. Different 5.1 channel soundtracks from movies have been played back and recorded using VALDEMAR. The microphone amplifier was adjusted in order to use most of the dynamic range while recording. A list of the used movie tracks can be seen together with more details about the measurement procedure in Appendix A.3.3.

## 4.5 Headphone transfer functions

It was decided in Section 2.3 to compare different equalization approaches for headphones, thus measurements of the headphones using VALDEMAR were needed. As described in Section 3.1 the headphone measurement should be done with blocked ear canals, when the recording was made with blocked ear canals. Two pairs of headphones (DT990pro) have been measured on VALDEMAR. An MLS based measurement system was used to determine their impulse responses. The frequency response of headphones varies with their position on the head, so an average over six different positions was made. The measurement report is attached in Appendix A.2.3.

The results are shown in Figure 4.7. It can be seen easily that there are differences between the two plots. This means that there are considerable variations between headphones even of the same model. A more detailed evaluation of these results follows in Section 5.2.

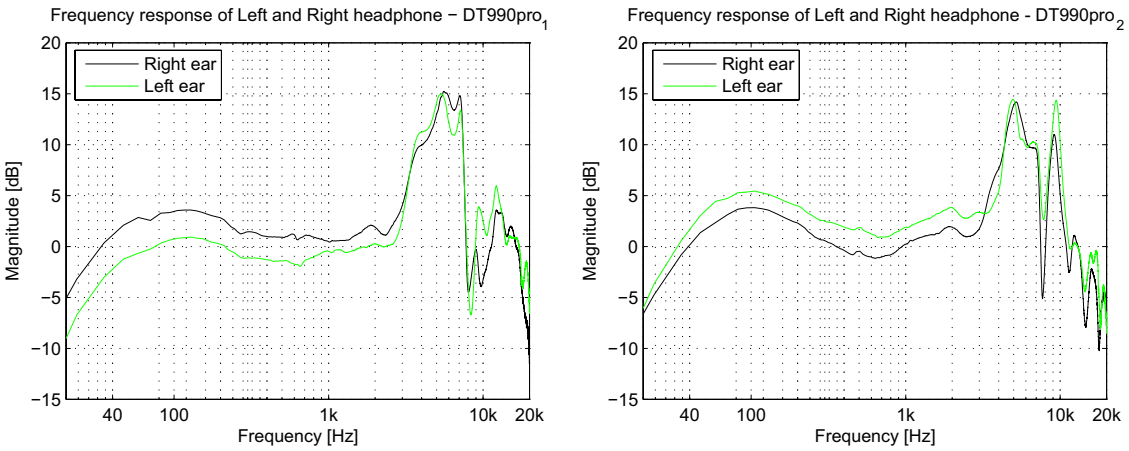



**Figure 4.7:** Transfer functions of the two measured pairs of headphones on VALDEMAR.

Chapter 5

# Implementation

By using the measurement results from previous chapter, the binaural synthesis algorithm will now be implemented. The goal is to create the offline system, in order to make the different binaural synthesis samples needed in the listening experiment. However, before the different elements required in the processing are combined, the measurement results must be post-processed. This is divided into two parts, first the BRIRs are analysed and processed in more details, followed by the headphone measurements from which the equalization filters are created.

## 5.1 Post-processing of BRIRs

The most important results from the measurements described in the prior chapter, are the BRIRs from the multi-channel room, as these are the fundamental signals containing all the necessary spatial information and the ones that are to be used in the listening experiment. If any errors have occurred in the measurement chain and thus distorted the BRIRs in any way, it will not be possible to perform a binaural synthesis that rivals the binaural recording. This section assesses the validity of the BRIRs by analyzing them first in frequency and then in time, according to what can be expected from responses like these. After this validation, the level throughout the signal processing chain will be assessed in order to find the proper gain adjustments.

### 5.1.1 Frequency domain assessment

The goal of analyzing the BRIRs in frequency is to verify that the magnitude responses have a correct overall shape individually and relative to each other. As described in the Appendix A.3.2, the output level of the loudspeakers were identical with a very small error margin. This means that the only gain adjustments necessary before comparing the BRIRs are corrections for different sensitivities of the two microphones used in the artificial head. If the setup used is completely symmetrical, then the BRIR from left front ($L$) loudspeaker to left ear should be identical to the BRIR from right front ($R$) loudspeaker to right ear. Thus, the symmetry line follows the median plane so that the two BRIRs from the center loudspeaker ($C$) should be the same. However, for this to be

the case, complete symmetry around the median plane must include the loudspeakers, the room, and the artificial head comprising head, torso, ears, and microphones. A simple evaluation of the loudspeakers and the artificial head are presented in Figure 5.1.

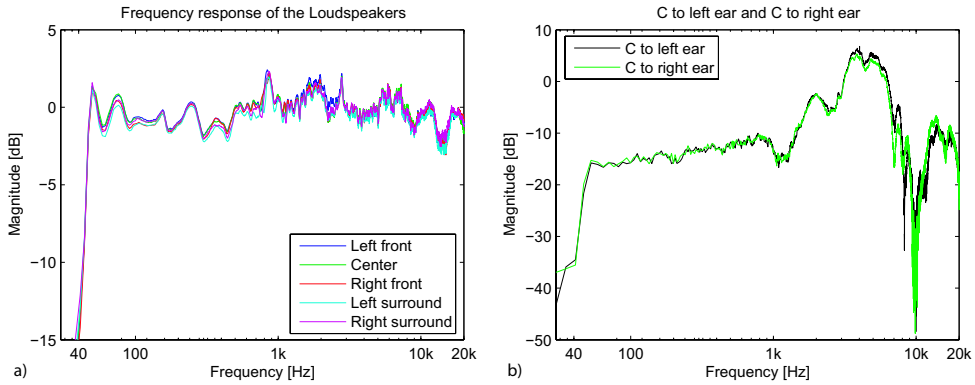


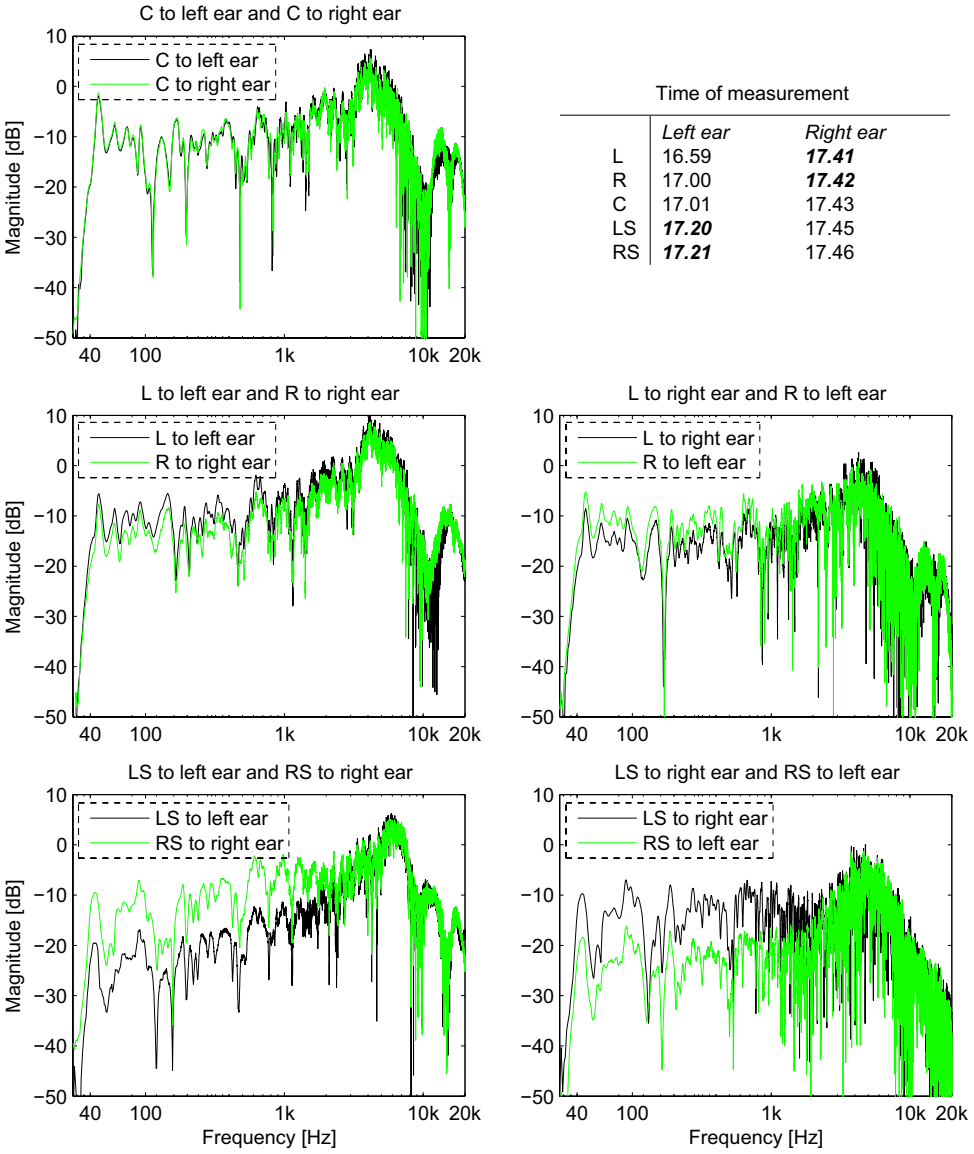

**Figure 5.1:** Symmetry evaluation of equipment used in the multi channel room: a) anechoic frequency response of the five Genelec 1031A loudspeakers used in the surround setup, and b) anechoic frequency response of center loudspeaker to left and right ear on the artificial head.

The frequency responses of the five Genelec loudspeakers measured in the anechoic chamber are shown in Figure 5.1a for which the maximum difference between loudspeakers is approximately 1 dB. It can be seen that the overall curve characteristics are the same for all loudspeakers, thus symmetry can be assumed with regard to these. Figure 5.1b shows the frequency response from 0° azimuth to left and right ear measured under anechoic conditions. If the artificial head is completely symmetrical, these two plots will be identical, which is more or less the case although variations occur in the higher frequencies, but this area is also more sensitive to small amount of asymmetry.

The multi-channel room in which the BRIRs were obtained is built as a symmetrical room, complete with an extra door to match the real door. However, small variations will always be present and the setup with regard to the median plane might not be 100 % centered in the room.

It is assumed that the multi-channel room setup is symmetrical to a certain degree so that the BRIRs can be evaluated on the basis of this. Variations are however expected, especially in the high frequencies. Figure 5.2 shows the ten BRIRs plotted in pairs which should be identical. The graphs are scaled to compensate for sensitivity differences between left and right ear microphones. The responses are not corrected for loudspeaker characteristics and they are normalized according to the gain adjustments presented later in section 5.1.3.

It is immediately apparent that something is wrong with the measurements from the two surround loudspeakers ($LS$ and $RS$), as a deviation of up to 10 dB occurs below 2 kHz while the curves above this frequency follow each other as would be expected. The same applies to front left and right loudspeakers although the deviations are smaller in this case. The two BRIRs for the center loudspeaker are identical to a degree that is expected, which means that the artificial head, with everything included, behaves as in Figure 5.1b, and that the room is more or less symmetrical. This means that the errors seen in the four other plots are not related to this part of the measurement chain. By comparing the two last plots for $LS$ and $RS$ it can be seen that the error is related to the recording




Time of measurement

| | *Left ear* | *Right ear* |
|---|---|---|
| L | 16.59 | ***17.41*** |
| R | 17.00 | ***17.42*** |
| C | 17.01 | 17.43 |
| LS | ***17.20*** | 17.45 |
| RS | ***17.21*** | 17.46 |

**Figure 5.2:** BRIRs from multi-channel room in pairs that under symmetrical conditions should be identical. $L$, $C$, $R$, $LS$, and $RS$ are left front, center, right front, left surround, and right surround loudspeakers, respectively. In the upper right corner there is a table listing at what time the individual measurements were made.

chain and not the loudspeakers, which is also the case for the $L$ and $R$ measurements. To determine which measurements are the correct ones, they are compared with the same kind of measurements performed by group 961 in 2003, which are shown in Figure 5.3.

The setup used for these measurements is roughly the same as the one used in this project although the measuring equipment differs. The measurements are scaled for microphone sensitivity and normalized in the same way as those in Figure 5.2. By comparing the two sets of data it can be concluded that for the $L$ and $R$ measurement the error is associated with the right ear and for the $LS$ and $RS$ loudspeakers it is the left ear. The time of measurement for these four error-prone measurements is marked in the table included in Figure 5.2, which shows that these are made one after the other. The process of

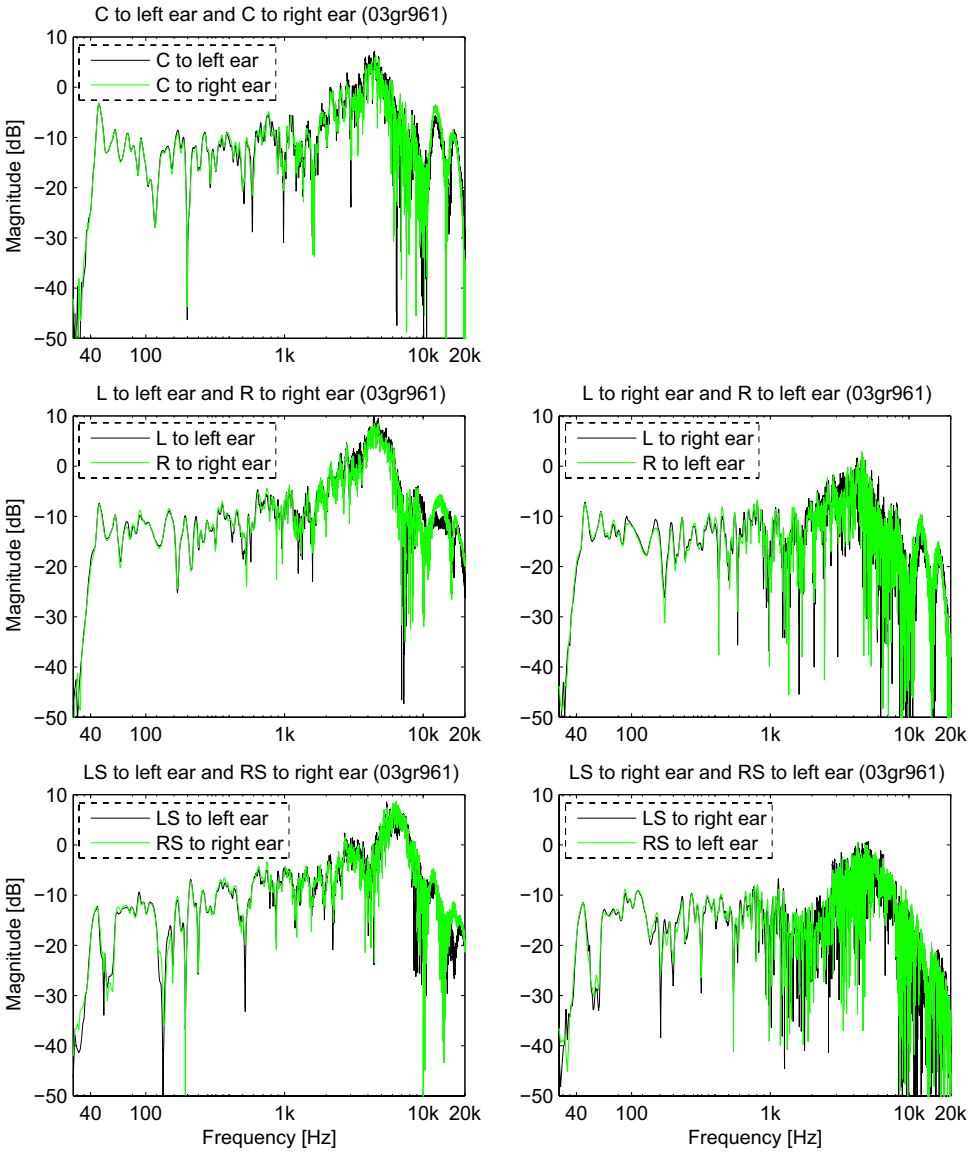



**Figure 5.3:** BRIRs from multi-channel room measured by group 961 in 2003 under conditions similar to the setup used in this project. $L$, $C$, $R$, $LS$, and $RS$ are left front, center, right front, left surround, and right surround loudspeakers, respectively.

switching between BRIR measurements was limited to moving cables in the control room as described in the previous section. The errors might be associated with one of these cables in which a loose connection in some cases could distort the signals.

As these errors were not identified within the time period when the multi-channel room was vacant, it is not possible to redo the measurements. A solution would be to use the data from previous year, but preliminary testing has revealed that these add a mid frequency colourization that is not present in the binaural recording or the BRIRs from this project, which will add an unwanted bias in a later listening experiment. However, it can be seen in Figure 5.2 that maximum one error-prone measurement is present in each plot, which means that a mirrored setup can be created. For example, instead of using the faulty $LS$ to left ear measurement the complementary $RS$ to right ear is used. Only

the two measurements from the center loudspeaker are excluded from such modifications.

### 5.1.2 Time domain assessment

The purpose of analyzing the BRIRs in time domain is to assess if they follow the expected decay curve and when the impulse response has died out. This is important information from a computational point of view, as it is desirable to limit the length of the BRIRs in order to minimize the computational cost. Figure 5.4 shows two plots of the response from left loudspeaker to left ear with different lengths.

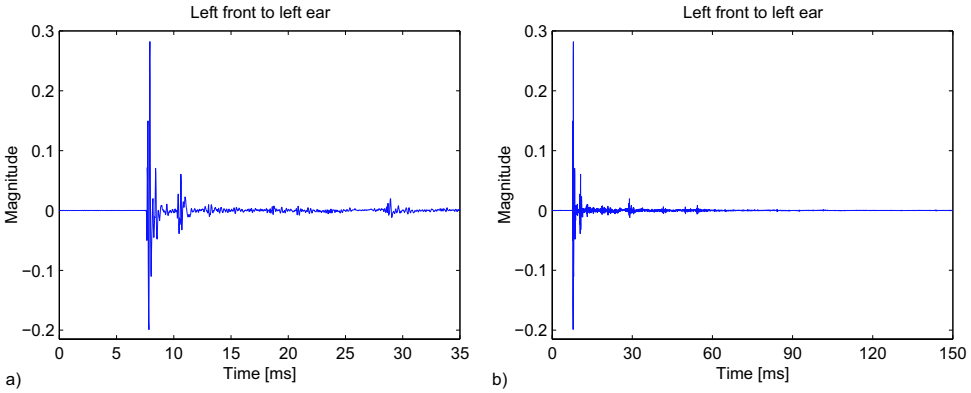



**Figure 5.4:** Response from front left loudspeaker to left ear of VALDEMAR plotted with two different time durations.

Figure 5.4a is limited to just above half the length of the simulation in Figure 2.8 which is based on roughly the same setup. This means that the above figure shows the direct sound together with all first order reflections and most of the second order as well. The earliest reflections from the floor and ceiling just above 10 ms is quite prominent, but then the response quickly decays and individual reflections are hard to spot. This however, is also what is expected from the dry environment in the multi-channel room. Figure 5.4b shows the same response up to 150 ms at which time mainly low frequency variation is detectable.

The reverberation tail can be seen in more detail in Figure 5.5a, which shows the response from 100 ms and up.

Quite prominent peaks relative to the noise floor can be seen throughout the response and a close up of one of these are shown in Figure 5.5b. These spurious reflections are assumed to come from some kind of distortion in the measuring chain. As it was made sure that overload did not occur during the measurements, these fake reflections cannot be assigned to this. Another cause could be that the excitation of the loudspeaker was too high. This can lead to distortion from too high cone excursion and variations in coil temperature. The spurious reflections are present in all the impulse responses from the multi-channel room, and it can be heard as clicks when listening directly to the response. The highest peak in Figure 5.5a around 0.9 s is only 30 dB lower than the direct sound. As a comparison, the noise floor is approximately 60 dB lower than the direct sound. As with the magnitude errors, this was not noticed in the time window when the multi-channel room was vacant, thus redoing the measurements has not been possible. If binaural synthesis is performed using the entire response, something like echoes can be heard, but

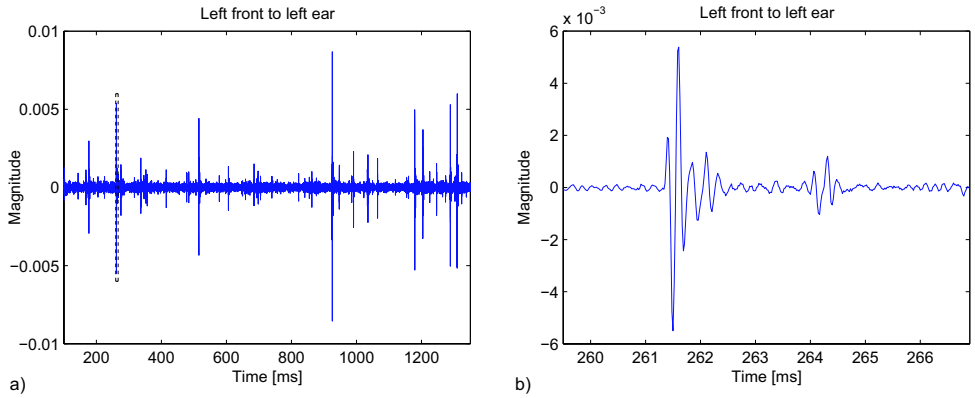



**Figure 5.5:** Reverberation tail of response from front left loudspeaker to left ear of VALDEMAR plotted with two different time durations.

this can be limited by making the BRIRs shorter. Preliminary testing with movie samples does not indicate noticeable artefacts when the BRIR lengths are shortened to 0.5 s, which should be sufficient to describe the room according to the measured reverberation time.

### Choice of BRIR length

The number of different BRIR lengths to be used in the listening experiment is limited to three, which are denoted, *long*, *mid*, and *short*:

**Long:** Here the BRIR length is set to 0.5 s in coherence with the longest average reverberation time measured in third octave bands in the multi-channel room (refer to Section 4.2).

**Mid:** This BRIR length is used to assess the consequences of shortening the BRIRs down to a length optimized for real time implementation. The filtering process is done through FFT convolution based on the *overlap-add* method described in Appendix B. The AC-3 filter, in which the binaural synthesis is to be implemented, is able to work with a delay of up to 0.5 s, as it can get the audio before the video is processed. However, the FFTs must have length according to $2^n$, with $n$ being a positive integer. The closest value to fulfill the maximum delay criteria is $2^{14}$ which corresponds to approximately 0.34 s when sampling at 48 kHz. When dividing the input into blocks the total delay can appropriately be set to two blocks, so that a block can be sampled and then processed in the time space when the next block is loaded. Based on these design criteria, two solutions are presented in Figure 5.6 that are either optimized for maximum BRIR length or minimum computational cost.

Both designs utilizes a FFT length of $2^{14}$ which corresponds to the zero-padded signal lengths $L + M - 1$. In Figure 5.6a the BRIR length is set to approx 0.17 s which is the maximum possible if $M$ is not to exceed $L$. Of course it would be possible to make the BRIRs even longer, but then the input blocks must be shorter which will increase the computational cost even more. Figure 5.6b shows the case when $L$ is maximized to 0.25 s to fully utilize the 0.5 s delay possibility and thus shortening the BRIRs to 0.09 s. As both methods uses the same FFT length, the second method will be approximately 50 % more efficient than the first. Preliminary

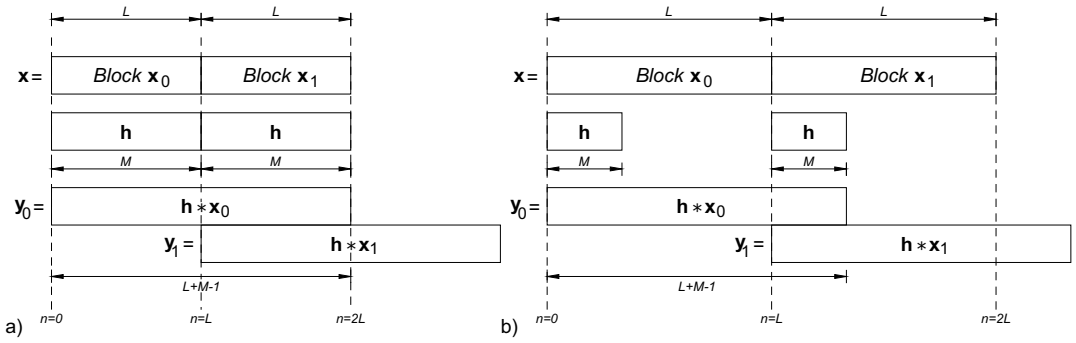



**Figure 5.6:** Design of overlap-add method when two blocks ($2L$) maximum can be $0.5\,\mathrm{s}$ and FFT length must match $2^{14}$ : **a)** maximum BRIR length of approx $0.17\,\mathrm{s}$ when $L \geq M$ and **b)** minimized computational cost by maximizing $L$

simulations show that differences in binaural synthesis when using either $0.17\,\mathrm{s}$ or $0.09\,\mathrm{s}$ BRIRs are very hard to distinguish, so the *mid* length of BRIRs are chosen to $0.09\,\mathrm{s}$.

**Short:** In order to investigate the influence of the room on the listening experience, the *short* BRIR is limited to $0.01\,\mathrm{s}$ and thus, only includes the direct sound. This should then correspond to the measurements from the anechoic room, which is illustrated for one loudspeaker in Figure 5.7.

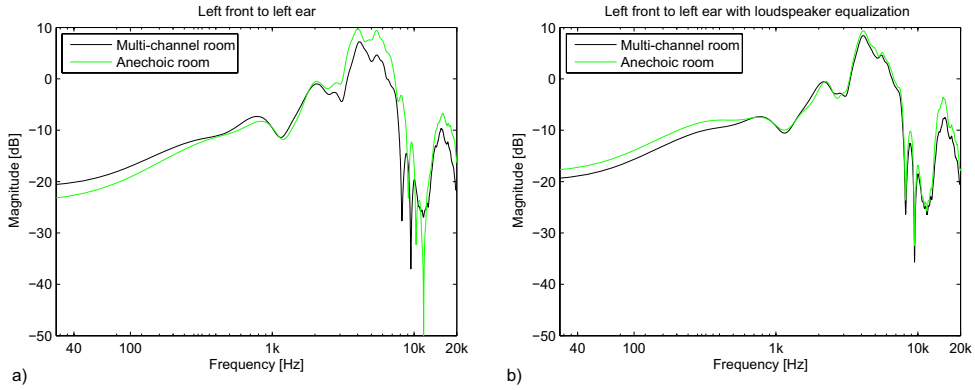



**Figure 5.7:** Response from left front to left ear with a signal duration of $0.01\,\mathrm{s}$ for measurements done in both multi-channel and anechoic room: **a)** and **b)** show with and without loudspeaker equalization, respectively.

When compensation is made for the different loudspeaker characteristics, the two responses become more or less identical.

### 5.1.3 Gain adjustments throughout the system

An analysis of the signal path in relation to the level is needed in order to make sure that the output level is within a valid range so that clipping is avoided. Potential changes in level can occur at several different stages throughout the signal chain, which elements are illustrated in Figure 5.8.

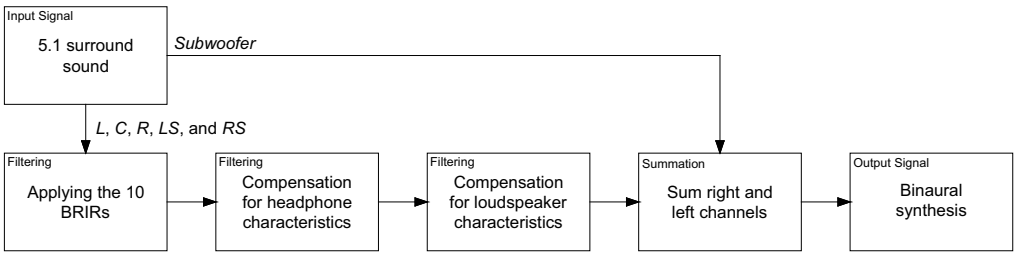



**Figure 5.8:** Simplified signal path for processing 5.1 surround sound into binaural sound.

The system is divided into convolution with BRIRs, headphone equalization, loudspeaker equalization, and summation of the six signals into left and right ear. The signal range of input and output is equal, so in principle the total system gain should be 0 dB. However, as each channel in the output is a sum of the filtered five channel input (plus subwoofer) some degree of attenuation is needed. It is of course possible to completely avoid overload by normalizing all filters to have a maximum gain of 0 dB and scaling each channel by 1/6. This will then resolve the case with a maximum output on all six channels at the same time with a frequency content matching the highest gain in the filters. This scenario is however highly unlikely to occur, plus the fact that some of the filters might boost and attenuate the same frequency components. Another method could be to analyse the entire input signal in order to locate a worst case situation for that given signal, and then adjusting the output gain according to this. This would result in optimum use of the dynamic range, but is computationally demanding and not possible for a real time implementation. Instead, the following steps are used to adjust the system gain:

1. The transfer functions of the headphones and loudspeakers are normalized so that they are roughly situated around 0 dB. This is more or less straightforward with the loudspeakers as the responses of these are quite flat, but the headphone responses are problematic because of the large peaks and dips that occur in these. The process of this equalization is described in more detail in Section 3.4.2.

2. When comparing the binaural synthesis with a binaural recording made in the multi-channel room it is necessary to have the same signal level in order to give a "fair" comparison. Further, the recording level was adjusted so that the dynamic range of the recording device was utilized best possible as described in the previous chapter. This means that the binaural recordings are suitable to be used as design targets when level adjusting the binaural synthesis. The adjustments are done on the BRIRs by scaling these and then comparing the RMS value between binaural synthesis and recording. Because the five loudspeakers are assumed to have the same sensitivity, individual scaling between these is not needed, and thus the same scaling factor must always be used on all BRIRs. In the end the best result was found by finding the maximum amplitude in the ten BRIRs and normalizing this to 10 dB, which corresponds to the plot shown in Figure 5.2.

   The approach of scaling the binaural synthesis according to a corresponding binaural recording should reduce clipping to a minimum as the two methods in principle should yield the same result. However, small variations might occur, especially because of the modifications made previously on the BRIRs in order to compensate for the errors in the measurements. The binaural synthesis program will be made

so that it can handle such situations, but this is assessed in the implementation phase as it does not encompass modifications of the measurement results.

3. According to preliminary testing, it is possible to add the LFE channel directly to left and right channel without scaling, although this is very subjective and some people might want more or less low frequency effect. Thus, it would be a good idea to make the LFE channel adjustable according to a given users preferences.

## 5.2 Equalization of signal chain

As described in Chapter 3, the signal chain comprising recording and reproduction must be equalized for undesired characteristics in the headphone and loudspeaker transfer functions. However, before the inverse filters are created according to the assessments made in Section 3.4, the appropriate target functions must be selected.

### 5.2.1 Choosing target functions

In the system specifications in Section 2.3, three different headphone equalization approaches are used. The general diffuse-field curve was assessed previously in Section 3.2 and is considered ready for the actual inverse filter design presented in the next section. However, the data for the two other, human and VALDEMAR equalization approaches will be analyzed in this section. The goal is to select the appropriate measurements to be used and evaluate how the two equalization approaches are realized.

Figure 5.9 shows the frequency response for two different DT990pro measured on VALDEMAR and human subjects, respectively. Those performed on the artificial head are the ones described in the previous measurement chapter, while the others are an average of measurements done on 27 human subjects by Emine Çelik.

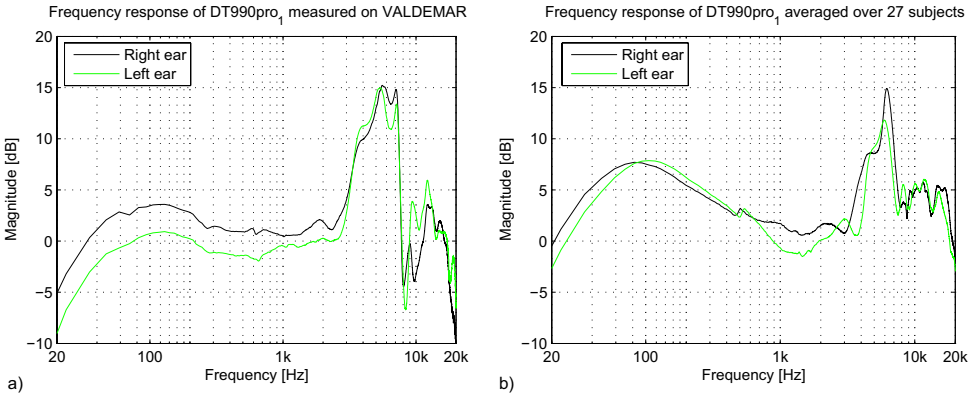



**Figure 5.9:** Frequency responses of DT990pro headphones measured on **a)** VALDEMAR and **b)** 27 human subjects. Note that the headphones used in the two measurements are not the same.

Prominent differences between the two measurements are apparent both in the low and high frequencies. This indicates that VALDEMAR is not a suitable representative of the average human when used for headphone equalization. It was also noticed during the measurements that the headphones did not fit very well on the artificial head, which might

be related to this. However, it should be taken into account that the headphones used in the two measurements are not the same, which might explain some of the differences. Figure 5.10 shows two sets of measurements, both performed on 27 human subjects by Emine Çelik, but with different headphones. Also in this case prominent differences exist.

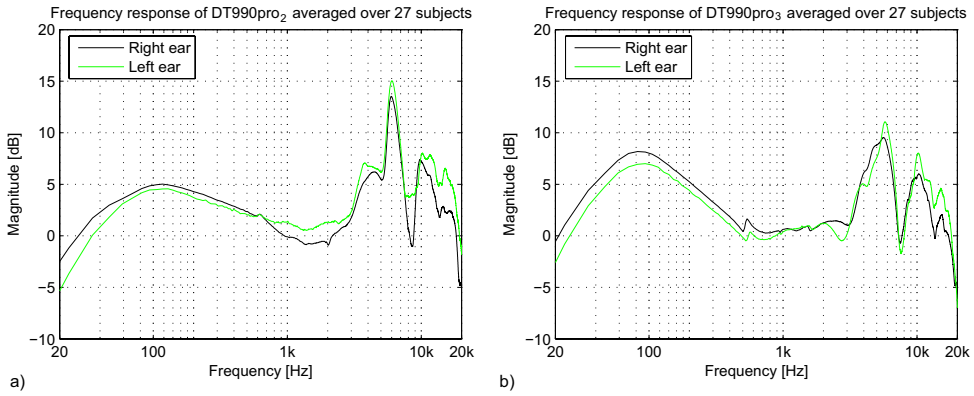



**Figure 5.10:** Frequency responses of two different DT990pro headphones measured on 27 human subjects.

This means that the different measurements should ideally be done with the same headphones, and these should then also be used in the listening experiment. Another solution could be to make an average over different headphones, separately for the VALDEMAR and human subject measurements, and then using these on a completely different set of headphones. However, the headphones used by Emine Çelik have also been measured on VALDEMAR. Based on this, one pair of these headphones will be used in the listening experiment, and the corresponding measurements from VALDEMAR and human subjects are used for the equalization filters. Figure 5.11 shows all four equalization targets, including the general diffuse field curve and the loudspeaker response.

Now that the measurements are from the same pair of headphones, the difference between VALDEMAR and human subjects is not as prominent. However, differences still exist especially for frequencies above 3 kHz. Comparing the two sets of measurements with the general diffuse-field curve, reveals that the measured DT990pro are profound different from the recommended diffuse-field characteristic.

The loudspeakers used in the multi-channel room was positioned in an existing setup (refer to Chapter 4) and thus, it was not possible to move them in order to measure their anechoic frequency response. However, these measurements have already been done by Sylvain Choisel and Florian Wickelmaier and the results were shown previously in Figure 5.1a. Here it was seen that the loudspeaker responses are identical except for small deviations of maximum 1 dB. For this reason, the target function for the loudspeaker equalization filter is chosen as an average of the five measurements. This average is shown together with the headphone responses in Figure 5.11.

The chosen equalization targets are used to design equalization filters according to the *invfreqz* method explained in Section 3.4.4 after shaping the target function as explained in Section 3.4.3. Plots of the target functions and the derived filters are shown in Appendix D on page D21.

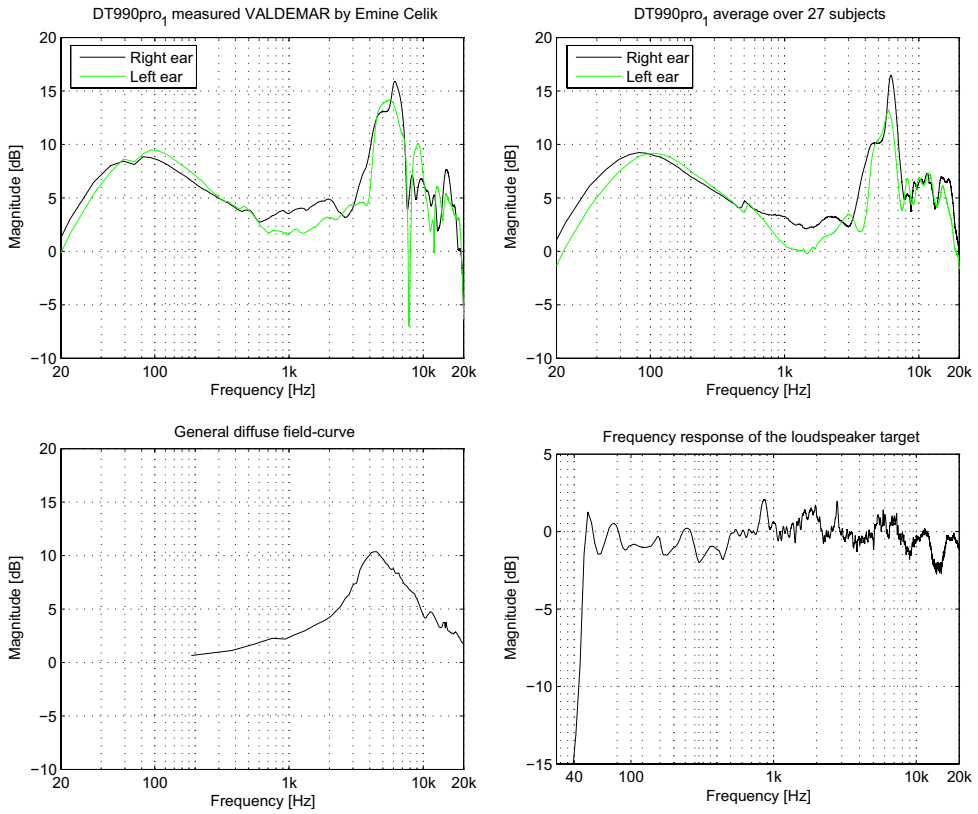



**Figure 5.11:** The four frequency responses that will be used as target functions in the inverse filter implementation. Note that data for the diffuse-field curve only exist down to about 200 Hz and then goes toward 0 dB at DC.

## 5.3 MATLAB processing

All the data from the measurements are now processed and ready to be combined into the complete binaural synthesis system. As described in the system specifications (see Section 2.3) a GUI should handle the user input with regard to the different parameters. Depending on these, the correct processing must then be applied to the selected input signal. The pre-processed data from the previous section is saved in individual MAT-files, which are loaded according to the settings made in the GUI. The GUI itself, and how this is used, is described in Appendix C, while this chapter only focuses on the signal processing.

### 5.3.1 Implementation approach

Figure 5.12 shows the complete signal path from a DVD player to the binaural synthesis played back in a pair of headphones. The grey area outlines the elements that are included in the processing, while elements outside are supervised by external components. This means that the input is a six-channel wave-file while the output is a two-channel wave-file. The system comprises four main processing blocks. First the influence of the loudspeakers is removed by filtering the individual channels with the inverse of the corresponding loudspeaker responses. Because each channel has two destinations (left and right ear),

the five channels are copied into ten and then convolved with the corresponding BRIRs. Following this, the ten signals must be down-mixed into left and right ear signals and then equalized for the headphone characteristics.

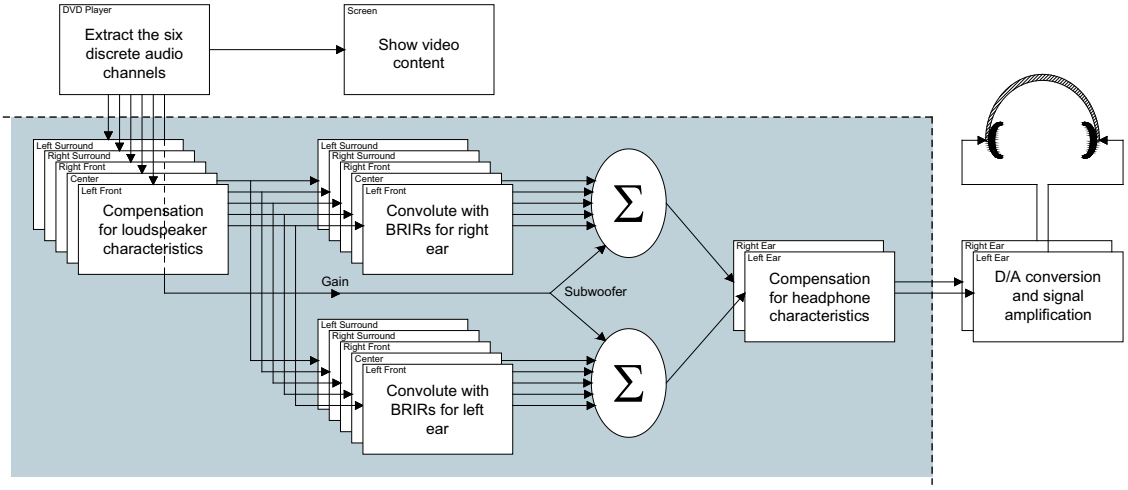



**Figure 5.12:** Complete processing scheme from the input of DVD-decoding device, in this case a DVD player, to the binaural synthesis played back in headphones.

The processing scheme presented in Figure 5.12 can be implemented in numerous ways. Some design criteria and considerations are made in the following, in order to make the binaural synthesis program as flexible as possible.

- Although the program is running offline, it will not be feasible to process all the data as one block. A six-channel PCM signal with a sampling frequency of 48 kHz has a bit rate of 4608 kbps, and thus quickly becomes unmanageable. This means that the input signal must be divided into blocks, which are then processed one at a time.

- Because a real time implementation is the final objective, it will be practical to include some of the same design approaches in the offline calculations. This mainly refers to enhancing the processing efficiency and minimizing the computational cost.

- As described in Section 5.1.3 the output level should match the level of the binaural recordings although overload might still occur. As the signal is processed in blocks, both on the input and output side, it is not possible to normalize the output to ensure proper use of the dynamic range. Instead the output will be monitored, and if clippings occur, these will be counted and a warning message will be presented for the user when the synthesis is finished. The severity of the clipping will also be assessed, so that a global gain option, working on all channels, can be used to adjust the level accordingly.

### 5.3.2 Realization

In order to decrease the computational cost in the convolution process, the filtering with the BRIRs are performed in the frequency domain according to the overlap-add method

described in Appendix B. The processing scheme in Figure 5.12 is mainly to give an overview of the different parallel processes needed for the binaural synthesis. With regard to computational cost, this scheme is highly ineffective and in order to optimize the processing, two different approaches are used. First of all, the system is linear so that the different elements can be shifted around, for example the headphone equalization can be applied to the ten separate signals before the down mixing. In this way the three different filters can be combined into one main filter. The second approach is to pre-process all the static elements which are only changeable from one simulation to another. Based on this, a new processing scheme is presented in Figure 5.13.

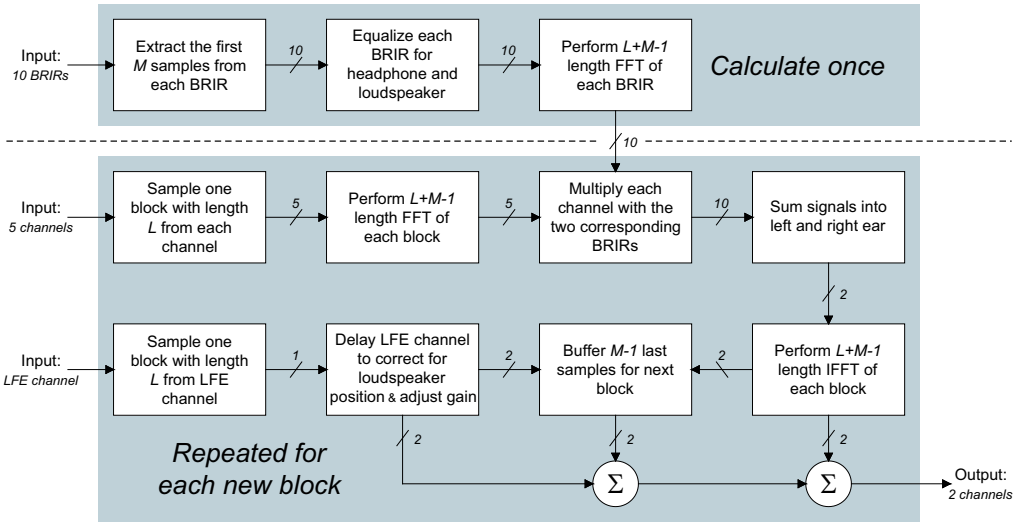



**Figure 5.13:** Improved processing scheme, with regard to computational complexity, in which all the static elements are pre-processed. The convolution is performed in the frequency domain according to the overlap-add method.

Here the ten BRIRs are filtered with the inverse responses of loudspeakers and headphones, and then shifted into the frequency domain so that they are ready to be applied on the input signal. The rest of the processing scheme is repeated for each new input block, which corresponds to the $L$ newest samples from all six channels. The five channels $L$, $C$, $R$, $RS$, and $LS$ are Fourier transformed, copied into ten channels, and then multiplied with the pre-processed filters. The ten channels are then down-mixed to a left and right channel directly in the frequency domain, thus minimizing the following number of IFFTs to two. The last $M-1$ samples are delayed one block in a buffer, while the first $L$ samples are sent to the output. Here the last $M-1$ samples from the previous block are added to the $M-1$ first samples of the new block. It has been verified that the output from the implemented frequency based overlap-add method, is the same as with the time domain convolution.

The LFE channel is passed almost directly to the output. However, to compensate for the 2.5 m from source to listener, the LFE channel is delayed, which also means that the last samples in the block corresponding to this delay must be stored in the buffer. As the LFE channel is passed around all the filtering, this part of the signal will not be equalized for the headphones, which have a small boost in the low frequency area (refers to DT990pro, see previous section). However, if necessary, compensation for this can be made by adjusting the LFE channel gain in the GUI (refer to Appendix C).

The six-channel input wave-file is read directly by using MATLAB's built-in function *wavread*, which can be used to read a specified block at a time comprising all six channels. In order to output one two-channel block at a time, a function is created that opens the destination wave-file and appends the current data to the end of this file. Before saving the data however, the signal is analyzed for possible clippings as described above. As this is only on one block at a time, compensation is not possible. When the whole input signal has been processed, the total number of clippings is displayed as an error dialogue together with the maximum clipping size in dB.

Chapter 6

# Listening Experiment

In Section 2.3 was specified how the system should be validated in terms of BRIR length, equalization, and overall quality. The different filters needed for realizing such a validation were implemented in the previous chapter.

To evaluate the preferred synthesis method, two different listening tests are done with 20 test subjects participating. First, a difference test is performed, which will give an overview of the similarity of the compared sounds. This is explained in Section 6.2. A preference test is then performed to find out which synthesis methods are favored. This involves Section 6.3. The process of choosing the movie sequences, and if a picture is necessary is discussed together with the whole design in Section 6.5. The results should provide information on the best settings for the real time system.

## 6.1 Goal of the listening experiment

The aim of the listening experiment is to investigate which synthesis method is the most appropriate. Depending on the property to be evaluated, the methods that have to be compared are split into groups, according to three criteria (see Section 2.3). From these groups, pairs of comparisons will be determined. A difference will be carried out, followed by a preference, using these pairs.

#### Evaluation of the global quality of the system

Ideally, if there is no error in the measurement chain, the reproduction chain, and in the processing system, the synthesis using the long BRIR should be identical to the recording. Hence, the following sounds are compared:

- Binaural recording referred to as Recording.
- Synthesis using the 0.5 s BRIR referred to as $\text{BRIR}_{Long}$.

If the synthesis is optimally calculated, it is expected that the population cannot make a difference between the two sounds. If there is a difference, it is not expected that the

subjects will prefer the binaural synthesis, as the recording is considered to be the best reproduction under these conditions.

### Evaluation of the BRIR length

The final system is expected to run in real time. It would be desirable to decrease the length of the BRIR, to save computation resources (see Section 5.1.2). But the sound quality should not be significantly lower compared to the best synthesis, which is assumed to be the synthesis using long BRIRs. By making a comparison between signals convolved with different lengths of the BRIR, the compromise between a shorter computation time and the best sound quality will be assessed. This will be evaluated with the following synthesis methods:

- Synthesis using the 0.5 s BRIR referred to as $\mathrm{BRIR}_{Long}$.
- Synthesis using the 0.09 s BRIR referred to as $\mathrm{BRIR}_{Mid}$.
- Synthesis using the 0.01 s BRIR referred to as $\mathrm{BRIR}_{Short}$.

### Evaluation of the headphone equalization

The headphone equalization for the synthesized signals will be done in three different ways: The signals will be equalized for first the artificial head using the DT990pro, then for human subjects wearing the DT990pro, and at last a general diffuse-field equalization for headphones has been done 5.2.1. The following synthesis methods are evaluated:

- Synthesis equalized for VALDEMAR and for the specific headphones DT990pro, referred to as $\mathrm{Eq}_{VALDEMAR}$.
- Synthesis equalized for human subjects and for the DT990pro, referred to as $\mathrm{Eq}_{Human}$.
- Synthesis for diffuse-field equalization, referred to as $\mathrm{Eq}_{DiffuseField}$.

For these comparisons VALDEMAR and human equalization should be the most accurate ones. The equalization for the diffuse-field is general and less appropriate to the specific headphones used (see Section 5.2.1).

### Choice of the pairs

According to the previous considerations, the different synthesis methods that have to be compared are matched into pairs which are listed in the following Table 6.1:

Ideally, a full matrix of all the possible comparisons for one criteria should be done. This means that, the two comparisons $\mathrm{BRIR}_{Long}$ vs. $\mathrm{BRIR}_{Short}$ and $\mathrm{Eq}_{Human}$ vs. $\mathrm{Eq}_{DiffuseField}$ are not included. This is done to reduce the number of pairs and thus, the listening experiment duration.

These five pairs are used in both the difference test and the preference test.

**Table 6.1:** The 5 pairs that will be used for the experiment to compare the different synthesis methods.

| # | Pair |
|---|---|
| 1 | Recording vs $\text{BRIR}_{Long}$ |
| 2 | $\text{BRIR}_{Long}$ vs. $\text{BRIR}_{Mid}$ |
| 3 | $\text{BRIR}_{Mid}$ vs. $\text{BRIR}_{Short}$ |
| 4 | $\text{Eq}_{VALDEMAR}$ vs. $\text{Eq}_{Human}$ |
| 5 | $\text{Eq}_{VALDEMAR}$ vs. $\text{Eq}_{DiffuseField}$ |

## 6.2 Difference test

First, a test based on the three alternative forced choice method is performed. It is an appropriate method to evaluate if the listener is able to detect any difference between the samples of one pair. There is no alternative, like "no difference", so the listeners are forced to choose a sound even if they have some doubts. This test gives an overview if the subjects can hear differences between the two compared sounds.

### Procedure

Three sounds are played, of which two are identical, and one is different. The listener has to find the different sound among the three proposed. In the case the subject is guessing, the answer will be given randomly. There are six different possible combinations for one pair:

AAB, ABA, BAA, BBA, BAB, ABB, where A represents one sound, and B the other.

In order to use the six combinations, each pairs are tested six time. The sample size for one pair is too small to have a significant statistical result for a single subject so that the data of the twenty subjects will be grouped for the calculations. The six combinations for the five pairs represent a block of 30 triples which are played in a totally random order.

### Statistical method and processing

In the test, the listener is asked which sound is different in on triple. The answer can either be correct (= 1) or wrong (= 0), so the answer is Bernoulli distributed, $X_i \sim \mathcal{B}(p)$ where $p$ is the detection probability. In addition, the 30 triples for comparing the five pairs are played in a random order. It can be assumed that each answer is independent and identically distributed (i.i.d).

1. **Probability of detection**:

   The sample size of the answers for one pair for the 20 subjects is 120. Hence, each sample for one pair has a binomial distribution $X_j \sim \mathcal{B}(120, p)$. The statistical method used to determine the probability of detection is the maximum likelihood estimation. In that case, the estimate is

$$\hat{p} = \frac{c}{n}$$

where $c$ is the number of correct answer(s), and $n$ the sample size. The higher the probability of detection the more the subjects can perceive a difference.

2. **Test hypothesis**:

   As the sample size is greater than 30, the approximation that the sample follows a normal distribution can be made, according to the central limit theorem [Papoulis, 1991]. A t-test can be performed to estimate whether the subjects can hear a difference. Under the null hypothesis,*there is no significant difference*, the sample should be normally distributed with a mean value equal to $\frac{1}{3}$. In other words, the probability that the subject chooses the correct sound among the three proposed is $\frac{1}{3}$. The significance level $\alpha$ used is 5%. That leads to reject the null hypothesis if the mean of the observed sample is outside the interval of $[0.25; 0.42]$, (see Appendix E.1).

The final results for the difference test return the pairs which are significantly different, and the detection probabilities. It is also calculated the $p - value$ of the test. The $p - value$ is the probability of type I error that is the probability of observing the given sample result under the assumption that the null hypothesis, $H_0$, is true. If $p \leq \alpha$, then the null hypothesis is rejected:

$$P(H_1/H_0) = p$$

## 6.3 Preference test

From the three forced choice test, the sounds which are significantly different can be estimated. A preference test is done, with the same comparison pairs as in the difference test. The subjects have to answer which sound they prefer. If the listener cannot make a choice, it can either be interpreted as *there is no preference* or *there is no difference*. The two tests can be compared, and part of the ambiguity will be removed. To reduced the test, only sounds which are significantly different have to be evaluated with the preference test. However, this is not possible because of the small sample size for a single subject.

### Procedure

For every trial, two different sounds are played once. The subjects have to choose if they prefer sound A or sound B. Here again, the thirty samples are randomized, and the answers are i.i.d.

### Statistical method and processing

1. **Proportion of the preferred sound**:

   As before, the maximum likelihood estimation from a binomial distribution is used to determine the proportion of the preferred sounds for all the subjects.

2. **Test hypothesis**:

   To determine more accurately if the population can effectively make a choice or not, a Mann-Whitney-Wilcoxon test will be used (see Appendix E.1). This test

compare the differences between the occurrence of sound A and the occurrence of sound B. The only requirement is that the two samples under comparison follow the same distribution, unlike the t-test. This condition is fulfilled as the two sample follow a Binomial distribution. The test evaluates if there is a significant difference between the mean of the two samples. For the listening test experiment, the null hypothesis is that there is no significant preference between the two sounds of one pair. For the confidence level of 5%, that correspond to the interval $[0.43; 0.57]$. If the proportion of the sound A or B are not in this interval, the null hypothesis is rejected.

In the results, the probability of the preferred sound is calculated. The test determines if the subjects can effectively make a choice, and the $p - value$ of the test is also given.

## 6.4 Audiometry test

Before the beginning of the listening test, it is necessary to ensure, that all subjects participating in the test have normal hearing. This can be checked with an audiometric test, where the hearing threshold level of each subject is measured for specific frequencies. The measuring stimulus is presented through headphones. The "hearing threshold level" is the level of the weakest sounds that person can hear over frequency. To make the results of threshold measurements easier to read, these are always plotted in form of an *audiogram*. An example is shown in Figure 6.1, where the hearing loss (HL) is plotted over frequency. The value of 0 dB HL corresponds to the frequency dependent *absolute threshold* [Moore, 2003, p. 55] and thus the vertical axis is relative to human perception of loudness.

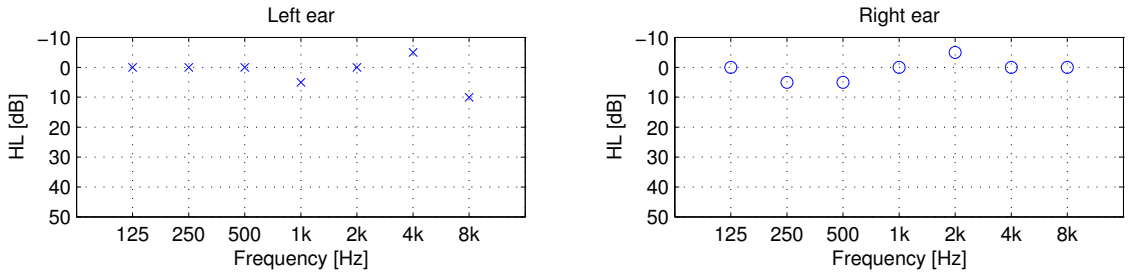



**Figure 6.1:** Example of one audiogram, where a HL of 0 dB corresponds to the absolute hearing threshold.

The hearing threshold measurement was carried out with *Madsen Orbiter 922 Clinical Audiometer* [Madsen, 1997]. This device complies with the standard IEC 645-1 [IEC645-1, 1992] and allows to use different automatic audiometric tests. For determining the hearing threshold, "Auto threshold" test (designed by Hughson & Westlake) or "Bekésy test" can be used. Finally, "Auto threshold" determination was chosen, because the method it uses is based on ascending method described in ISO 8253-1 standard [ISO8253-1, 1991]. Automatic mode also offers short measuring time. More details about this method are given in Appendix E.2.

### Results of audiometry test

People are considered to have a normal hearing if they can hear the sounds of 20 dB or lower across the measured frequency range as stated in [Audiometry, 1993]. All 20 subjects were tested for hearing conditions and the results are depicted in audiograms, which can be seen in Appendix E.2. There was no subject with hearing threshold level worse (higher) than 20 dB and this level itself was reached very rarely. Thus, all 20 subjects could participate in the main listening experiment.

## 6.5 Design of the listening test

After making the decisions about the aims of the listening experiment and proper testing methods, which can be used, it is necessary to specify the key elements of the listening test design. These are:

- Specification and selection of the movie sequences to be used.
- Time chart of the experiment.
- Testing MATLAB program and interface for subjects.
- Selection of the population sample to be tested.
- Choice of the final listening test setup.

These features of the design are discussed and described in the following sections together with some further considerations.

### Basic testing conditions

Even though the listening experiment is made only for the sound evaluation, the surround sound is almost always associated with the picture, e.g. in movies or computer games. That is why there were several binaural recordings of movie sequences made as a reference, which is described in the Appendix A.3.3. If the subject in the experiment is not able to see the scene, it might change the perception of the sound because all visual cues concerning the sound are absent. Therefore it is reasonable to use the picture also in the listening experiment.

The two testing methods to be used in the listening experiment are described in the Sections 6.2 and 6.3. In order to obtain a balanced results out of those proposed tests it is necessary to make the subjects familiar with the testing procedure before the actual test. If the subject is not sure about what to do, this can cause confusion and can have influence on the subject's response. The character of testing samples should not take the subject by surprise either. In order to make the subject familiar with further presented samples, these are introduced before the actual test. This introduction is done only before the difference test because the preferences should not be affected by previous familiarity of testing samples. To make sure that the tested person is skilled in the testing procedure, first of all, the trial test is run. This looks exactly like the true test but include only a couple of repetitions and uses different movie sequences from those used in the following valid test.

### 6.5.1 Selection of movie samples

When evaluating the quality of the binaural synthesis it should ideally be done on as many different movie soundtracks as possible. However, to include all thinkable scenarios would be highly time consuming, not only in experiment duration, but also in finding such representative samples. Instead, only a limited number of samples are selected which must fulfill the following criteria:

- The samples must have a suitable length. If it is too long it will be difficult for the subjects to identify small differences between the samples as they have to remember more, and the listening experiment itself will be unpractical long as each sample must be repeated many times. Because all the subjects are unpaid volunteers, they cannot be expected to come more than once, and the duration of the listening experiment should not exceed one hour. However, the sample duration must be long enough for the subjects to be enveloped in the movie sequence, so that they, through the picture, can get an idea of what the sound should be.

- It is inherent that some of the samples should contain surround effects in which all the channels are active, as this is the main feature that the system should improve compared to normal stereo reproduction. It is also important that the system is able to reproduce natural sounding dialogue as this is contained in all movies.

- In principle, the movie sequences can be chosen from any movie containing scenes fulfilling the above criteria. However, in order to make comparisons with the binaural recordings made in the multi-channel room (refer to Appendix A.3), the sequences chosen must also be included in these recordings.

According to the above criteria, four movie sequences are selected:

1. **Pearl Harbor - Trial:** This sequence is used in the trial session to familiarize the subjects with the experimental procedures. It is 4.171 s long and contains two airplanes in a spinning motion with clear surround effects from the rear channels.

2. **Matrix - Action:** This scene from the first Matrix movie is known as the ”Bullet Time” sequence and is here cut down to 6.924 s. It contains impact sound in form of gunshots, flying bullets in slow-motion, and camera movements relative to these bullets which yield a high amount of surround effects.

3. **Pearl Harbor:** This action sequence follows a group of airplanes flying down between two ships and is 6.590 s long. It includes large amounts of surround effects from gunfire and explosions.

4. **Matrix - Speech:** This sequence is 4.379 s long and contains both male and female speech with a slight low frequency rumbling fading in and out in the middle of the scene.

### Processing the movie samples

Each of the four movie samples must be processed into the seven samples needed in the listening experiment. These were listed in Section 2.3 and comprises the binaural recording, three different BRIR lengths, and three different types of headphone equalizations. But as only one parameter is varied when comparing two samples, all other parameters must be selected. The settings used for all four movie sequences are listed in Table 6.2.

**Table 6.2:** Parameter settings for each of the seven different samples created from each movie sequence.

| # | Sample Tag | Lenght of BRIRs | Headp. Eq. | Loudspeaker Eq. | LFE |
|---|---|---|---|---|---|
| 1 | Recording | - | VALDEMAR | - | - |
| 2 | $\text{BRIR}_{Long}$ | 0.5 s | VALDEMAR | None | None |
| 3 | $\text{BRIR}_{Mid}$ | 0.09 s | VALDEMAR | None | None |
| 4 | $\text{BRIR}_{Short}$ | 0.01 s | VALDEMAR | None | None |
| 5 | $\text{Eq}_{VALDEMAR}$ | 0.5 s | VALDEMAR | None | None |
| 6 | $\text{Eq}_{Human}$ | 0.5 s | Human | None | None |
| 7 | $\text{Eq}_{DiffuseField}$ | 0.5 s | Diffuse-field | None | None |

The loudspeaker equalization option is not used on any of the samples, as it is not a parameter of interest in the listening experiment. Further, the movie samples used in the binaural recording were not pre-equalized for the loudspeakers, and this cannot be done correctly afterward unless all the loudspeakers have exactly the same transfer functions.

Only the movie sequence from Pearl Harbor used in the preference test has noticeable sound emission from the LFE channel. However, this cannot be heard in the binaural recordings so if it is added to the binaural synthesis this will create an unwanted bias.

In Section 5.1.3 it was described how the gain in the binaural synthesis algorithm was adjusted in order for the output level to match the level of the recording. This is a crucial factor in listening tests, as the subjects otherwise will choose the sample with the highest level, and thus creating bias in the experiment. Because of small deviations in overall level in the equalization filters, it is necessary to perform an additional level adjustment. This was done by applying an individual scaling factor to each of the three headphone equalization methods. These factors were adjusted subjectively until the output level was the same for all three equalization filters, and at the same time also matched the binaural recording.

## 6.5.2 Time scheduling of the experiment

As mentioned in Sections 6.2 and 6.3 each out of 5 sample pairs is repeated 6 times for each subject. Total number of samples played in both tests is then:

- For the **difference test**: 3 samples are compared at one moment repeated 6 times for 5 different pairs, that amounts to $3 \times 6 \times 5 = \mathbf{90}$ single movie samples to play for one complete difference test.
- For the **preference test**: only 2 samples are presented at one moment and also repeated 6 times for 5 different pairs $\Rightarrow 2 \times 6 \times 5 = \mathbf{60}$ movie samples to play for one complete preference test.

If the necessary length of the sequence is approximately 6 seconds (as mentioned in Section 6.5.1), then the pure play-time is 9 minutes for the difference test and 6 minutes for the preference test. Response time is assumed to be 3 seconds, which extends the tests to 10.5 minutes and 7.5 minutes respectively. When the testing procedure is quite tiring for the subject, breaks (approx. 10 minutes) should be done at least every 20 minutes. Trial test and introduction to samples can last approximately 12 minutes. If, in ideal case, both tests are made for all three different movie sequences listed in Table 6.5.1, the total time needed for one subject is approximately 80 minutes. This still does not include time for reading instructions and the subject's potential questions. Thus, the duration of one test (means testing of one subject) should be shortened so it does not exceed one hour much.

In the preference test, different movie samples should be present because preferences may vary according to different character of the sound (different content) in each sample. Against this, if the subject can distinguish between two different soundtracks using one movie sample then it can be assumed that the difference is audible also in another movie samples. Thus, it is not necessary to play more than one track in the difference test. If this simplification is used then the representative sample have to be chosen carefully. Finally action sequence from the Matrix "Bullet time" was chosen because it has both low and high frequency content and uses many surround effects.

Now it is possible to draft a preliminary time structure of the whole listening experiment. This is depicted in Figure 6.2 below.

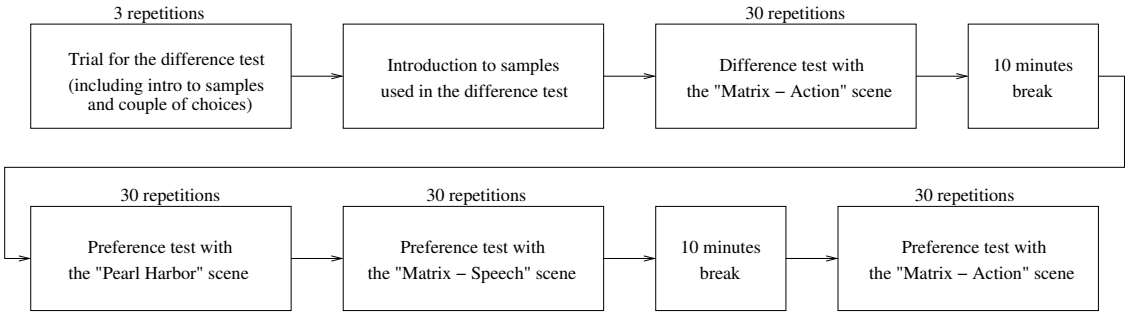



**Figure 6.2:** Time structure of the listening experiment.

As each of the 20 subjects was tested in total for 1.5 or 2 hours (depending if the audiometry test was done before), it was possible to test five subjects a day.

### 6.5.3 MATLAB interface

An interface was created in MATLAB to guide the subject through the listening test process. This program includes all three main parts mentioned in previous sections: trial test, difference test and preference test. For every subject the order of the triples in difference test is rearranged according to a Latin Square ($30 \times 30$). In that way 30 users can participate in the test without having the same sample after another for any user. Each window of the program was created to be as simple as possible for the user. It either contains buttons that start playing movie samples (in the introduction) or letters A - B - (C) highlighted when the corresponding movie is playing, together with the checkboxes

marked also A - B - (C) (both in the difference and the preference test). A more detailed description can be found in Appendix E.3.

Special MATLAB files have been created to store the data in the bin-files, to create the Latin-Square distribution, to get back the calculated data and to play the samples with the media player. Every time when a sequence is played, the media player is started in full screen mode on the TV. While the samples are played, all buttons to handle the user interface are disabled to guarantee that the user cannot abort the testing process.

The user should follow the instructions on the touch-screen and select the item by touching the corresponding checkbox after listening to the signals. The sequence of the signals can only be heard once before the decision has to be taken by the test subject.

The whole setup of the listenig experiment is described in Appendix E.4.

### 6.5.4 Selection of the subjects

There were 20 subjects participating in the listening experiment in the age range from 21 to 32 years and they were mostly students. The whole testing group consisted of 5 women and 15 men. The testing group was not meant to represent the population, there was only need to make the synthesis testing with people who were not involved it the project work. It was required that the subjects had normal hearing, which was checked by audiometry tests.

## 6.6 Results

The results from the listening experiment are presented in the following section. First, the difference test results are presented followed by the results of the three preference tests corresponding to the three different movie sequences. After presenting the results, they will be discussed in a separate section.

### 6.6.1 Difference test

The data from the test were analyzed with MATLAB *ttest*, to estimate if the people can hear a difference, and MATLAB *binofit* to calculate the detection probability. The results obtained are listed in the Table 6.3. The pairs marked with the sign $(*)$ are not significantly different from each other. The means of the samples are plotted in Figure 6.3 with their confidence intervals, with a confidence level of 5%. If the $p-value$ is lower than $\alpha = 5\%$, there is a significant difference (see Section 6.2).

From this results, it follows that the only pair that has no significant difference is the $\text{BRIR}_{Long}$ synthesis method vs. the $\text{BRIR}_{Mid}$ one.

The probability of detection between the $\text{Eq}_{VALDEMAR}$ method and the $\text{Eq}_{DiffuseField}$ method is higher than between the $\text{Eq}_{VALDEMAR}$ method and the $\text{Eq}_{Human}$ method.

The $\text{BRIR}_{Mid}$ method and the $\text{BRIR}_{Short}$ method are clearly distinguishable.

**Table 6.3:** Detection probability and $p-value$ for the difference for the 20 subjects. The pairs marked with $(*)$ are not significantly different.

| Pair | Detection probability | $p-value$ |
|---|---|---|
| Recording vs $\text{BRIR}_{Long}$ | 58% | $\leq 0.01\%$ |
| $\text{BRIR}_{Long}$ vs. $\text{BRIR}_{Mid^*}$ | 39% | $19.5\%^*$ |
| $\text{BRIR}_{Mid}$ vs. $\text{BRIR}_{Short}$ | 83% | $\leq 0.01\%$ |
| $\text{Eq}_{VALDEMAR}$ vs. $\text{Eq}_{Human}$ | 53% | $\leq 0.01\%$ |
| $\text{Eq}_{VALDEMAR}$ vs. $\text{Eq}_{DiffuseField}$ | 90% | $\leq 0.01\%$ |

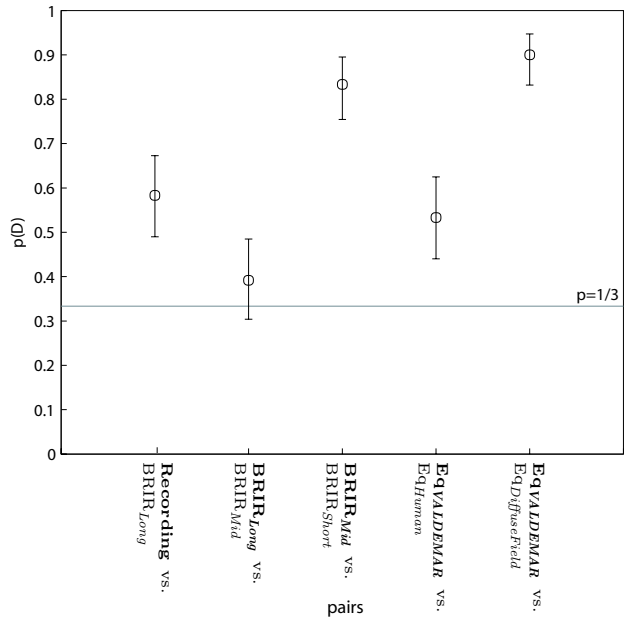



**Figure 6.3:** Probability of detection for the 20 subjects (three forced choice). The circles represent the means and the segments the confidence intervals.

### 6.6.2 Preference test

The *ranksum* algorithm from MATLAB is used to perform the Wilcoxon test. MATLAB *binofit* is used again to determine the maximum likelihood estimate of the proportion of the chosen sound. The case that half of the population choose sound A and half choose sound B can occur. It does not mean that people cannot make a difference, but that the population cannot make a choice between the two sounds, as explained in Section 6.3.

The results of the preference test are presented in Table 6.4 for each movie sequences. The $p-value$ of the test is presented. The means and the confidence intervals of the maximum likelihood are plotted in Figure 6.4

The results vary depending on the movie sequence. For example, for the speech, there is only one significant difference between the equalization using $\text{Eq}_{VALDEMAR}$ method and the equalization using $\text{Eq}_{DiffuseField}$ method, whereas in the Pearl Harbor scene, there are three significant differences.

**Table 6.4:** Preference test results for the 20 subjects. The $p-values$ of the test are given together with the proportion of the preferred sound. This proportion is estimated with the maximum likelihood estimation. When there is no significant preference, the cells remain empty.

| **Pair** | **Sound preferred** | **Proportion** | **p-value** |
|---|---|---|---|
| *Pearl Harbor, sequence 1* | | | |
| Recording vs $\text{BRIR}_{Long}$ | Recording | 82.5% | $\leq 0.01\%$ |
| $\text{BRIR}_{Long}$ vs. $\text{BRIR}_{Mid}$ | – | – | 13.4% |
| $\text{BRIR}_{Mid}$ vs. $\text{BRIR}_{Short}$ | $\text{BRIR}_{Mid}$ | 79.2% | $\leq 0.01\%$ |
| $\text{Eq}_{VALDEMAR}$ vs. $\text{Eq}_{Human}$ | $\text{Eq}_{VALDEMAR}$ | 78.3% | $\leq 0.01\%$ |
| $\text{Eq}_{VALDEMAR}$ vs. $\text{Eq}_{DiffuseField}$ | – | – | 60.3% |
| *Speech, Matrix, sequence 2* | | | |
| Recording vs $\text{BRIR}_{Long}$ | – | – | 30.0% |
| $\text{BRIR}_{Long}$ vs. $\text{BRIR}_{Mid}$ | – | – | 80.0% |
| $\text{BRIR}_{Mid}$ vs. $\text{BRIR}_{Short}$ | – | – | 16.7% |
| $\text{Eq}_{VALDEMAR}$ vs. $\text{Eq}_{Human}$ | – | – | 83.6% |
| $\text{Eq}_{VALDEMAR}$ vs. $\text{Eq}_{DiffuseField}$ | $\text{Eq}_{VALDEMAR}$ | 75.8% | $\leq 0.01\%$ |
| *"Bullet Time", Matrix, sequence 3* | | | |
| Recording vs $\text{BRIR}_{Long}$ | – | – | 13.4% |
| $\text{BRIR}_{Long}$ vs. $\text{BRIR}_{Mid}$ | – | – | 17.0% |
| $\text{BRIR}_{Mid}$ vs. $\text{BRIR}_{Short}$ | $\text{BRIR}_{Mid}$ | 75.0% | $\leq 0.01\%$ |
| $\text{Eq}_{VALDEMAR}$ vs. $\text{Eq}_{Human}$ | – | – | 18.7% |
| $\text{Eq}_{VALDEMAR}$ vs. $\text{Eq}_{DiffuseField}$ | $\text{Eq}_{VALDEMAR}$ | 79.0% | $\leq 0.01\%$ |

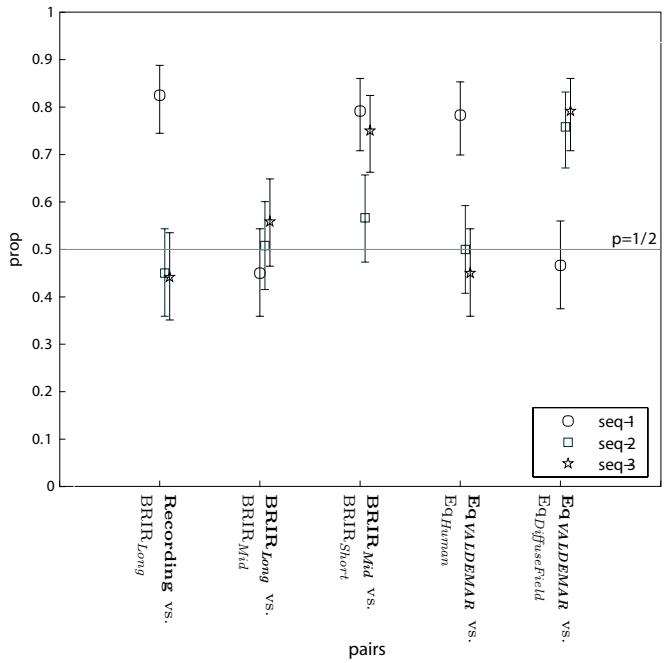



**Figure 6.4:** Proportion of the sound in a pair. The proportion plotted corresponds to the probability of the sound which is bold. The signs represent the means of the samples and the segments the confidence intervals.

## 6.7 Discussion

The results from the listening experiment, presented in the previous section, will in the following be discussed and analyzed in more detail. First, the results from the difference and preference tests are compared followed by a cross-comparison of the different movie samples.

### Correlation between difference and preference test

If the subjects were unable to identify any difference between two given samples, this should also be reflected in the preference test. However, a significant difference in the first test will not necessarily result in a significant preference in the other. It can be assumed that if the difference between two samples is relatively small (low detection probability), then it will be difficult for the subjects to give a uniform answer with respect to preference. While a clearly perceivable difference will more likely lead to coherent preferences. This is also reflected in the results from the difference and preference test conducted on the "Bullet Time" sequence, which is the only common movie sample between the two tests. The subjects did not have a preference when comparing Recording with $\mathrm{BRIR}_{Long}$ and $\mathrm{Eq}_{VALDEMAR}$ with $\mathrm{Eq}_{Human}$, which were the two pairs with the lowest detection probability while still being significant different. That there was no significant difference between $\mathrm{BRIR}_{Long}$ and $\mathrm{BRIR}_{Mid}$ is also reflected in the preference test, which is a very important result when the binaural synthesis is to be implemented in real-time.

### Correlation between movie samples

As expected, the subjects could clearly hear a difference between $\mathrm{BRIR}_{Mid}$ and $\mathrm{BRIR}_{Short}$ and had a strong preference towards $\mathrm{BRIR}_{Mid}$, which emphasizes the importance of having a room (reflections and reverberation) included in the binaural synthesis. However, no significant preference was associated with the "speech" sequence, which is properly related to the fact that only the center loudspeaker is used, and thus limiting the spatial perception. This means that the best approach, with regard to the BRIR length, is to use $\mathrm{BRIR}_{Mid}$ which is truncated after 90 ms.

For the two Matrix sequences the subjects had a clear preference towards equalization for VALDEMAR compared to the general diffuse-field, while no significant preference was associated with this comparison when using the Pearl Harbor sequence. Because the general diffuse-field curve corresponds poorly to the calibration characteristics of the DT990pro, a clearly audible difference exists between these two approaches. The VALDEMAR approach gives the most accurate sound reproduction and will in most cases sound more natural, while the diffuse-field curve does not correct for the low frequency boost in the DT990pro, which might be a preferred characteristic for some subjects. When comparing $\mathrm{Eq}_{VALDEMAR}$ with $\mathrm{Eq}_{Human}$ a significant preference was only found with the Pearl Harbor sequence, where the subjects again choose equalization for VALDEMAR. This is properly correlated with the fact that $\mathrm{Eq}_{Human}$ attenuates low frequencies slightly more than $\mathrm{Eq}_{VALDEMAR}$, which can be seen in the plotted equalization filters in Appendix D. That the Pearl Harbor sequence deviates from the two others in both "equalization tests", because of low frequency deviations between the three equalization approaches,

corresponds to the fact that the Pearl Harbor sequence has the highest low frequency content of the three used movie samples.

When comparing the Recording with $\text{BRIR}_{Long}$, the subjects noticeably preferred the recording when the Pearl Harbor sequence was used, and had no significance preference with the two other movie samples. When listening to the two samples in more detail, it is apparent that these are very different. It sounds like whole elements of the soundtrack are missing in the binaural synthesis, especially noticeable is some missing gunfire from an airplane to the left in the start of the sequence. This seems highly peculiar as such errors are not noticeable in the "Bullet Time" sequence. Thus, it cannot be directly associated with the binaural synthesis and seems more related to errors in the used signals. The original source signals for the binaural recording and synthesis are the same but unknown elements in the signal chain do exist. So far, it has been assumed that the output signals from the DVD-player, used in the multi-channel room, are identical to the signals contained in the six-channel wave-file used for the binaural synthesis. This implies that the AC-3 decoding in the DVD-player is the same as the decoding performed on a PC (refer to Appendix C). By playing back the given Pearl Harbor sequence on different equipment, it has been verified that the binaural recording is the correct sounding version. The six-channel wave-file has been analyzed in more details and the individual channels have been compared with other decoded AC-3 streams. From this it could be concluded that the left and right front channels have an incorrect attenuation of approx 12 dB. If this is compensated for in the binaural synthesis algorithm, the output will be more or less identical to the binaural recording. This error only appears when decoding the AC-3 stream from the Pearl Harbor movie. The results from comparing Recording with $\text{BRIR}_{Long}$, for this particular movie, can thus be discarded, which means that the binaural synthesis does not deteriorate the sound quality compared to a binaural recording.

Chapter 7

# Real-Time Implementation

The only practical application approach with the developed binaural synthesis system is to have it running in real-time. This is also the reason for the constant desire to limit the computational cost, which has been approached throughout the previous implementation phase. In the listening experiment it was found that the BRIR lengths can be truncated to 90 ms without causing a significant deterioration in sound quality. Together with the other result regarding headphone equalization and suitable parameter settings found through informal listening tests, the specifications for the final system can be listed:

1. The BRIRs are truncated after 90 ms.
2. The system is equalized for DT990pro headphones according to measurements performed on VALDEMAR.
3. The system is equalized for the loudspeakers used when obtaining the BRIRs.
4. The LFE channel is added unfiltered to left and right headphone channel with a gain adjustment.

The measured BRIRs contain a time delay corresponding to the 2.5 m between loudspeaker and recording position. This delay is removed from all the measurements (same number of samples from each BRIR), while the overall length is maintained as 90 ms. This also means that no delay must be added to the LFE channel (refer to Section 5.3).

As specified in Section 2.3, the real-time binaural synthesis is to be implemented in an open-source AC-3 decoder codec for Windows. AC3Filter has been selected as the basis, because it is both open-source and an AC3-decoder designed for Windows-based video players (using DirectVideo), like *Windows Media Player*. AC3Filter acts as a filter pin in DirectSound, thus it should work with all video players supporting DirectSound.

The playback takes place with DirectSound passing the raw AC-3 stream into AC3Filter, and AC3Filter passing the uncompressed sound samples back to DirectSound, which will then pass it on to the sound card. The sound sample blocks are equipped with a timestamp, allowing DirectSound/DirectVideo to syncronize the audio and video. AC3Filter has built-in support for up to $\pm 0.5$ s delay, which basically alters the timestamp on the

output sample block. As the binaural synthesis has a fixed block size, it is possible to adjust the delay to match the block size in the binaural synthesis, making sure that the video and audio will stay syncronized.

#### Compiling AC3Filter

Before starting the work on AC3Filter, all the relevant compilers, SDKs and source code files must be installed correctly and compile without errors.

The following software was used (on a PC running Windows 2000):

- AC3Filter 0.70b source-code.
- DirectX SDK 8.1.
- Microsoft Visual C++ 6.0.

Visual C++ and DirectX SDK was configured according to the instructions in the AC3Filter help file. Minor alterations had to be done in the DirectX SDK source-code in order for it to compile properly.

After succesfull compilation, the generated *ac3filter.ax* file is registered automatically by the compiler and is ready for immediate use or debugging. After slight modifications to the AC3Filter GUI, the program was compiled and tested. Due to the modifications, it was apparent that the newly compiled version was running.

## 7.1 Implementation approach

The block diagram for the implementation of the binaural synthesis is shown in Figure 7.1.

The general idea is to intercept the decoded sound, just before it exits AC3Filter and is passed on to DirectSound. This should make sure that all processing performed in AC3Filter is preserved in the output. For the standard 5.1 surround setup, most processing settings will be disabled, but it should still be possible to enable for example dynamic range compression, which can be useful in noisy environments.

The final stage in AC3Filter is the *block()* function in the *AC3Filter* object, which prepares a 256-sample block (convert to correct format) and passes it on to DirectSound. The function is modified to pass the block of samples to a FIFO-buffer (ringbuffer) and then retrieve another block of samples from another buffer, which it will then pass on to DirectSound. If no samples are available in the output buffer, it will simply pass on zeros (no sound). A buffer is needed for each input-channel and each output-channel as depicted in Figure 7.1.

The next step is to implement a function that will take an appropriate number of samples from the input-buffer, apply the binaural synthesis algorithm and place the output in the output-buffer. As the binaural synthesis method works on much larger block sizes than AC3Filter ($2^{14}$ samples relative to $2^8$ samples), the buffers should be large enough to accomodate at least $2^{14}$ samples, preferably $2^{15}$ samples or more to avoid losing samples.

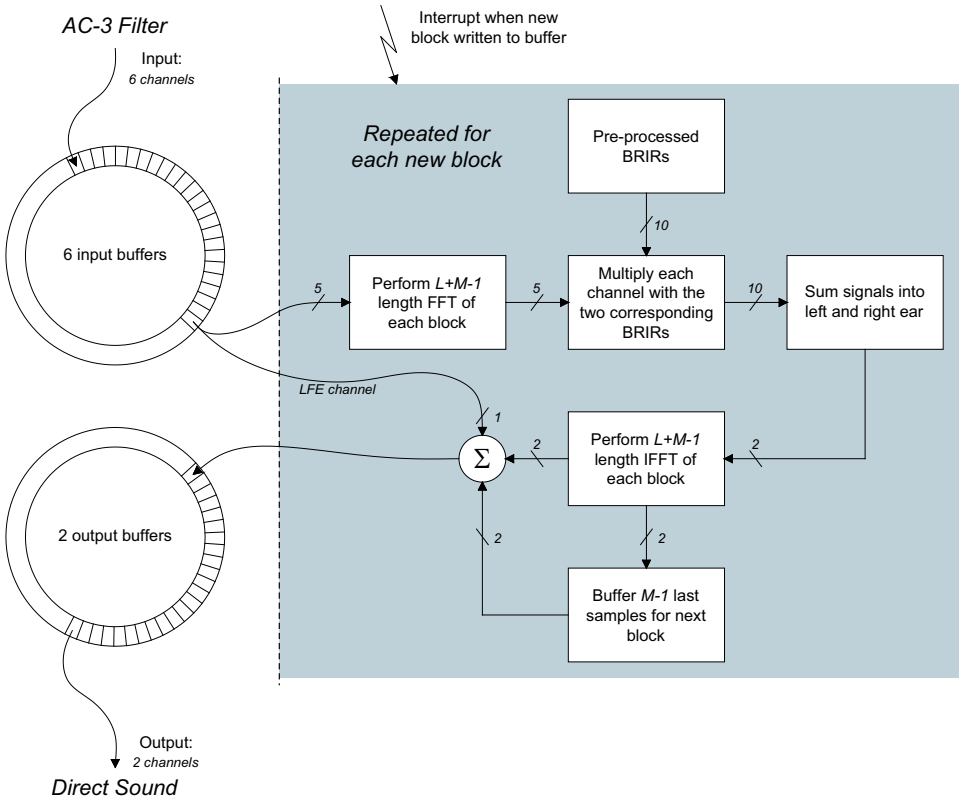



**Figure 7.1:** Approach for implementing the binaural synthesis into AC3Filter, which performs the Dolby Digital 5.1 decoding.

Eight buffers of each $2^{15}$ samples equals 512 kb memory usage, which is small compared to the available memory in modern computers.

The processing will be done in a seperate thread, as the main program needs to transfer 256-sample blocks regularly to DirectSound. The *block()* function will count the number of 256-sample blocks transferred, and when the counter reaches $2^6$ blocks, it means that there are $2^{14}$ samples available in the input-buffer. Then *block()* will trigger a software interrupt and reset the counter. The software interrupt will run in a seperate thread, and involves a $2^{15}$-point FFT, multiplication with the precalculated $2^{15}$-point BRIRs, summing the channels seperate for each ear plus the remainder from the previous block, and in the end an IFFT to go back to the time domain. The first half of the time signal ($2^{14}$ samples) is copied to the output buffer, and the second half is saved as the new remainder for addition to the next block. The fast convolution technique is explained in more detail in Appendix B.

## 7.2 Preliminary testing

The implementation was not finished before the deadline, but some preliminary testing have been done. First of all, the program can be compiled and runs without problems. The first test was to silence the channels one by one by forcing all samples in an output buffer to zero. It was verified that this worked for all channels, and also revealed the

channel order.

A buffer of $2^{15}$ samples have been implemented together with the appropriate functions that will handle the buffers. It have been tested by delaying the sound in one of the input-channels with 1 s., and then playing back an AC3 test file. It was clearly audible that the sound in one channel had been delayed, so the first step with creating the input and output buffers has been verified.

Chapter 8

# Conclusion

The desire for reproduction of surround sound through headphones has formed the basis for developing an algorithm suited for this purpose. The approach has been to add the acoustical filtering from loudspeakers to listener, which is removed when the sound is played directly into the ears of the listener through headphones. The first goal was to create an offline program in which different settings and simulation approaches could be tested. A listening experiment was then to identify the best parameters, which were to be used in a real-time implementation.

An analysis was made to assess some of the key elements regarding the acoustical path between loudspeakers and listener. First the standard 5.1 surround setup was presented with regard to loudspeaker positioning, listening room specifications, and coding technique. Following this, the human spatial perception was analysed. First the general theory on sound source localization was presented, which was then related to the specific situation with a listener situated in a surround sound environment. The principles of Head Related Transfer Functions (HRTFs) were presented, and consequences of using an artificial head for obtaining these was discussed. The importance of room effects was assessed, and it was found that the acoustical filtering can be synthesized by convolving the given signals with corresponding Binaural Room Impulse Responses (BRIRs).

A design phase assessed the recording and reproduction chain needed for reproducing the surround sound experience in headphones. Here the method for achieving the correct sound pressure at each ear was derived, in which the main factor is compensating for the headphone transfer functions. This led to a further assessment of how to realize such an equalization and possible methods for creating inverse filters. The two MATLAB functions, *yulewalk* and *invfreqz* were compared, and *invfreqz* was found to be the most appropriate method.

HRTFs and BRIRs were found for the recommended loudspeaker positions in a standard surround setup, under anechoic conditions (room B4-111), and in a dedicated multi-channel listening room (B5-108), respectively. All the head related measurements were made using the binaural recording head VALDEMAR. The reverberation time for the multi-channel room was found to be lower than stated in the recommendations, which was excepted as this lowered the necessary BRIR length. Headphone transfer functions were measured for two pairs of beyerdynamic DT990pro using VALDEMAR. Finally, different binaural recordings were made to be used as reference in a later listening experiment.

The results from the measurements were analyzed and post-processed before the actual implementation was carried out. Two types of errors were identified in the measured BRIRs. A deviation between BRIRs, that should have been more or less identical, could not be explained and was solved by mirroring the BRIRs, and thus discarding the faulty measurements. This was possible as only some of the measurements were error-prone. The second type of error was spurious reflections visible throughout the response. At the given time nothing could be done about this, but it was not distinctly audible when truncating the BRIRs to the necessary length. Three different lengths were chosen for testing in the listening experiment. One representing the reverberation time of the room, one was optimized for the real-time implementation, and the last only containing the direct sound. Three different headphone equalization filters were also created for the later listening experiment. One was measured on VALDEMAR, one measured on human subjects, and the last one was a general diffuse-field characteristic. A MATLAB program capable of handling the different parameters was developed in order to make the different binaural synthesis files needed for the following testing.

A listening experiment was conducted with 20 subjects who were presented different movie samples, which had been processed according to the different parameters of interest. A three forced choice test should identify differences, and a two sample comparison was used to find possible preferences, all tested with a significance level of 5 %. When comparing the binaural recording with the synthesis, the subjects were capable of detecting a difference and had, for one movie sequence, a preference towards the recording. However, an error has been associated with the AC-3 decoding of this particular movie, and thus making it necessary to discard these results. The two other movie samples did not result in a significant preference, and it can be concluded that the binaural synthesis does not deteriorate the sound quality compared to a binaural recording. No significance difference was found between the *long* and *mid* BRIR lengths that was 0.5 s and 0.09 s, respectively, which is a highly feasible result with regard to decreasing the computational cost in the binaural synthesis. When comparing *mid* and *short* BRIRs the subjects clearly preferred the mid length version when the movie sequence contained surround effects, while no preference was associated with the "speech" scene. This result was expected, as the synthesis with short BRIRs sounds dry and unnatural compared to the longer version. The headphone equalization for VALDEMAR was preferred over both human and general diffuse-field. The largest difference was between VALDEMAR and diffuse-field, which was expected as the latter deviated significantly from the used headphones, and emphasizes the importance of having correct equalization throughout the reproduction chain.

The possibility of creating a real-time system by using an open-source AC-3 filter was assessed and a design approach was made. Preliminary tests have shown positive results, but a complete real-time binaural synthesis system has not been achieved within the time frame.

Overall, it can be concluded that it is possible to reproduce a standard 5.1 surround setup through headphones with a satisfactory result. If it is a suitable substitution for the regular loudspeaker setup depends on the user, but for mobile applications it is a good alternative to normal stereo down-mix. The approach of using BRIRs to compensate for the removed acoustic filters has been appropriate, especially because it was possible to reduce their length considerably. It might be possible to reduce the length even more, which could be assessed through additional listening tests.

# Bibliography


[ATSC, 2001] *Doc. A/52A, Digital Audio Compression (AC-3), Revision A* (August 2001). Advanced Television Systems Committee.

[Audiometry, 1993] *Stig Arlinger and Swedish Audiometrics Method Group, Manual Of Practical Audiometry* (1993). Whurr Publishers ltd. ISBN 1-870332-02-4.

[Bickel and Doksum, 2001] Bickel, P. J. and Doksum, K. A. *Mathematical Statistics* (2001). Prentice Hall, second edition. ISBN 0-13-850363-X.

[Blauert, 1997] Blauert, J. *Spatial Hearing - The Psychophysics of Human Sound Localization* (1997). MIT Press, third edition. ISBN 0-262-02413-6.

[BS775-1, 1992] *ITU-R BS.775-1: Multichannel stereophonic sound system with and without accompanying picture* (1992). International Telecommunications Union - Radiocommunication Assembly.

[Chen et al., 2004] Chen, Z., Bai, Z., and Sinha, B. K. *Ranked Set Sampling* (2004). Springer. ISBN 0-387-40263-2.

[Hammershøi and Møller, 1996] Hammershøi, D. and Møller, H. *Sound Transmission to and within the Human Ear Canal* (July 1996). J. Aco. Soc., Am. 100.

[IEC645-1, 1992] *IEC 645-1: Audiometers - Part 1: Pure-tone audiometers* (1992). International Electrotechnical Commission, Technical Committee No.29: Electroacoustics.

[ISO1996, 1975] *ISO1996: Acoustics. Assessment of noise with respect to community response* (1975). International Organization for Standardization.

[ISO3382, 1997] *ISO 3382, Acoustics - Measurement of the reverberation time of rooms with reference to other acoustical parameters* (1997). International Organisation for Standardization, Technical Committee ISO/TC 43, Acoustics. Subcommittee SC 2, Building Acoustics.

[ISO8253-1, 1991] *DS/EN ISO 8253-1: Acoustics - Audiometric test methods - Part 1: Basic pure tone air and bone conduction threshold audiometry* (1991). Danish Standards Association.

[Kinsler et al., 2000] Kinsler, L. E., Frey, A. R., Coppens, A. B., and Sanders, J. V. *Fundamentals of Acoustics* (2000). Wiley, fourth edition. ISBN 0-471-84789-5.

[Madsen, 1997] *Operation Manual for ORBITER 922-2 Clinical Audiometer (ver. 2.x)* (1997). MADSEN ELECTRONICS Copenhagen, Denmark.

[Maekawa and Lord, 1993] Maekawa, Z. and Lord, P. *Environmental and Architectural Acoustics* (1993). Spon Press. ISBN 0-419-15980-0.

[Minnaar et al., 2001] Minnaar, P., Olesen, S. K., Christensen, F., and Møller, H. *Localization with Binaural Recordings from Artificial and Human Heads* (May 2001). J. Audio Eng. Soc., Vol. 49, No. 5, pp. 323-336.

[Moore, 2003] Moore, B. C. J. *An Introduction to the Psychology of Hearing* (2003). Academic Press, fifth edition. ISBN 0-12-505628-1.

[Møller et al., 1995a] Møller, H., Hammershøi, D., Jensen, C. B., and Sørensen, M. F. *Transfer Characteristics of Headphones Measured on Human Ears* (April 1995a). J. Audio Eng. Soc., Vol. 43, No. 4, pp. 203-217.

[Møller et al., 1999] Møller, H., Hammershøi, D., Jensen, C. B., and Sørensen, M. F. *Evaluation of Artificial Heads in Listening Tests* (March 1999). J. Audio Eng. Soc., Vol. 47, No. 3, pp. 83-100.

[Møller et al., 1995b] Møller, H., Jensen, C. B., Hammershøi, D., and Sørensen, M. F. *Design Criteria for Headphones* (April 1995b). J. Audio Eng. Soc., Vol. 43, No. 4, pp. 206-218.

[Møller et al., 1996a] Møller, H., Jensen, C. B., Hammershøi, D., and Sørensen, M. F. *Using a Typical Human Subject for Binaural Recording* (May 1996a). 100th Audio Eng. Soc. Conv., Copenhagen, Preprint 4157, pp. 1-18.

[Møller et al., 1996b] Møller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D. *Binaural Technique: Do We Need Individual Recordings?* (June 1996b). J. Audio Eng. Soc., Vol. 44, No. 6, pp. 451-468.

[Oppenheim et al., 1998] Oppenheim, A. V., Schafer, R. W., and Buck, J. R. *Discrete-Time Signal Processing* (1998). Prentice Hall, second edition. ISBN 0-13-754920-2.

[Orfanidis, 1996] Orfanidis, S. J. *Introduction to Signal Processing* (1996). Prentice Hall, first edition. ISBN 0-13-209172-0.

[Papoulis, 1991] Papoulis, A. *Probability, Random Variables, and Stochastic Processes* (1991). McGraw-Hill, Inc., third edition. ISBN 1-870332-02-4.

[Rumsey et al., 2001] Rumsey, F., Griesinger, D., Holman, T., Sawaguchi, M., Steinke, G., Theile, G., and Wakatuki, T. *Multichannel surround sound systems and operations* (2001). Audio Eng. Soc.

[SMPTE, 1991] *RP-173: Loudspeaker placements for audio monitoring in high definition electronic production* (1991). Society of Motion Picture and Television Engineers.

# Appendix

Appendix A

# Measurements

This appendix contains the reports for all measurements that have been done throughout the project and a description of a method that assesses the signal-to-noise ratio (SNR) in the measurement results.

## A.1 Calculation of the signal-to-noise ratio

The SNR is plotted for the measured impulse responses over frequency. It is defined as shown in the following equation:

$$\mathrm{SNR} = 20 \log \frac{S}{N} \tag{A.1}$$

with $S$ being the frequency content of the signal and $N$ being the frequency content of the noise.

The signal is defined as the first part of the impulse response, which contains most of the information. It can have maximum half the length of the impulse response. The noise tail from the second half of each impulse response was used to extract a segment of the same length as the signal and is defined as noise. A Fourier transformation is applied to both and the amplitude is plotted in dB over frequency, normalized to 0 dB for the signal amplitude at 1 kHz. The distance between those two curves is then the SNR. Because the impulse response is used to extract signal and noise, information about the SNR in the measurement result is obtained and not about the SNR during the measurement in itself.

To make the SNR easier to assess, it is calculated as an average over one-third-octave bands and plotted as a bar graph in a second plot.

If the SNR is calculated like explained above, it changes with the length used to extract signal and noise from the impulse response. This has two reasons. First, the level in the amplitude spectrum of the noise changes considerable depending on the length used to calculate the FFT. The second one is that the frequency content for low frequencies of the signal changes with this length. Thus the SNR will be more representative for low frequencies if the chosen length is long enough compared to the assessed wavelength. On

the other hand, the SNR will be more characteristic for high frequencies if the chosen length is short, but still includes the high frequency information of the analyzed impulse response.

For this reason, the SNR plots for HRTFs and BRIRs are shown for two different lengths to extract signal and noise. The longer one, which is half a second, shows the behaviour for low frequencies. The shorter one, which is 0.09 s, represents the high frequency restrictions. For the loudspeaker and the headphone measurement results, the SNR was evaluated in the same way but the plots are only shown for one length.

## A.2 Measurements in the anechoic chamber

To measure the HRTFs for all needed directions, a single loudspeaker was used. Its frequency response was measured in order to compensate for it afterwards. Then the pressure at the blocked ear canal of VALDEMAR was measured for all needed angles between the artificial head and the loudspeaker. The responses of the used headphones were measured afterwards on VALDEMAR.

### A.2.1 Loudspeaker response

The frequency response of an active Genelec 1031A loudspeaker was measured.

The measurement was performed in the anechoic room B4-111 in the AAU acoustics laboratory and controlled from the adjacent control room S.

An MLS system analyzer (MLSSA) was used with a external 96 kHz clock generator and an MLS order of 16.

To reduce the noise floor, each measurement was made with an average over 20 impulse responses, which lowers the noise floor by approximately 15dB.

The loudspeaker and the measurement microphone was positioned at a height of 1.2 m from the floor, facing each other. The distance between microphone and loudspeaker was 3 m. The grid where the microphone and the loudspeaker were placed on was covered by foam to avoid reflections.

The microphone was calibrated together with its preamplifier and amplifier using the calibration function in the MLSSA system and an acoustic calibrator. The calibrator produced 94 dB sound pressure level at the microphone at 1 kHz and the system was adjusted to that value.

The used equipment is listed in Table A.1. It was checked, that the impedances at the connections between different devices matched so that they did not destruct the results. The SNR was calculated as described above in Section A.1 and is plotted in Figure A.1. It can be seen that the amplitude of the loudspeaker response decays at approximately 50 Hz, thus the SNR decreases for lower frequencies. This is no problem, as the SNR is adequate over the frequency range of the loudspeaker, which is of interest in this measurement.

A sketch of the measurement setup can be seen in Figure A.2.

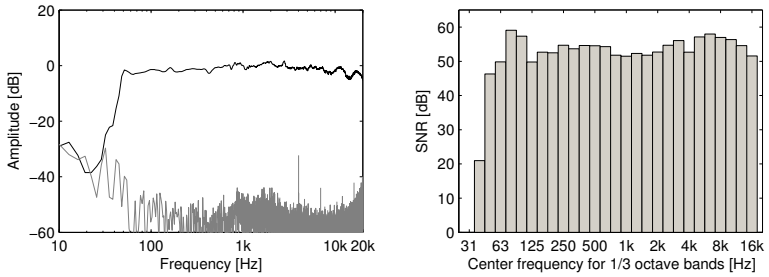



**Figure A.1:** The SNR for the loudspeaker measurement. It was calculated with an impulse response divided into two segments (signal and noise) of 30000 samples each.

**Table A.1:** Equipment used to measure the impulse response of the loudspeaker.

| Description | AAU no. | Type |
|---|---|---|
| Microphone | 08132 | B&K 4165 |
| Microphone preamp | 06560 | B&K 2619 |
| Measuring amplifier | 08022 | B&K 2636 |
| Acoustic calibrator | 33691 | B&K Type 4231 |
| MLSSA | 37493 | on PC Akulab33 |
| Ext. 96 kHz clock generator | 08125 | Philips PM5193 |
| Active loudspeaker | 33986 | Genelec 1031A |
| Active subwoofer | 33994 | Genelec 1094A |
| Misc. loudspeaker, mic. and signal cables | - | - |

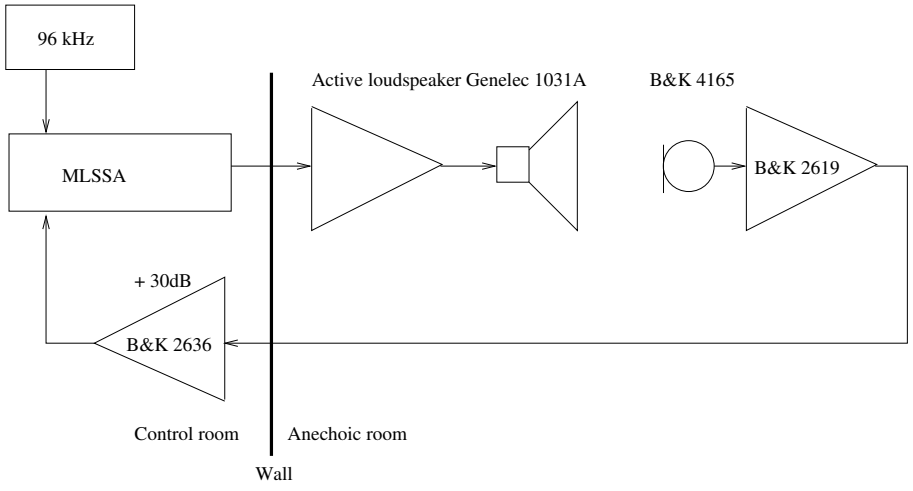



**Figure A.2:** A flowchart of the equipment used to measure the loudspeaker impulse response.

### A.2.2 Obtaining the HRTFs

To measure the HRTFs for sound with different incidence angles, only one loudspeaker was used. A sketch of the setup can be seen in Figure A.5. Part of the metal grid on the floor was covered with acoustic foam (Acoustilux) in order to reduce reflections.

The two microphones inside the artificial head were calibrated one after another together with their preamplifier and amplifier. Again an acoustic calibrator was used with the calibration function in the MLSSA system.

Instead of moving the loudspeaker, the artificial head was rotated to measure $0°, \pm 30°, \pm 110°$ and $180°$ relative to the loudspeaker. This refers to the standard angles in a surround setup (see Section 2.1.3).

The used equipment is listed in Table A.2 and a flowchart of the measurement setup can be seen in Figure A.6.

The HRTFs for each ear and each position were measured one after another. The signal recorded by the two microphones in the ears of VALDEMAR were both sent to the control room, so that it was possible to switch between them from there. Figure A.3 shows the SNR for the measurement from the center loudspeaker to the left ear. The impulse response was truncated after half a second. The SNR for low frequencies is around 25 dB. It can be seen that the SNR degrades below 50 Hz with the amplitude response of the loudspeaker. In Figure A.4 the SNR was calculated, using a shorter length for extracting signal and noise of 0.09 s. The SNR for midrange frequencies is approximately 40 dB and worsening towards higher frequencies. This is because of the amplitude drop for those frequencies in the HRTF from this direction. The SNRs for all other measurements were similar corresponding their specific amplitude shape of the HRTF, thus only these two plots are shown here.

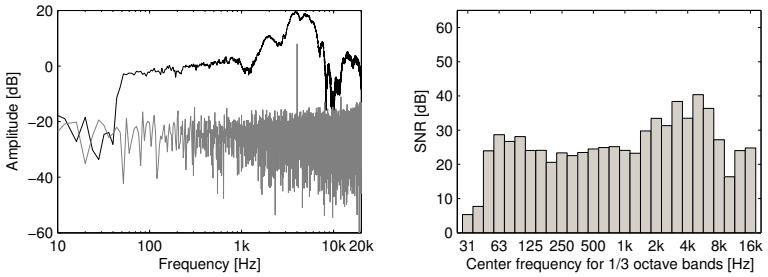



**Figure A.3:** The SNR for the HRTF measurement from center loudspeaker to left ear. It was calculated separating the impulse response into two segments (signal and noise) of half a second each.

To adjust the artificial head in the correct angle, a protractor mounted below the torso was used in combination with a fixed laser pointer on the stand.

The same measurement was repeated afterwards with a subwoofer replacing the full bandwith loudspeaker. The sampling frequency was reduced to 4 kHz. Figure A.7 shows the HRTFs from all surround loudspeaker positions to the right ear for low frequencies. It can be seen that the deviations between the HRTFs are in a range of maximum 3 dB. This means that the HRTFs in this frequency range are not considered to give any significant contribution to the auralization of a surround sound setup.

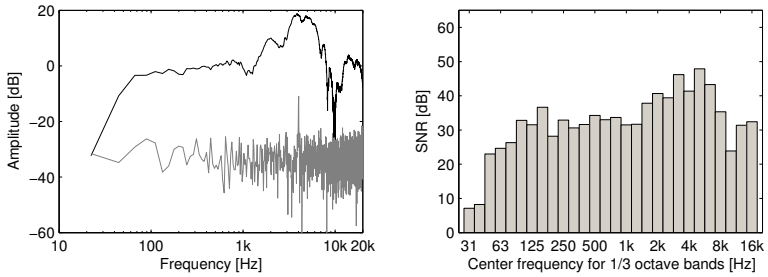



**Figure A.4:** The SNR for the HRTF measurement from center loudspeaker to left ear again. This time it was calculated extracting two segments from the impulse response (signal and noise) of $0.09\,\mathrm{s}$ each.

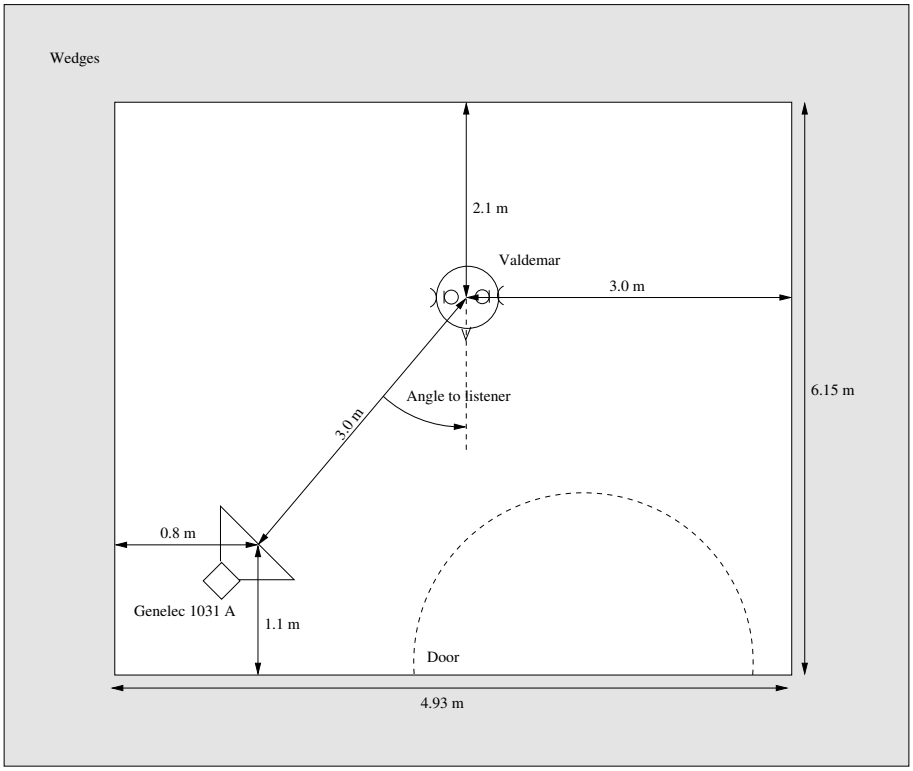



**Figure A.5:** Setup used to measure the HRTFs under different angles.

### A.2.3 Headphone transfer functions

The measurement of the headphone transfer functions were carried out using VALDEMAR. The artificial head is equipped with two condenser microphones, which were calibrated before the beginning of the measurements.

The used setup was similar to the one used in Section A.2.2 and is shown in Figure A.9. The equipment including the additional headphone amplifier and the measured headphones are listed in Table A.3. It was checked that the impedances at the connections between different devices matched so that they did not destruct the results. The SNR for the measurement on the left ear can be seen in Figure A.8. The one for the right ear is similar, thus it is not plotted here. The SNR is not crucial for this measurement as the headphone transfer function is smoothed later in the processing.

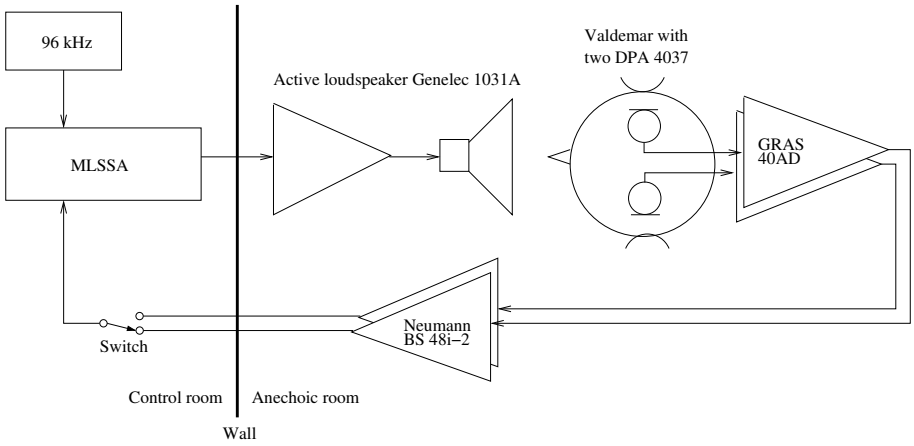



**Figure A.6:** A flowchart of the equipment used to measure the anechoic HRTFs.

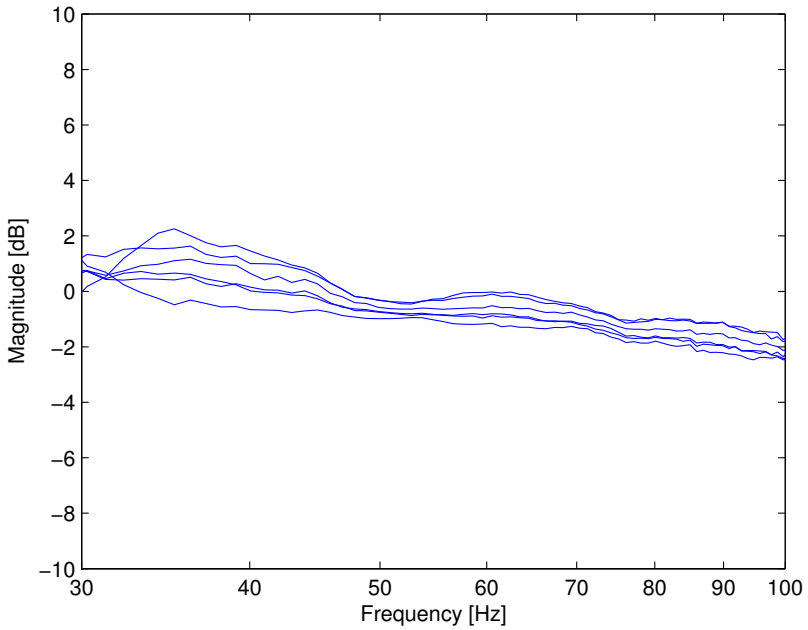



**Figure A.7:** HRTFs for low frequencies from all surround positions to the right ear, measured with a subwoofer.

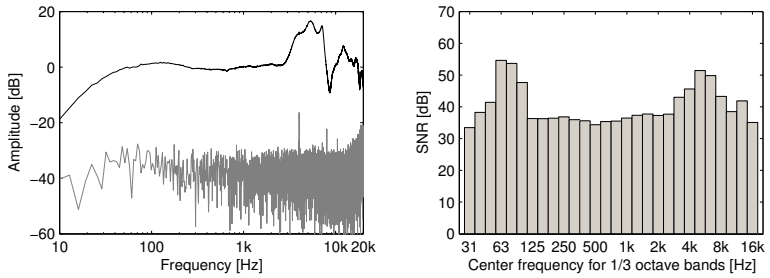



**Figure A.8:** The SNR for the measurement of the headphone transfer function on the left ear. It was calculated by dividing the impulse response into two segments (signal and noise) of 30000 samples.

**Table A.2:** Equipment used to measure the HRTFs.

| Description | AAU no. | Type |
|---|---|---|
| Binaural recording head | 2150-01 | VALDEMAR |
| Left microphone | 56517 | DPA 4037 |
| Right microphone | 56516 | DPA 4037 |
| Left microphone preamp | 56521 | GRAS 40AD |
| Right microphone preamp | 56520 | GRAS 40AD |
| Microphone amplifier | 2018 | Neumann BS 48i-2 |
| Acoustic calibrator | 33691 | B&K Type 4231 |
| MLSSA | 37493 | on PC Akulab33 |
| Ext. 96 kHz clock generator | 08125 | Philips PM5193 |
| Active loudspeaker | 33986 | Genelec 1031A |
| Active subwoofer | 33994 | Genelec 1094A |
| Misc. loudspeaker, mic. and signal cables | - | - |

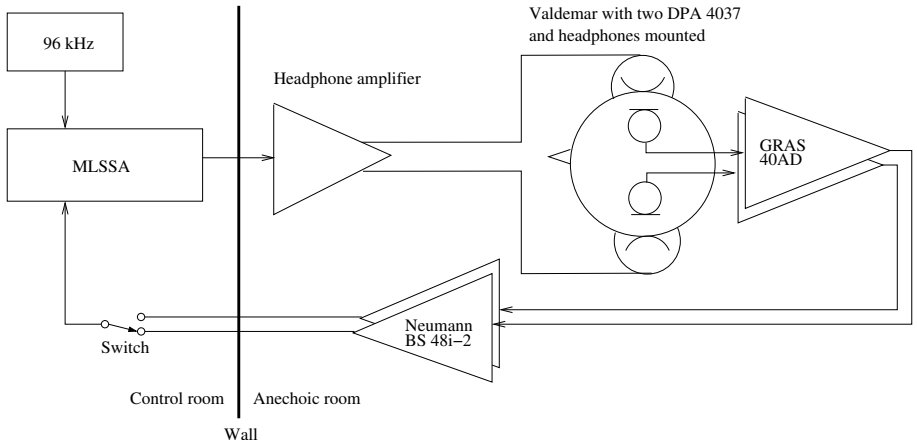



**Figure A.9:** A flowchart of the equipment used to measure the transfer functions of the headphones.

The measurement was made at the entrance of the blocked ear with the headphones mounted on the head. The impulse responses of the headphones were measured with MLSSA. The used MLS order was 16 and the sampling rate was 96 kHz.

As the position of the headphones has an influence on the transfer function, an average over 6 measurements was made as follows: After the acquisition of the signal, the headphone was removed from the artificial head and repositioned on its ears. The headphones were placed so that they enclose the ears as good as possible to avoid leakage.

## A.3 Measurements in the multi-channel room

The measurements took place in the multi channel listening room in the laboratory of the acoustics department at AAU. For the measurements an existing surround setup was utilized which could not be moved, as it was part of a listening experiment in progress. This setup consisted of seven loudspeakers of which five were used, which can be seen in Figure A.15. It was assumed that the influence of the two remaining loudspeakers did

**Table A.3:** Equipment used to measure the headphones on the artificial head.

| Description | AAU no. | Type |
|---|---|---|
| Headphone 1 | 2036-08 | beyerdynamic DT990pro |
| Headphone 2 | 2036-19 | beyerdynamic DT990pro |
| Headphone amplifier | 33240 | HA 903 |
| Binaural recording head | 2150-01 | VALDEMAR |
| Left microphone | 56517 | DPA 4037 |
| Right microphone | 56516 | DPA 4037 |
| Left microphone preamp | 56521 | GRAS 40AD |
| Right microphone preamp | 56520 | GRAS 40AD |
| Microphone amplifier | 2018 | Neumann BS 48i-2 |
| Acoustic calibrator | 33691 | B&K Type 4231 |
| MLSSA | 37493 | on PC Akulab33 |
| Ext. 96 kHz clock generator | 08125 | Philips PM5193 |
| Misc. loudspeaker, mic. and signal cables | - | - |

not disturb the sound field or the acoustical impression of the room. The whole setup remained in the room during all measurements.

First the reverberation time of the room was acquired. Then the BRIRs were measured for all needed directions. Finally, a binaural recording was made in the setup for several surround samples.

### A.3.1 Reverberation time of the multi-channel room

The reverberation time of the room in which the surround setup is built, was compared to the recommendations for multi-channel setups given by the AES [Rumsey et al., 2001]. A measurement was conducted to verify the reverberation time.

A WinMLS system was used which is capable to measure the reverberation time in one-third-octave bands.

A flowchart of the used equipment can be seen in Figure A.10 and a list of the equipment is shown in Table A.4.

To equalize for the AD/DA converter, its frequency response was measured for channel one. This was done by connecting the input directly to the output. The calibration is included in WinMLS and called loop-back.

According to the international standard for reverberation time measurements, the measurement was averaged over three source and three receiver positions [ISO3382, 1997]. This is illustrated in Figure A.11.

### A.3.2 Obtaining the BRIRs

All loudspeakers had the same distance to the listener, which is 2.5 m. This is half a meter less than in the anechoic setup. This radius was chosen before to move the loudspeakers far enough away from the wall. The surround loudspeakers are positioned at 110° in reference to a line connecting the listening position and the center loudspeaker, which is

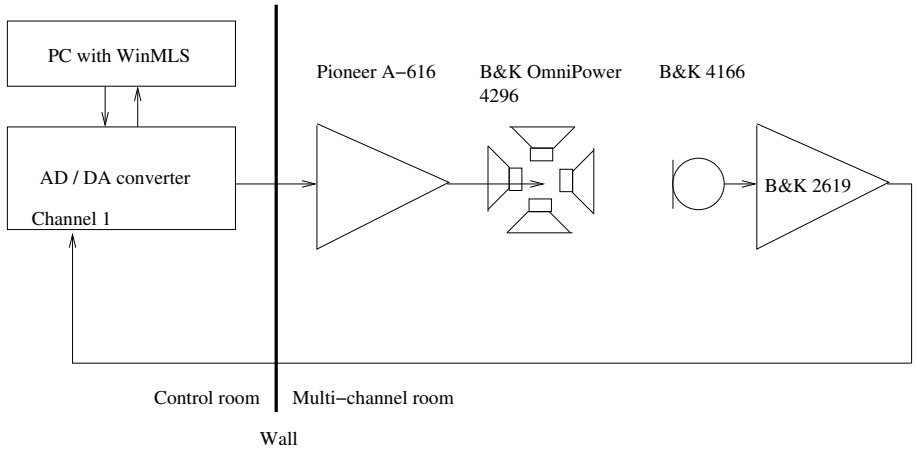



**Figure A.10:** The used equipment to measure the reverberation time.

**Table A.4:** Equipment used to measure the reverberation time in the multi-channel listening room.

| Description | AAU no. | Type |
|---|---|---|
| Amplifier | 08341 | Pioneer A 616 |
| Omni-directional loudspeaker | 33950 | B&K OmniPower 4296 |
| Microphone | 08602 | B&K 4166 |
| Microphone preamp | 06560 | B&K 2619 |
| WinMLS | 53441 | on Fujitsu-Siemens PC Akulab5 |
| AD/DA converter | 33967 | RME TDIF-1 |
| Misc. loudspeaker, mic. and signal cables | - | - |

within the required recommendation, described in Section 2.1.3. In addition a rear center loudspeaker and a subwoofer were added. A sketch of the setup can be seen in Figure A.15.

The loudspeakers used in the setup are Genelec 1031A.

Before any measurements could be performed, the gain of all channels had to be calibrated to an equal level. This was done using a B&K diffuse-field calibrated measurement microphone placed at the listener position (same distance to all loudspeakers), pointed towards the ceiling. Each loudspeaker played in turn band pass filtered white noise, and the RMS-value of the microphone output was calculated. Deviations have been corrected on the individual loudspeaker sensitivity control knobs, until all loudspeakers measured within 0.5 dB. In order not to overload the loudspeaker, the output gain of the PC was set to $-16$ dB.

The next step was to place the artificial head in the room and calibrate the two microphone channels to the same sensitivity. In a fully symmetrical room and with a symmetrical head placed in the center, a signal played back on the front center channel should have the same amplitude at both ears. Using this method with a white noise signal from the center loudspeaker, the two channels on the microphone amplifier were adjusted manually until the input level of the two channels on the A/D converter had the same level. This coarse alignment serves to maximize the SNR of both channels.

Finally, the sensitivity of each microphone channel (including amplifier) was measured by recording the signal from a B&K acoustic calibrator and calculating the RMS-values

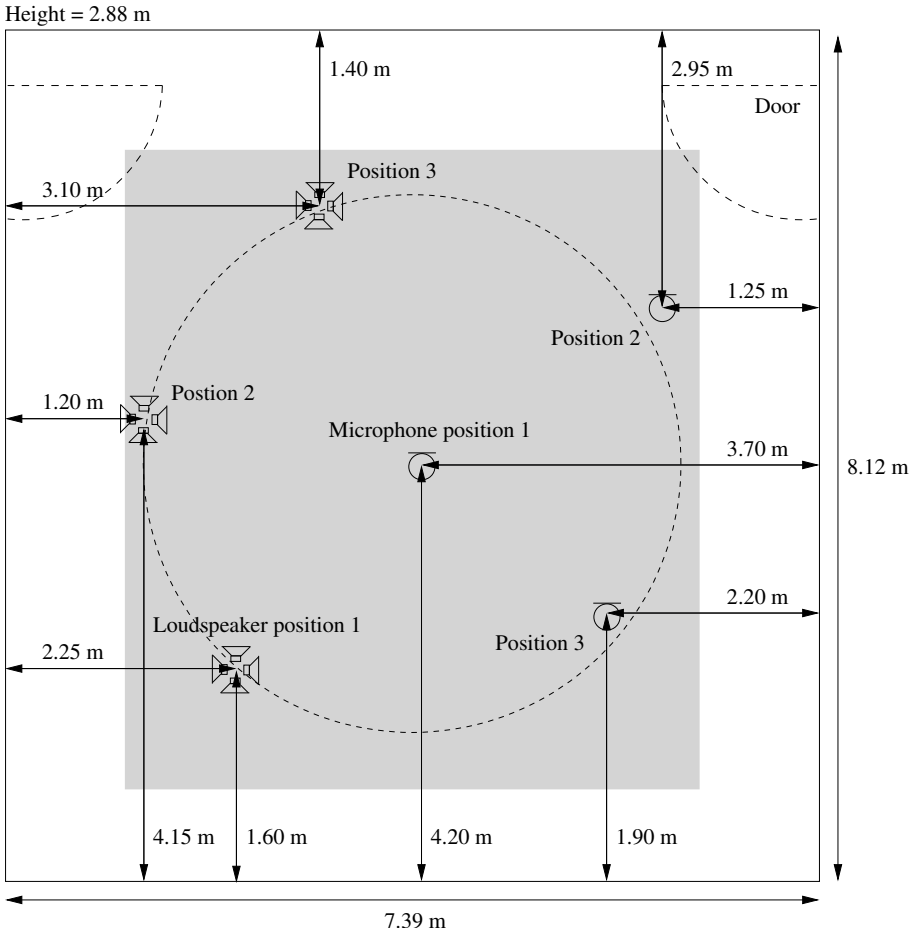



**Figure A.11:** The setup to measure the reverberation time in the multi-channel room. The surround loudspeaker setup was present during the measurement, but is not drawn here.

of these two signals.

Channel one of the AD/DA converter was always used, as this one was calibrated before. The BRIRs for the six loudspeaker positions were measured one after another, one ear at a time. A flowchart of the measurement chain can be seen in Figure A.14.

A WinMLS system was used with 48 kHz sampling frequency. The MLS order was 16. For each BRIR measurement 16 averages were made. The resulting SNR for the the BRIR from center loudspeaker to left ear can be seen in Figure A.12 for a signal and noise length of half a second each. The SNR decreases below the low frequency cutoff of the loudspeaker. In Figure A.13 the SNR is plotted, after this length is reduced to 0.09 s. It can be seen, that the noise spectrum follows the frequency content of the signal. This might be due to errors in the MLS measurement, which are described in detail in Section 5.1.

For the BRIR measurement with the subwoofer, the MLS order was raised to 18 and 64 averages. The used equipment is listed in Table A.5.

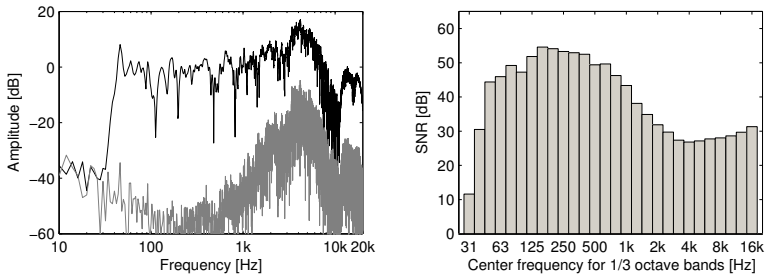



**Figure A.12:** The SNR for the measurement of the BRIR from center loudspeaker to the left ear. It was calculated extracting two segments of the impulse response (signal and noise) of half a second each, which corresponds to the used $\mathrm{BRIR}_{long}$.

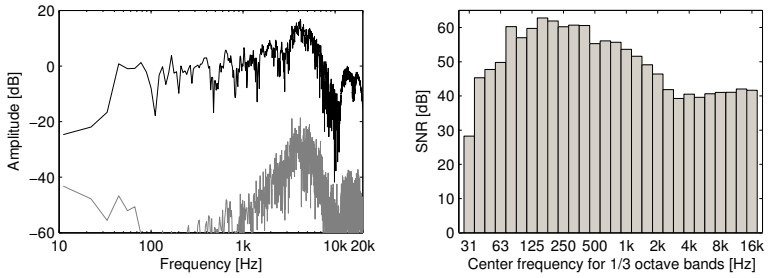



**Figure A.13:** The SNR for the measurement of the BRIR from center loudspeaker to the left ear again. It was calculated extracting two segments of the impulse response (signal and noise) of $0.09\,\mathrm{s}$ each, which corresponds to the length of the used $\mathrm{BRIR}_{mid}$.

**Table A.5:** Equipment used to measure the BRIRs in the multi-channel listening room.

| Description | AAU no. | Type |
|---|---|---|
| Binaural recording head | 2150-01 | VALDEMAR |
| Left microphone | 56517 | DPA 4037 |
| Right microphone | 56516 | DPA 4037 |
| Left microphone preamp | 56521 | GRAS 40AD |
| Right microphone preamp | 56520 | GRAS 40AD |
| Microphone amplifier | 33106 | Rostec LMA 4 |
| Acoustic calibrator | 33691 | B&K Type 4231 |
| WinMLS | 53441 | on Fujitsu-Siemens PC Akulab5 |
| AD/DA converter | 33967 | RME TDIF-1 |
| Active loudspeaker left | 33984 | Genelec 1031A |
| Active loudspeaker right | 33985 | Genelec 1031A |
| Active loudspeaker center | 33990 | Genelec 1031A |
| Active loudspeaker rear | 33986 | Genelec 1031A |
| Active loudspeaker left surround | 33987 | Genelec 1031A |
| Active loudspeaker right surround | 33988 | Genelec 1031A |
| Active subwoofer | 33994 | Genelec 1094A |
| Recording software | - | Adobe Audition |
| Misc. loudspeaker, mic. and signal cables | - | - |

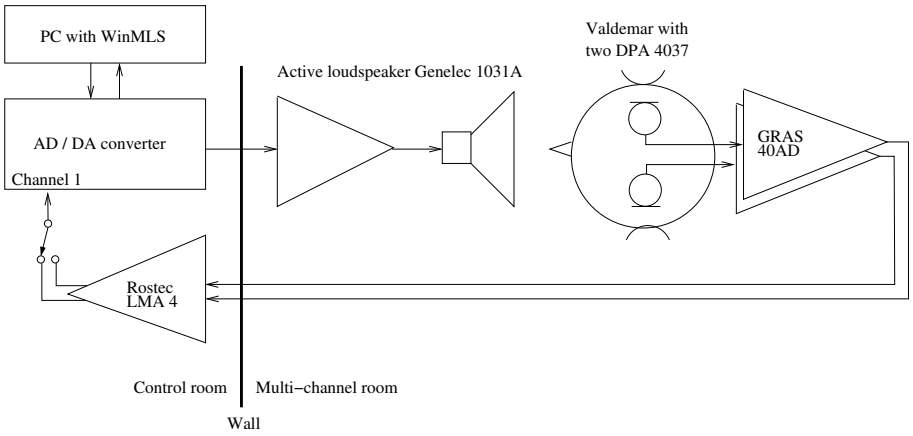



**Figure A.14:** The equipment used to measure the BRIRs in the multi-channel room.

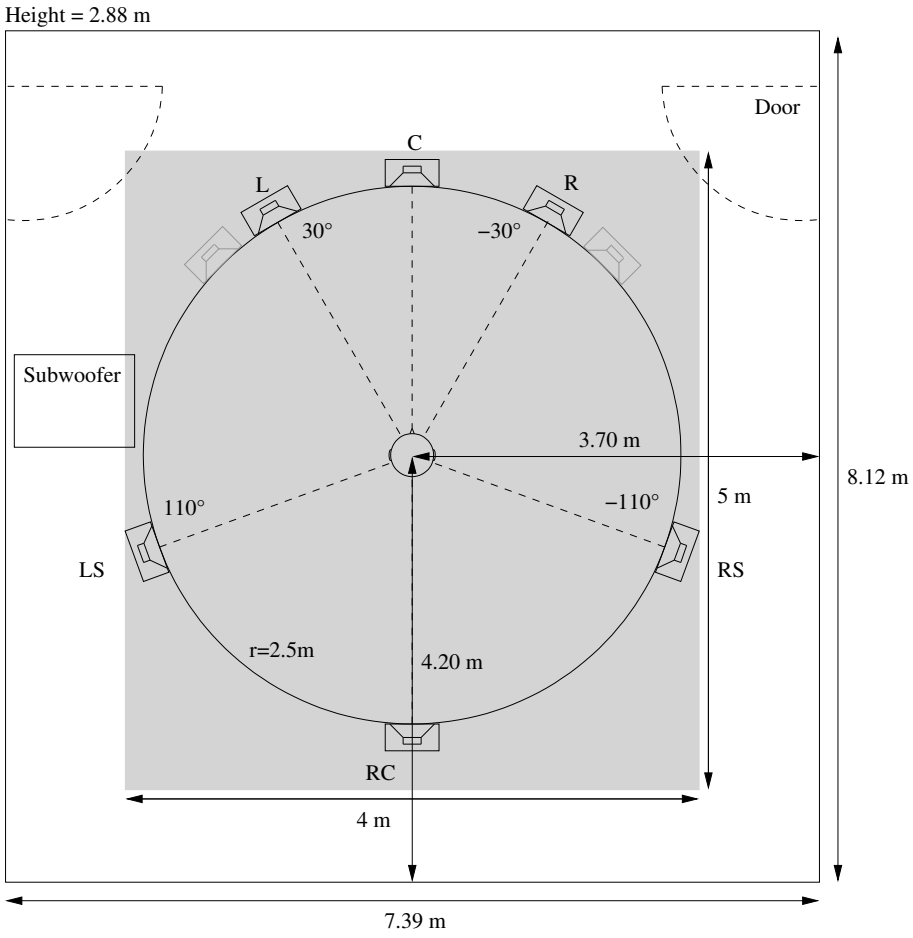



**Figure A.15:** The setup in the multi-channel room to measure the BRIRs.

### A.3.3 Binaural recordings

Several binaural recordings have been made with VALDEMAR in the listening position of the surround setup. Different sequences from movies with a Dolby Digital soundtrack have been played back through the 5.1 setup, using the subwoofer only for the LFE channel.

Two recording channels were needed for recording the binaural signals. Thus the channel two of the AD/DA converter was used in addition to channel one. The gain for the rest of the two different recording channels was roughly aligned before, using the adjustable gain of the microphone amplifier (see previous section). Figure A.16 shows the SNR for the left channel of the recording made with VALDEMAR in the multi-channel room. It is calculated using a recording of the noise floor made with Adobe Audition. The signal is defined as full scale of a wav-file, as the system was adjusted to use the whole dynamic range for the recording. The SNR for the right channel is similar, so it is not plotted here.

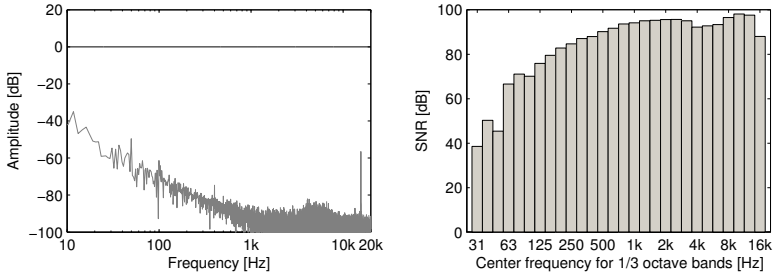



**Figure A.16:** The best possible SNR for the left ear binaural recording with VALDEMAR.

Table A.6 lists the different DVD-movies used for the binaural recording with the approx starting and ending time.

**Table A.6:** The different movie sequences used for the binaural recording. Note that the start and stop times for Pearl Harbor refers to disc 1.

| Movie | Region | Start | Stop |
|---|---|---|---|
| Pearl Harbor - Director's cut | 1 | 1.27.41 | 1.37.46 |
| The Matrix | 1 | 1.40.59 | 1.52.50 |
| The Fifth Element - Superbit | 3 | 1.27.44 | 1.34.26 |
| Saving Private Ryan | 2 UK | 0.05.36 | 0.15.36 |
| Titan A.E. | 1 | 1.02.28 | 1.08.45 |

A sketch of the setup can be seen in Figure A.17 and the used equipment is listet in Table A.7.

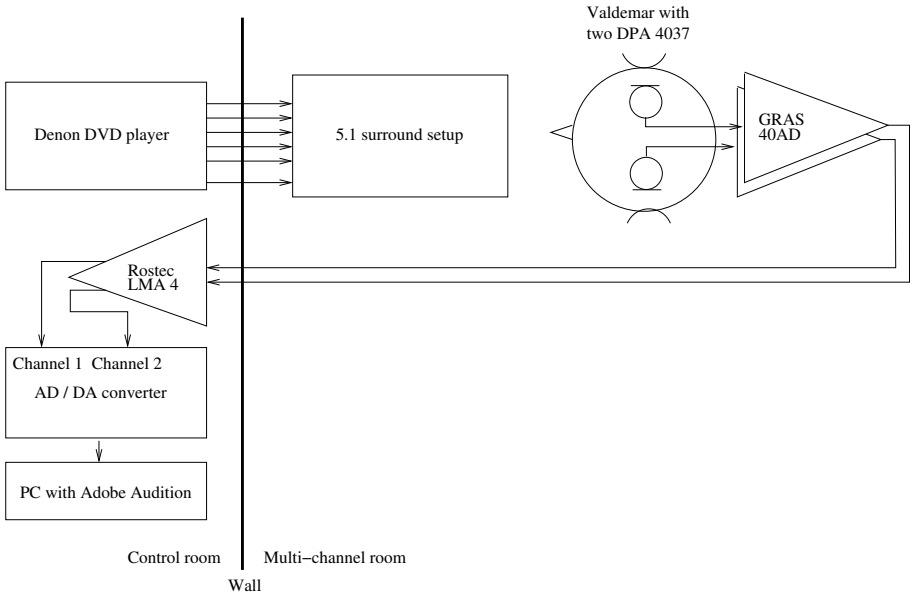



**Figure A.17:** A flowchart of the equipment used to make the binaural recordings in the multi-channel listening room.

**Table A.7:** Equipment used to make the binaural recordings in the multi-channel listenening room.

| Description | AAU no. | Type |
|---|---|---|
| DVD player | 56550 | Denon DVD-2200 |
| Binaural recording head | 2150-01 | VALDEMAR |
| Left microphone | 56517 | DPA 4037 |
| Right microphone | 56516 | DPA 4037 |
| Left microphone preamp | 56521 | GRAS 40AD |
| Right microphone preamp | 56520 | GRAS 40AD |
| Microphone amplifier | 33106 | Rostec LMA 4 |
| AD/DA converter | 33967 | RME TDIF-1 |
| Active loudspeaker left | 33984 | Genelec 1031A |
| Active loudspeaker right | 33985 | Genelec 1031A |
| Active loudspeaker center | 33990 | Genelec 1031A |
| Active loudspeaker rear | 33986 | Genelec 1031A |
| Active loudspeaker left surround | 33987 | Genelec 1031A |
| Active loudspeaker right surround | 33988 | Genelec 1031A |
| Active subwoofer | 33994 | Genelec 1094A |
| Adobe Audition | 53441 | on Fujitsu-Siemens PC Akulab5 |
| Misc. loudspeaker, mic. and signal cables | - | - |

Appendix B

# Convolution Techniques

When reproducing a given loudspeaker position through binaural synthesis the signal emitted from this source must be copied into two signals, which are then convolved with a matching pair of HRIRs or BRIRs. This means that a 5.1 surround setup will require 12 convolutions, which results in a high computational load when the room impulse responses are used. This appendix evaluates different convolution techniques with regard to implementation procedure, computational cost, and input/output delay. The methods are limited to block based processing methods as this is the most appropriate implementation option for the given purpose.

### B.0.4 Time domain convolution

Consider the two sequences $x[n]$ and $h[n]$ with $x[n]$ being the input signal to be convolved with the filter $h[n]$. Then the output $y[n]$ can be written as:

$$y[n] = h[n] * x[n] = \sum_{m=0}^{M} h[m]x[n-m] \tag{B.1}$$

The input sequence $x[n]$ is denoted with the length $L$ while the filter order of $h[n]$ is $M$, which means that the output will have length $L + M - 1$. The filtering process through convolution can be described as shifting the filter in inverse order across the input signal. This means that for each new output sample $M$ multiplications and $M$ accumulations ($M \cdot$ MACs) is required. When working with a sampling frequency of 48 kHz and long filters, this quickly becomes a very computational demanding task.

As the input signal is split up into blocks of length $L$, the signal composition following the convolution must compensate for the additional length of the output blocks. This can be done by using the *overlap-add* method, which is illustrated in Figure B.1 [Orfanidis, 1996]:

The output blocks are added together according to their absolute timing so that $\mathbf{y}_1$ starts at $n = L$, $\mathbf{y}_2$ starts at $n = 2L$, and so on. The overlapping parts consisting of $M - 1$ samples are added together.

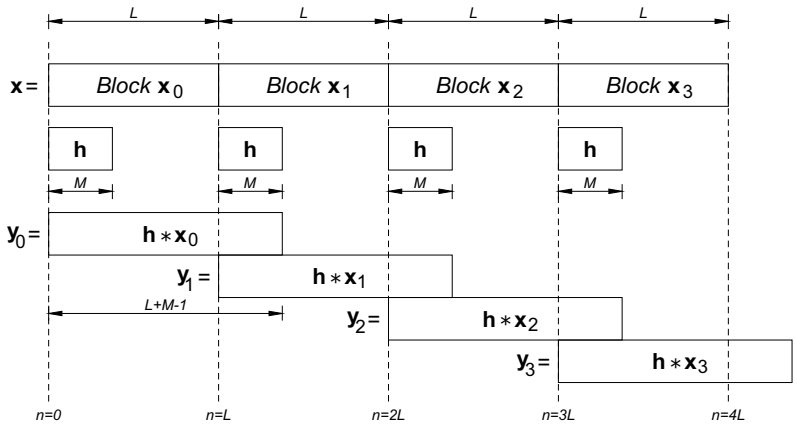



**Figure B.1:** Overlap-add method: The filter represented by the vector $\mathbf{h}$ is convolved with each block in $\mathbf{x}$ and the output is found by overlapping-adding the individual output blocks.

### B.0.5 Frequency domain convolution

In order to decrease the computational complexity the filtering can be performed in the frequency domain. Equation B.1 can be rewritten as:

$$y[n] = h[n] * x[n] = \text{IFFT}(\text{FFT}(h[n]) \cdot \text{FFT}(x[n])) \tag{B.2}$$

To avoid circular wrap-around, the output from Equation B.2 must at least have the same length as in Equation B.1, which means that the FFT length $N$ have to fulfill the requirement: $N \geq L + M - 1$ [Orfanidis, 1996]. The computational cost of performing an FFT is $N\log_2(N)$ complex multiplications [Oppenheim et al., 1998, pp. 638]. Because the Fourier transform of $h[n]$ only needs to be calculated once, it can be pre-calculated, so that the computational cost per $L$ samples is limited to two FFTs plus $N$ complex multiplications. The total number of MACs is then:

$$\frac{\text{MACs}}{L \text{ samples}} = 8N\log_2(N) + 4N \tag{B.3}$$

This means that the filtering will be a compromise between low latency and low computational cost. If the system is to run in real time, then the delay between input and output must be minimized, but so must the computational cost in order to make the product feasible.

Appendix C

# GUI for Binaural Synthesis

This appendix describes how to use the graphical user interface developed to give a fast and reliable parameter control for the binaural synthesis. Basically the GUI handles the different parameter settings made by the user, and then calls the function described in Section 5.3 on page 59 with these. The GUI ensures that this function is called with valid parameters, as error handling is applied to all user inputs. A screenshot of the GUI is shown in Figure C.1.

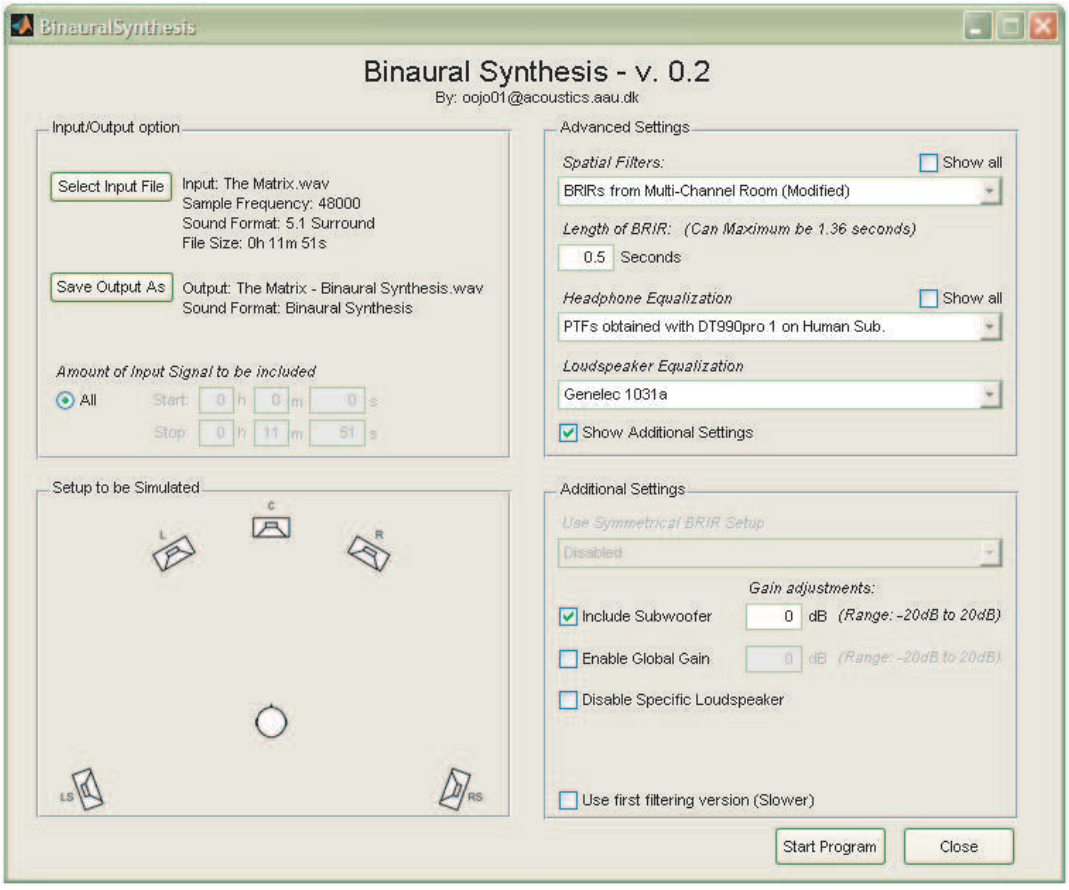



**Figure C.1:** The GUI developed to handle all user inputs to the binaural synthesis system.

The following points describe how the program is used, which functionalities are included, and what the limitations are.

- The program is included on the supplied CD in the "GUI" directory. Everything in this folder is needed to sustain all the program's facilities. The program is started by running the "BinauralSynthesis.m" file, which will open a simplified version of the window shown in Figure C.1. Note that the program is made with MATLAB 7.0 and compatibility with older versions is not guaranteed.

- The first step is to select input/output destinations. The input must be a six channel wave-file with the channels ordered as $L$, $C$, $R$, $LS$, $RS$, and $LFE$. A sampling frequency of 48 kHz is required. Note that binaural synthesis is also possible on two-channel stereo recordings in which case $L$ and $R$ loudspeakers are simulated. When the output file name and destinations have been selected it is possible to start the program with the default settings.

- If only a part of the input signal is to be processed, it is possible to specify a start and end time.

- Different filter options are found under the *Advanced Settings*. Here the type of spatial filters can be selected, including BRIRs from multi-channel room and HRTFs from anechoic chamber. The length of the impulse responses can be specified in seconds, which is then transformed into an integer number of samples according to the sampling frequency. Different equalization approaches to compensate for headphone characteristics are available including a *none* equalization option. Either binaural synthesis with or without loudspeaker equalization is possible.

- Under *Additional Settings* it is possible to set a global gain on all channels between $\pm 20$ dB, which can be used if the output is too low or clipping has occurred (refer to Section 5.3). However, a 0 dB gain should result in an output level matching the binaural recording. It is possible to separately adjust the LFE channel with a gain factor between $\pm 20$ dB relative the global gain. There is also an option of mirroring the spatial filters around the median plane with regard to either left or right ear. Note however, that this option is not possible for the *modified* BRIRs as these are already mirrored (refer to Section 5.1).

If for some reason an error occurs after pressing the "Start Program" button, the GUI must be restarted before it can be used again.

### Additional programs used

The program described above takes a six-channel wave-file as input, which is far from the audio format found on a DVD. However, by using different freeware programs it is possible to extract and convert the needed signals. When the binaural synthesis has been applied, the output must be combined with the picture again. The process of going from the DVD to the finished movie samples used in the listening experiment, is illustrated in Figure C.2.

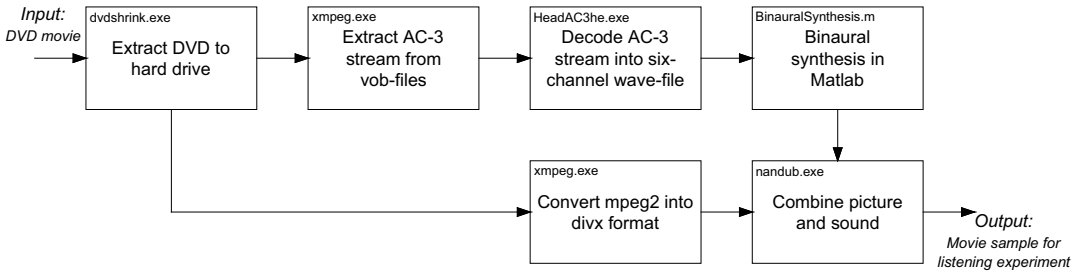



**Figure C.2:** Different programs used for creating the movie samples needed in the listening experiment, when the source material is a DVD-movie.

All the programs illustrated in Figure C.2 are included on the CD, attached in the back of the report.

Appendix D

# Equalization Filters

In Section 5.2.1 on page 57, the equalization targets were chosen to be the following four systems:

- DT990pro measured on human subjects.
- DT990pro measured on Valdemar.
- Generalized diffuse-field design goal.
- Genelec 1031A loudspeaker.

The equalization targets and the derived equalization filters are shown in Figures D.1, D.2, D.3 and D.4. For the 2-by-2 plots, the top row is the left channel, and bottom row right channel.

As no individual optimization have been performed, the order of all filters are equal and set to 50 and 100 for the **a** and **b** polynomias respectively.

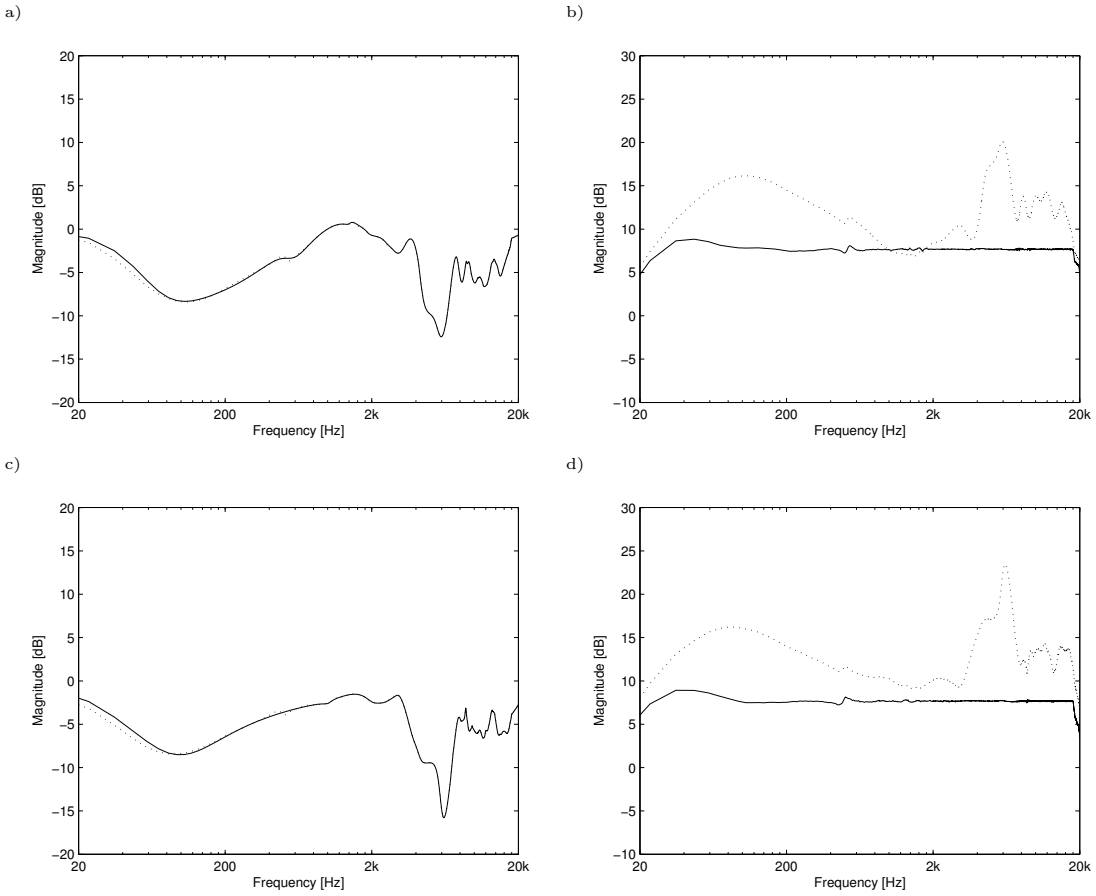



**Figure D.1:** Equalization filters for DT990pro headphone measured on human subjects; **a)** and **c)** shows the target function (dashed) and filter response (solid) for left and right channel, **b)** and **d)** shows the headphone response (dashed) and filtered response (solid) for left and right channel.

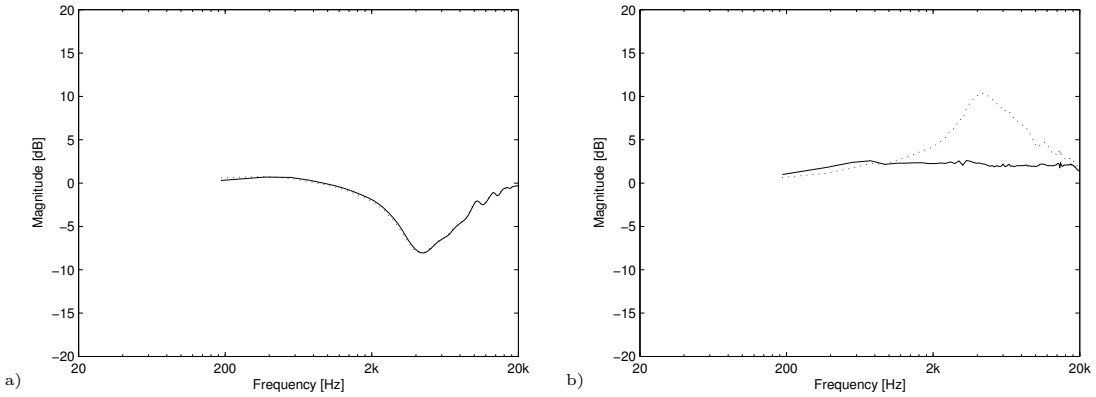



**Figure D.2:** Equalization filter for the diffuse-field design goal; **a)** the target function (dashed) and designed filter response, **b)** the ideal diffuse-field response (dashed) and filtered response (solid).

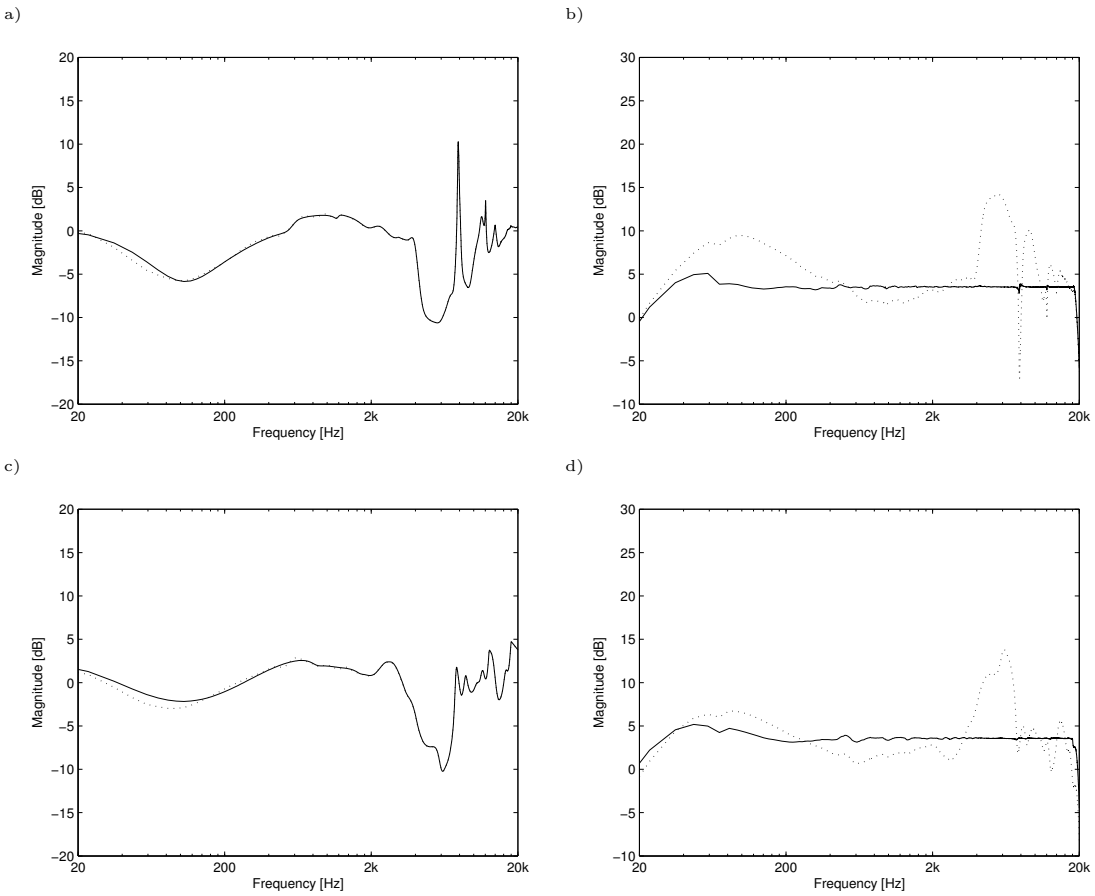



**Figure D.3:** Equalization filters for DT990pro headphone measured on Valdemar; **a)** and **c)** shows the target function (dashed) and filter response (solid) for left and right channel, **b)** and **d)** shows the headphone response (dashed) and filtered response (solid) for left and right channel.

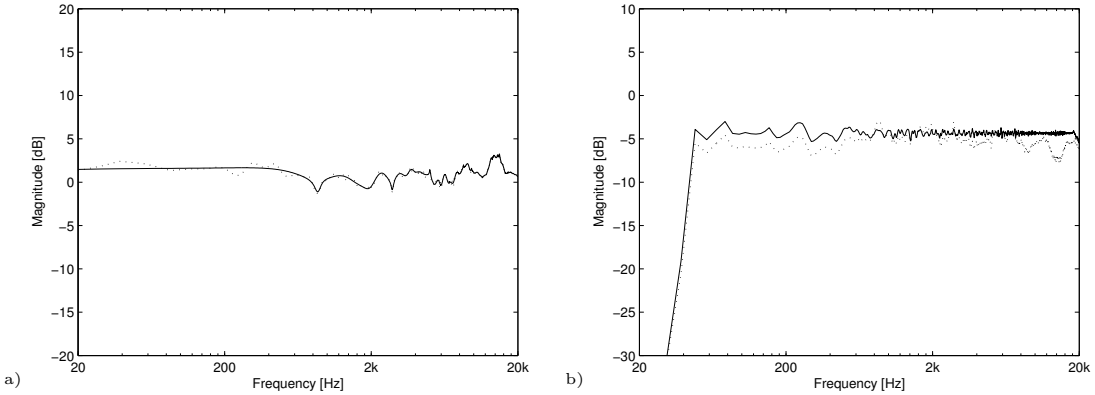



**Figure D.4:** Equalization filter for the Genelec 1031A loudspeaker, averaged across five loudspeaker measurements; **a)** the target function (dashed) and designed filter response, **b)** the average loudspeaker frequency response (dashed) and filtered response (solid).

Appendix E

# Listening Experiment

The listening experiment appendix goes into details for the different statistical methods used. All the audiograms from the 20 subjects are listed. The MATLAB interface program is briefly explained and the setup is described. The interface goes with instructions for the subjects, so they learn what the task is. A copy of those instructions are included at the end of this part.

## E.1 Statistical methods

#### t-test

The t-test requires the sample to be independent and identically distributed (i.i.d.) and normally distributed. These conditions are fulfill because the triples were randomly played and the sample size is greater than 30. To determine the critical value of the sample, a statistical pivot and a level of significance must be determine. The mean and the standard deviation of the sample are unknown, so the correct statistical pivot is $T(\mu)$, which has a Student distribution with n-1 degree of freedom:

$$T(\mu) = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

The significance level $\alpha$ is 5%. It means that there is a risk to reject the null hypothesis when there is no difference of 5%:

$$P(H_1/H_0) = 5\%$$

The critical area can be then calculated:

$$P(t_{n-1,1-\alpha/2} \geq \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \geq -t_{n-1,1-\alpha/2}) = 1 - \alpha$$

where

- n=120 is the sample size.
- $-t_{n-1,1-\frac{\alpha}{2}} = -t_{119,0.975} \sim 1.98$ is the quantile from the student law (according to statistical tables, [Bickel and Doksum, 2001]).
- $\bar{X} = \sum X_i/n$ is the mean of the observed sample.
- $S = \sqrt{(\frac{1}{n-1}\sum (X_i - \bar{X})^2)}$ is the standard deviation of the observed sample. $X_I$ is the realization of one triple, 0 if the answer is wrong, 1 if the answer is correct.
- $\mu = \frac{1}{3}$ is the mean under the null hypothesis.

This equation lead to the following interval:

$$[\bar{X} - s.t_{n-1,1-\alpha/2}; \bar{X} + s.t_{n-1,1-\alpha/2}]$$

In reality, one triple follow a Bernoulli distribution. So the standard deviation can be estimated by $\sqrt{(\bar{X}(1-\bar{X}))}$ ( [Bickel and Doksum, 2001, pp236-237]). Under the null hypothesis, the mean value of the sample must be equal to $\frac{1}{3}$. So $\bar{X} = \frac{1}{3}$.

$$\left[\frac{1}{3} - \frac{1.98\sqrt{(\frac{1}{3}(1-\frac{1}{3}))}}{\sqrt{(n)}}; \frac{1}{3} - \frac{1.98\sqrt{(\frac{1}{3}(1-\frac{1}{3}))}}{\sqrt{(n)}}\right] = [0.25; 0.42]$$

If the observed mean of one sample is within this interval, the null hypothesis is not rejected. The confidence interval for each pair is calculated from the results of the test. The means and the confidence intervals have been plotted in Figure 6.3 on page 73 in the result section.

### Wilcoxon test

The Wilcoxon test is a non parametric test which compares two groups together. The only requirement for this test is that the samples of the two group follow the same statistical distribution, whatever it is. Under the null hypothesis, the two groups are assumed to have the same mean. Unlike the three forced choice method, only two different sounds are presented, which allows to use the Wilcoxon test. The difference between the occurrence of the sound A and the sound B is calculated for each individual. The test-statistic for the Wilcoxon test is calculated as follows:

$$U = \sum\sum F\{X_i - Y_j \leq 0\}$$

where

- $F$ is the estimate of the distribution function.
- $X_i$ is the answers of the evaluation of "sound A" for the $i_{th}$ (either 1 if 'Sound A' is chosen or 0).
- $Y_i$ is the answer of the evaluation of "sound B" for the $i^{th}$.

According to [Chen et al., 2004], the evaluated mean and the evaluated standard distribution of the test-statistic under the null hypothesis are:

- $E(U) = \frac{1}{2}MN$ where $M$ and $N$ are respectively the sample size of sound A and B. In the case of $M = N = 120$, $E(U) = 7200$.
- $Var(U) = \frac{1}{12} = MN(M + N + 1)$. For the listening test experiment condition, $Var(U) = 289200$.

Applying the central limit theorem, $\frac{U-E(U)}{\sqrt{Var(U)}}$ follows a normal distribution. In that case, the quantile of the median correspond with the quantile of the standard distribution $u_{1-\alpha}$. For a significant level of $\alpha = 5\%$, $u_{1-\alpha} \sim 1.980$. The confidence interval of the test-statistic under the null hypothesis correspond to

$$\left[E(U) - u_{1-\alpha}.\sqrt{Var(U)}; E(U) + u_{1-\alpha}.\sqrt{Var(U)}\right] = [6135.2; 8264.8]$$

The sample size of the test-statistic is $MN = 120 \times 120$, as the 120 variables of sound A are compared with the 120 variables of sound B. In that case, the test-statistic represents the probability of choosing the sound A rather than the sound B:

$$F(x - y) = F(0) = P(X - Y \leq 0)$$

In other words, if the two sounds are not significantly different, the proportion of the sound A is in this interval:

$$\left[\frac{6135.2}{120 \times 120}; \frac{8264.8}{120 \times 120}\right] = [0.426; 0.574]$$

## E.2 Audiometry test

The audiometry tests took place in a well sound insulated room Cabin B (B5-104), where only a subject, an experimenter and the measuring device was present during the test. The instructions for the subjects were simple: to press a button whenever they can hear any sound. The subjects were asked to sit down so that they cannot see the audiometer screen and adjust the headphones to make it fit best to the ears. Then they got two response buttons for both left and right ear. For the stimuli constant pure tones at certain frequencies were used, which are 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, and 8 kHz.

### Auto Threshold test procedure

The device was set to use only frequencies listed above during the measurement. One tone at 1 kHz is played to the first ear at the level of 50 dB. If the subjects can hear the tone, they give a response, that means presses the corresponding button (left/right) within 1.7 s from the beginning of the stimulus. If the tone is audible, it is attenuated in 10 dB steps until the subject cannot hear the tone any more. When this occurs, the level is raised in steps of 5 dB, until the subject can hear the tone again and presses the button. This procedure is repeated until the response is given at the same level twice consecutively, while the tone is in the increasing phase. Then the reached level is accepted as a threshold and the whole procedure is started again at the next frequency with a starting level 20 dB higher than this threshold.

The sequence of tested frequencies starts at 1 kHz and increases. After testing the highest frequency the program automatically returns and continues with frequencies from 1 kHz downwards (excluding 1 kHz). Finally, the threshold is measured at 1 kHz again and the result has to be the same as at the beginning of test. The used sequence is thus 1 kHz, 2 kHz, 4 kHz, 8 kHz, 500 Hz, 250 Hz, 125 Hz, and 1 kHz again. The procedure is repeated for the second ear.

If there are any doubts about some result after the test is finished, it is possible to measure the hearing threshold again for specific frequencies.

### Results of the audiometry test

The normal hearing is in between 20 dB and −10 dB HL for the range of frequencies from 125 Hz to 8 kHz. The audiometry test has been done before the listening test for the subjects who did not have their hearing checked yet. Those who have already passed an audiometry test were asked to bring their data. The results of the audiometry tests from the twenty subjects are plotted in Figure E.1 and Figure E.2. In the graphics, the initials marked with star are the subjects whose audiometry tests have been proceeded previously, but where the 125 Hz frequency measurement is missing. As low frequencies are not crucial for source localization and spaciousness perception, it has been chosen not to repeat the audiometry test. The value of the threshold level for that frequency is plotted as 0 dB.

## E.3 MATLAB interface

The interface used in the listening experiment was programmed in MATLAB. It consists of several parts: the trial test, the difference test and the preference test.As initialization, the experimenter has to choose a number for every test subject. Depending on this number, the test sequences will be adjusted for the user. The order of these sequences is stored in a matrix which represents a $n \times n$ Latin Square, to guarantee that no sequence will be played after the same one twice for up to $n$ users and $n$ different test sequences. Playing a sample means that its picture will be shown on the TV-screen and the sound is played through the headphones. This can be achieved by starting a media-player through MATLAB's "dos"-command on the enlarged screen (TV).

After this initialization a trial test of the first task, the difference test, is done. It consists of only three sequences.

The difference test starts after that by presenting all possible samples used in this test to the subject. Every sample can be repeated by the users if they want to watch it again. A screenshot of the program when all signals are introduced can be seen in Figure E.3.

After this familiarization the test sequences are played, three samples forming a test triple. The sample playing at the moment is highlighted by setting the background colour of the letter belonging to it to blue. The users have to check the box for the sample that they think is different. A screenshot of this task in the experiment can be seen in Figure E.4. The next sequence is played when the "Next" button is pressed. When this procedure has been repeated 30 times, the difference test is finished and a break is recommended.

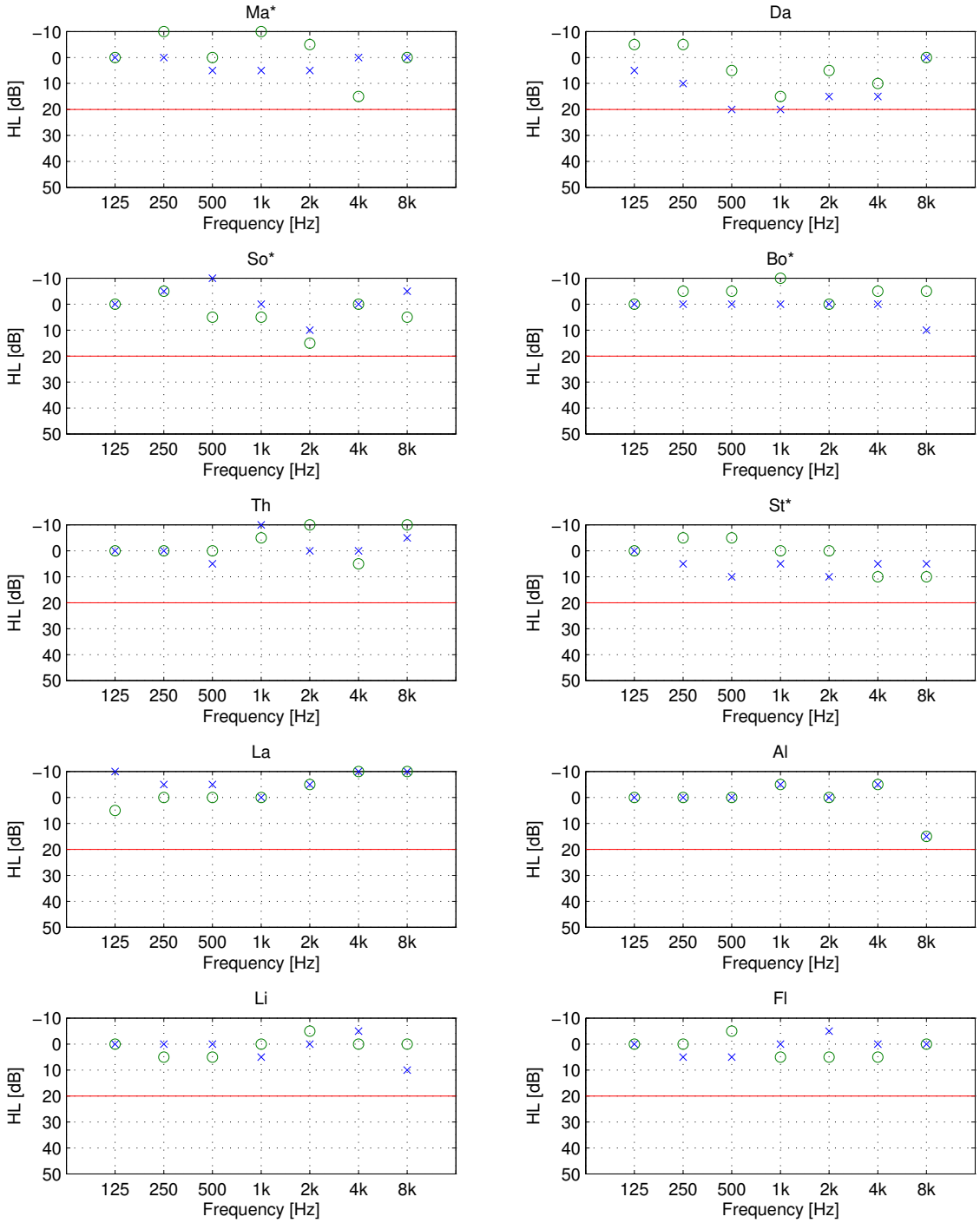



**Figure E.1:** Results of audiometry test for subjects 1-10. In each plot "×" denotes values for the left ear, "◯" denotes values for the right ear, the solid line at $20\,\mathrm{dB}$ denotes the normal hearing limit.

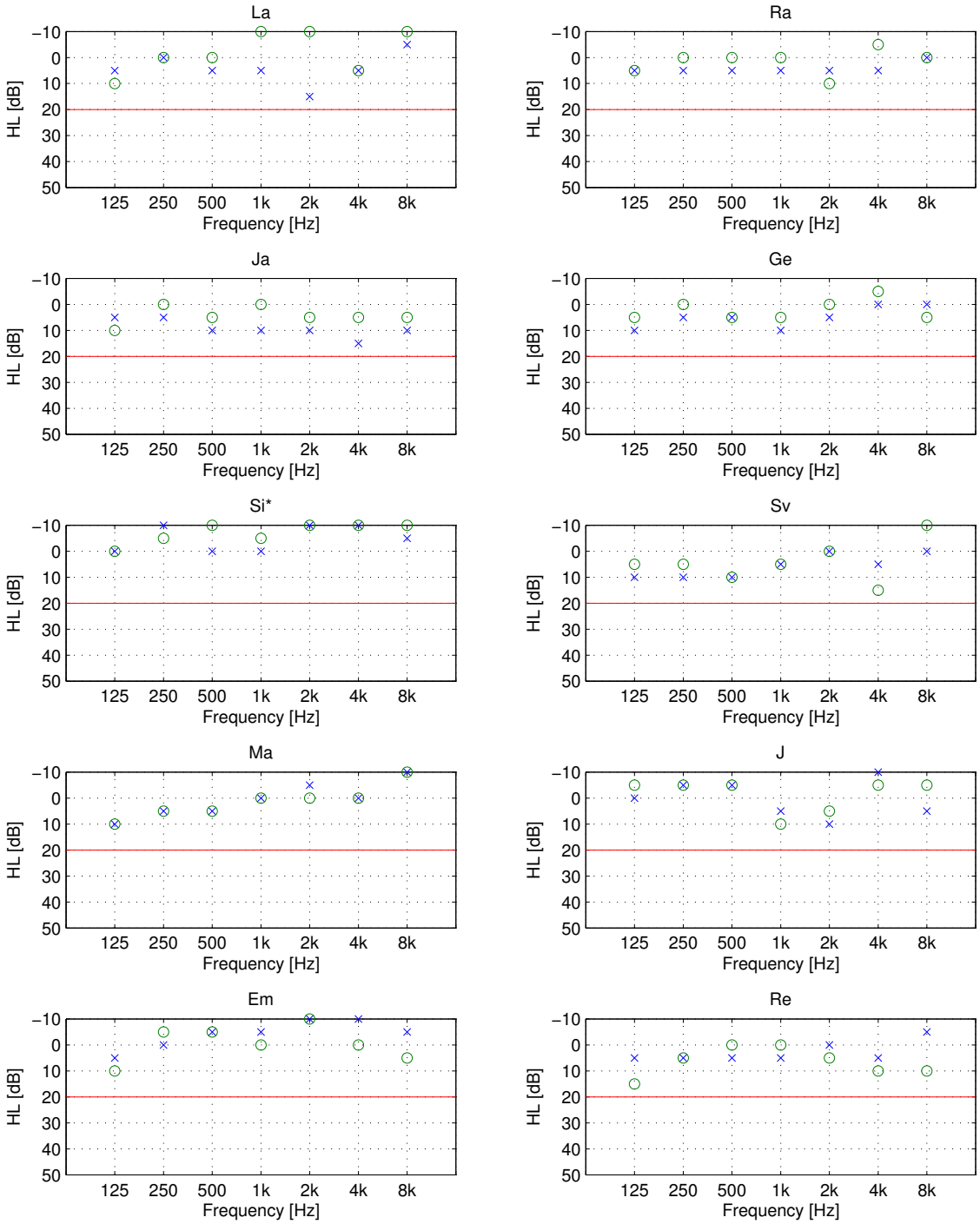



**Figure E.2:** Results of audiometry test for subjects 11-20. In each plot "×" denotes values for left ear, "◯" denotes values for right ear, the solid line at $20\,\mathrm{dB}$ denotes the normal hearing limit.


**Figure E.3:** Introduction of the samples in the MATLAB program.

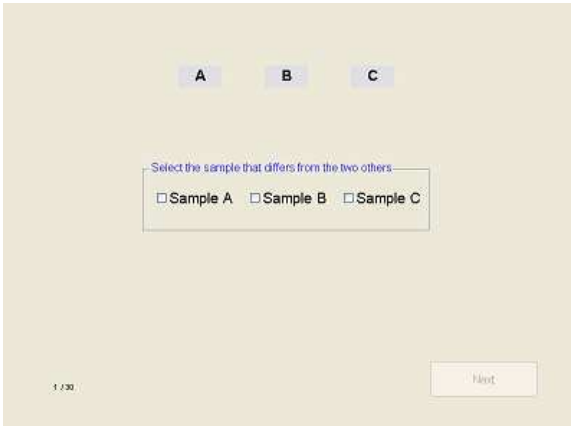



**Figure E.4:** Screenshot of the difference test.

For the preference test, the experimenter has to input the movie number in the start window. The subjects will then listen to a pair of samples and have to check the box for the preferred one. Again, the played sample will be indicated by changing the background colour of the appendant letter. After pressing "Next", this procedure will be repeated 30 times as well. A screenshot of the preference test is presented in Figure E.5. The whole preference test has to be run three times, once for every movie sample.

When the program was created, the most important practical aspect of it, besides storing the data, was, to make it "foolproof". To achieve that, all the check boxes are only enabled after the sequence of the sounds to compare has been played. Again, the "Next" button is enabled after a choice has been made. The test subject should never be able to stop the program before the demanded tasks have been finished. For that reason, the program on the touch screen had been running in full screen mode. Another reason for that was that the test subjects should not be distracted by any ongoing screen action as for example a Windows background.

The data for every user is stored in a "bin"-file for every part of the test separately, labeled after the user number and the part of the experiment (difference test, preference test 1, preference test 2, or preference test 3). Special MATLAB files have been written to read this data and provide it as a matrix to the calculations.

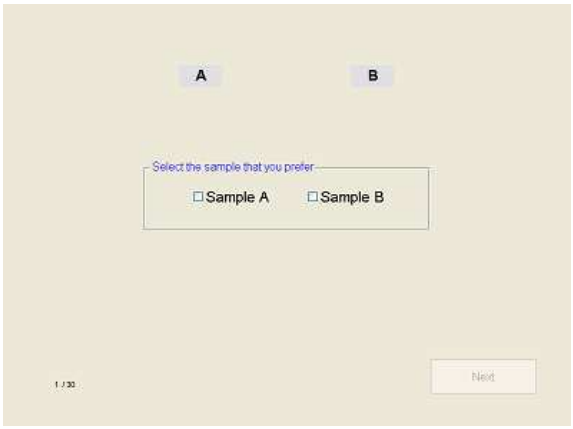



**Figure E.5:** Screenshot of the preference test.

## E.4 Setup for the listening experiment

The listening experiment took place in the listening room (B4-107) which conforms to the IEC 268-13 standard and thus is acoustically close to "average living room". This room is big enough to place the TV-screen in distance of approx. 2.5 m from the subject, corresponding to the distance between the loudspeakers and VALDEMAR during the measurements. All facilities present in the listening room during the test are an armchair, pair of headphones (1), the touch-screen (2), the intercom unit (3), and the TV-screen (4). All equipment used is depicted in the Figure E.6 and listed in the Table E.1. During the test, the subject was sitting in the armchair with the touch-screen nearby on the small table and with the headphones on the head. Armchair, touch-screen and TV-screen were all placed on the longitudinal axis of the room to prevent the subject from turning the head. The program was presented on the touch-screen and subjects were selecting their answers through this interface. The TV-screen served only for showing the movie samples and was visible at the same time with touch-screen. The experimenters could hear the subject's questions through the intercom.

The other part of the setup was placed in the control room A (B4-105). This consisted of the PC (8), where the MATLAB program was running, the digital/analog and analog/digital converter (6), the headphone amplifier (5), the monitor headphones, the KVM extender (7) and the monitor TV-screen. The PC is equipped with an ATI Radeon 8500 graphics adapter, which has two video outputs. These two video outputs were used, one (VGA out) for bringing the program interface picture to the touch-screen, the second (Composite Video out) carried only the movie picture to both the TV-screen in the listening room and the monitor TV-screen in the control room. The digital audio stream is passed through a D/A converter and then amplified and distributed into two pairs of headphones, one for the subject and one for the experimenter. The headphones used by the subjects were the ones, for which the equalization filters were made.

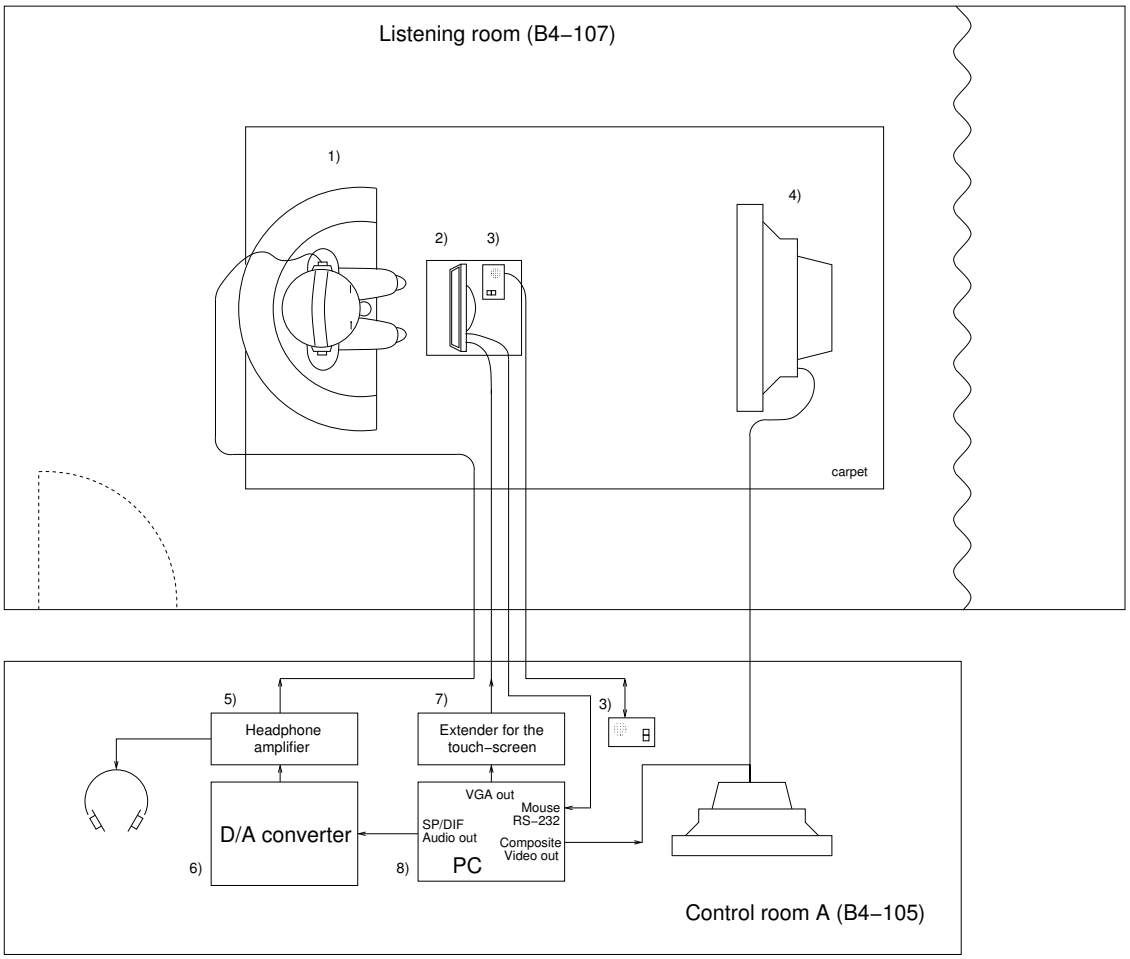



**Figure E.6:** Setup for the listening experiment.

**Table E.1:** Devices used in the listening experiment setup

| # | Device | Brand and Model | AAU no. |
|---|---|---|---|
| 1 | Headphones | beyerdynamic DT990pro | 2036–0 |
| 2 | Touch screen | Elo Touchsystems 1224L | 47234 |
| 3 | Intercom | Bouyer | 2156–03 |
| 4 | TV screen | Bang & Olufsen Avant | 52671 |
| 5 | Headphones Amplifier | Behringer PowerPlay HA 4000 | 33527–00 |
| 6 | D/A – A/D converter | Tracer BIG DAADI | 33694–00 |
| 7 | KVM Extender | Danbit KVM-Extender | 2153 –04 |
| 8 | PC | - | private |

## E.5 Instructions for the subjects

Two instruction papers which were given to subjects undergoing the listening test are inserted below. The first one "Instructions for subjects in listening test - part 1" was given to subject at the very beginning of the whole test. After the subjects were asked if they understand all tasks listed and potential questions were answered. The second paper "Instructions for subjects in listening test - part 2" was given before the preference test.

## Instructions for subjects in listening test - part 1

During the test certain movie sections will be showed on the bigger screen (TV), the sound will be played through headphones. On the PC-screen you will see the program, which will lead you through the whole testing procedure. You will also give your answers through this screen - it is touch-sensitive.

The test is divided into two main parts which are described on this and next page:

**Introduction of the used samples**
At the very beginning of the listening test 7 movie samples will be played back automatically for you. After that you can play whichever sample again by pressing one of 7 buttons. This part should make you familiar with the samples, which will be used later in the test. The picture of the movie will be always the same for all 7 samples.

**Difference test**
In the next part the listening test begins. The program will play always 3 video samples in sequence **without repeating**, so try to concentrate. Two of them are exactly the same, one is different. Your task is to decide, which of those three samples differs from the others. It is possible that sometimes the difference is not evident but anyway, you have to select one to go further. After selecting one sample press "Next" to play another triple.

To make you familiar with testing procedure, there will be a short Trial (containing both parts mentioned above) just at the beginning.

## Instructions for subjects in listening test - part 2

**Preference test**
In this part only 2 samples will play in sequence **without repeating**. Your task here is to choose that sample, which you prefer, that means which sounds better to you.

The preference test will be repeated 3 times with the tracks from different movies.

There is no introduction in this part of the test.