HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Electronics, Communications and Automation
Department of Signal Processing and Acoustics

Mikko-Ville Laitinen

Binaural Reproduction for Directional Audio Coding

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, May 26, 2008

Supervisor: Matti Karjalainen
Instructor: Ville Pulkki

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Mikko-Ville Laitinen
Name of the thesis:	Binaural Reproduction for Directional Audio Coding
Date:	May 26, 2008 **Number of pages:** vii + 66
Faculty:	Electronics, Communications and Automation
Professorship:	S-89
Supervisor:	Prof. Matti Karjalainen
Instructor:	Docent Ville Pulkki

We can hear the directions of sound sources in three dimensions, but also we perceive other spatial attributes such as the auditory sense of space. In order to reproduce spatial sound correctly, the directions of sound sources must be reproduced accurately and also the perception of space must be reproduced realistically. One recently proposed method for spatial sound reproduction is Directional Audio Coding (DirAC). It is currently implemented for loudspeaker reproduction. In this thesis it is investigated if DirAC could be implemented for headphone listening.

In DirAC analysis the direction and the diffuseness of sound are computed using B-format signals. The analysis and the synthesis are performed separately for each critical band of hearing. In DirAC synthesis sound is divided into nondiffuse and diffuse parts. The nondiffuse part is reproduced with amplitude panning. In headphone listening this was implemented by using virtual loudspeakers. Sound was positioned to the analyzed direction using vector base amplitude panning (VBAP). The virtual loudspeakers were created using head-related transfer functions (HRTF). The aim of the diffuse sound is to produce perception of surrounding sound lacking prominent direction. This was done by reproducing differently decorrelated versions of the signal with a few virtual loudspeakers. The directions of the virtual loudspeakers were chosen so that they covered the whole sphere around the listener.

In informal testing of the headphone version of DirAC, it was found that the auditory sense of space is reproduced well, and the directions of sound sources are perceived naturally. The main problem with this technique is that especially frontal sound sources are not properly externalized.

As a part of this work, a HRTF measurement system and a DirAC-based head tracking system were designed and constructed. HRTFs measured with the measurement system were used to create virtual loudspeakers. With head tracking in binaural reproduction, auditory objects can be positioned to a fixed direction even though a listener moves his/her head.

Keywords: DirAC, HRTF, binaural, spatial sound, headphones

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Mikko-Ville Laitinen
Työn nimi:	Binauraalinen toisto Directional Audio Coding -tekniikassa
Päivämäärä:	26.5.2008 **Sivuja:** vii + 66
Tiedekunta:	Elektroniikka, tietoliikenne ja automaatio
Professuuri:	S-89
Työn valvoja:	Prof. Matti Karjalainen
Työn ohjaaja:	Dos. Ville Pulkki

Ihminen kuulee äänen suunnan kolmessa ulottuvuudessa, mutta äänestä voi havaita myös muita tilaan liittyviä ominaisuuksia, kuten tilantuntu. Jotta tilaääni tulee toistetuksi oikein, äänilähteiden suunta sekä tilantuntu tulee toistaa realistisesti. Directional Audio Coding (DirAC) on eräs äskettäin esitetty menetelmä tilaäänen toistamiseen. Tällä hetkellä se on toteutettu kaiutinkuunteluun. Tässä diplomityössä tutkitaan, voitaisiinko DirAC-tekniikkaa käyttää kuulokekuuntelussa.

DirAC-analyysissä äänen suunta ja diffuusisuus lasketaan käyttämällä B-formaattisignaaleja. Analyysi ja synteesi suoritetaan taajuuskaistoittain, jotka vastaavat kuulon kriittisiä kaistoja. DirAC-synteesissä ääni jaetaan ei-diffuusiin ja diffuusiin osaan. Ei-diffuusi ääni toistetaan amplitudipanoroinnilla. Kuuloketoistossa käytettiin virtuaalisia kaiuttimia, joiden avulla ääni sijoitetaan haluttuun suuntaan käyttämällä vector base amplitude panning -tekniikkaa (VBAP). Virtuaaliset kaiuttimet toteutettiin käyttämällä head related transfer function -tekniikkaa (HRTF). Diffuusin äänen tarkoituksena on luoda havainto äänestä, joka ympäröi kuulijan eikä sen suuntaa voi havaita. Tämä toteutettiin toistamalla eri tavalla dekorreloituja signaaleja muutamilla virtuaalisilla kaiuttimilla. Virtuaalisten kaiuttimien suunta valittiin siten, että niitä oli joka puolella kuuntelijaa.

DirAC:in kuulokeversiota testattiin epäformaalisti. Havaittiin, että tilantuntu välittyy hyvin ja äänen suunta toistuu luonnollisesti. Suurin ongelma tässä tekniikassa on se, että äänilähteiden ei aina havaita olevan pään ulkopuolella, varsinkin edessä olevilla lähteillä.

Työn osana toteutettiin myös HRTF-mittausjärjestelmä ja DirAC-tekniikkaan pohjautuva päänseurantajärjestelmä. HRTF-mittausjärjestelmällä mitattuja vasteita käytettiin virtuaalisten kaiuttimien luonnissa. Päänseurannan avulla voidaan binauraalisessa toistossa pitää äänilähteet paikallaan, vaikka kuuntelija liikuttaisi päätään.

Avainsanat: DirAC, HRTF, binauraalinen, tilaääni, kuulokkeet

Acknowledgements

The research for this thesis was carried out in the Department of Signal Processing and Acoustics. The project was funded by the Emil Aaltonen foundation.

I want to thank my instructor docent Ville Pulkki for guidance and support during my research and writing process. He had lots of ideas and answers to questions. I would also like to thank my supervisor professor Matti Karjalainen. My gratitude also goes to my co-workers at the acoustics lab. Especially I would like to thank Juha Vilkamo, Jukka Ahonen and Timo Hiekkanen for cooperation and discussions.

Without these people the research work and writing the thesis would have been a lot harder and duller, so thank you!

Otaniemi, May 26, 2008

Mikko-Ville Laitinen

Contents

Abbreviations

DirAC	Directional audio coding
EDT	Early decay time
ERB	Equivalent rectangular bandwidth
FEC	Free-air equivalent coupling
FFT	Fast Fourier transform
FIR	Finite impulse response
HPTF	Headphone transfer function
HRTF	Head-related transfer function
IACC	Inter-aural cross correlation
IC	Inter-aural coherence
IIR	Infinite impulse response
ILD	Inter-aural level difference
ITD	Inter-aural time difference
PDR	Pressure division ratio
PTF	Headphone transfer function
SIRR	Spatial impulse response rendering
SPL	Sound pressure level
STFT	Short-time Fourier transform
VBAP	Vector base amplitude panning

Chapter 1

Introduction

1.1 General Discussion

We use directional hearing in our everyday life. Together with vision it is used to decide the information of where different objects or events are situated. For example in traffic we can hear if a car is coming from behind us or we can call out to a friend and he/she can spot us.

Hearing works in three dimensions. The direction of sound in left-right direction can be detected quite precisely, but also if the sound is coming from the front or behind or above or below us. Also spatial aspects can be perceived, whether we are in a small or large room or if the space is echoic or not. Let us consider a rock concert. It sounds completely different if it is in a stadium or in a small bar.

So there is a need for a technique that can store a recording of the sound in a certain space including also directional information and spatial aspects. This recording should be reproduced using loudspeakers or headphones. A sensation of being in that space should be reproduced.

The most used audio format nowadays is two-channel stereophonic reproduction. Typical stereophonic reproduction uses two loudspeakers that are placed in front of the listener and the spacing between them is 60 degrees. With this kind of system sound sources are in the horizontal plane. The sound sources are in front of the listener, between loudspeakers. If headphones are used, sound is localized inside the head of the listeners. Sound can be panned from the left ear to the right ear, but the sound remains inside the head.

With many applications two-channel stereophonic reproduction is sufficient. It is used for example in the reproduction of music and in the television. For many people this kind of system is enough. The most used multichannel system in the home audio is 5.1 channel surround that has three loudspeakers in front of the listener and two on the sides of the listener and one channel for low-frequency reproduction. A better perception of space is achieved using multi-channel reproduction. 5.1 channel surround systems have loudspeakers also behind the listener, so in theory a sound source can be positioned anywhere around the listener in the horizontal plane.

1.2 Aim of the Thesis

A recently proposed method for storing and reproducing spatial sound is Directional Audio Coding (DirAC) [1]. It is currently implemented for loudspeakers. In this thesis it is investigated if DirAC could be implemented for headphone listening.

The ultimate goal is that one can record and store sound in some space and later, using headphones, he/she can perceive the sound equally as in the original space, see Fig. 1.1. The recording is made using suitable microphone techniques and the reproduction using conventional headphones. One should be able to localize sounds in three dimensions. Also the feel of the space should be reproduced, for example if sound is recorded in a small or a large room or outdoors. This should be done without changing the timbre of the sound.

Figure 1.1: On the left a symphony orchestra is playing in a concert hall. This concert is recorded using a B-format microphone. This recording is DirAC-encoded and -decoded using binaural reproduction of DirAC. Later a listener is listening to this concert at home using headphones and he feels as he would have been in the concert for real. Modified from [2] and [3].

Possible purpose of use could be for example to record a rock or symphony orchestra concert and to reproduce it naturally using headphones. Also it could be used to get multi-channel reproduction of movies and television into mobile devices. DirAC could also be used in real-time, for example in teleconferencing. With it different talkers could be positioned to different directions and the intelligibility of speech would be increased. Also naturalness of the conference would be increased because the space is also reproduced.

DirAC could also be used for "being there". The idea is that one should at least be able to see and hear what happens in another space. Video image would be reproduced with suitable technique and DirAC would be used to reproduce sound. For example vehicles that operate in dangerous situations could be steered with a help of remote being.

1.3 Outline of the Thesis

The theoretical background for the project is presented in Ch. 2. It starts with the characteristics of hearing that are important to binaural DirAC. Basics of spatial hearing are explained a bit more carefully. DirAC and binaural reproduction are based on principles presented there. The theory of head-related transfer functions is explained in the beginning of Ch. 3. In the end of the chapter it is explained how they were measured in practice and implemented in this project. Principles of DirAC are presented in Ch. 4. Only the implementation for loudspeakers is considered there. It functions as a background information about DirAC for the binaural implementation.

In Ch. 5 it is presented how binaural reproduction of DirAC was implemented. It also includes what parameters were chosen and how different parts of DirAC were implemented in detail. Results of the project are presented in Ch. 6. The quality of binaural reproduction for DirAC is considered there. Ch. 7 includes conclusions about the binaural DirAC and discussion what could be done in the future to make it work better.

Chapter 2

Background

Sound can mean a physical phenomenon or a subjective auditory sensation. In this chapter it is briefly explained what sound is as a physical phenomenon. The human hearing system is very complex. The structure of the human hearing organ and some principles of hearing are briefly explained in this chapter. Spatial hearing is covered in more detail. Emphasis is in the properties that are important to DirAC.

2.1 Sound as Physical Phenomenon

As a physical phenomenon sound can be described as a disturbance in a medium that can cause an auditory sensation [4]. As a physical phenomenon sound does not include actual auditory sensation. In air or in other fluids sound consists of longitudinal wave motion of particles of fluid. In solid materials sound can also move as transverse waves.

Sound consists of waves that can have arbitrary frequencies. Waves with different frequencies have different wavelengths, calculated from

$$\lambda = \frac{c}{f} \tag{2.1}$$

where λ is wavelength, c is the speed of sound and f is frequency. The speed of sound in air is about 343 m/s [4].

Sound can be presented using two quantities: pressure and particle velocity. The pressure of sound is a scalar quantity. Our hearing system is able to hear sound pressure. Particle velocity is a vector quantity. It stands for the velocity of a particle in a medium. This should not be confused with the speed of sound which means the speed of the sound waves. Actual particles do not move at this speed. Instantaneous intensity is defined as the product of the sound pressure $p(t)$ and the particle velocity vector $\boldsymbol{u}(t)$ [5]

$$\boldsymbol{I}(t) = p(t)\boldsymbol{u}(t). \tag{2.2}$$

The instantaneous energy density can be calculated as

$$E(t) = \frac{1}{2}\rho\left[\frac{p^2(t)}{Z^2} + \boldsymbol{u}^2(t)\right] \tag{2.3}$$

where ρ is the density of the medium and Z is the acoustical impedance of the medium [6].

The behavior of sound pressure in a fluid can be presented using the general wave equation

$$\frac{\partial^2 p}{\partial t^2} = c^2 \frac{\partial^2 p}{\partial x^2} \tag{2.4}$$

where p is pressure, x position and t is time [7]. This equation gives sound pressure in one-dimensional case at any given location and time instant.

2.2 Sound in Rooms

Sound in a room can be divided into three parts: direct sound, early reflections and late reverberation [4], see Fig. 2.1. Direct sound is sound coming straight from the source. Early reflections are first few reflections from a floor, ceiling or walls. The number of reflections is propositional to t^2, where t is time [8]. After a short time the number of the reflections is so huge that they cannot be considered one by one, but instead are considered as late reverberation.

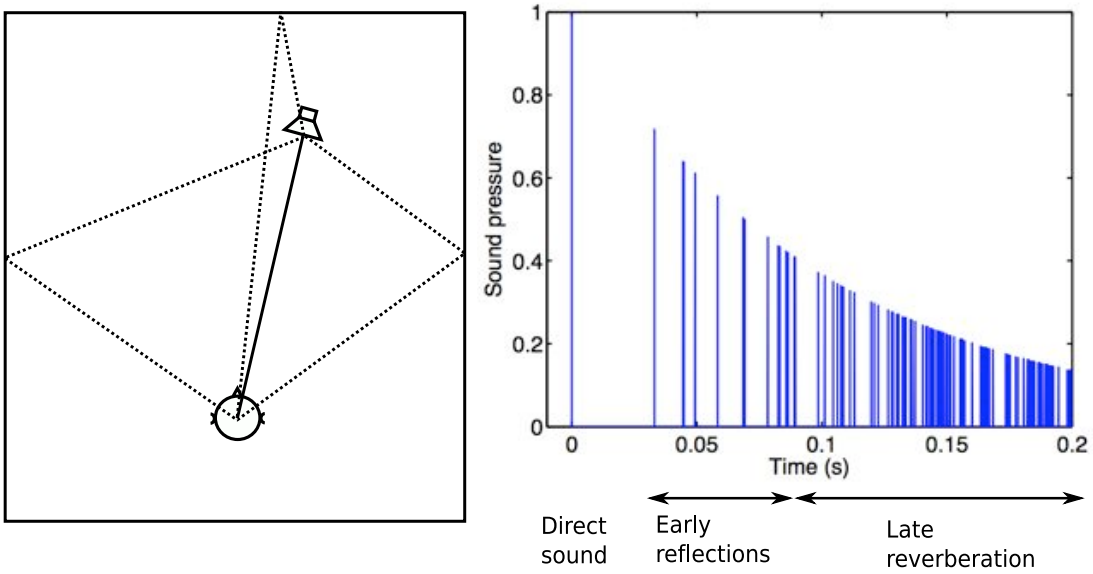

Figure 2.1: On the left: paths of the direct sound and few early reflections in a room. On the right: a theoretical example of an impulse response. Direct sound, early reflections and late reverberation are separated in the figure. In reality decaying is not so smooth.

Late reverberation consists of multiple reflections in a room. Sound field is so complex in late reverberation that it typically is considered statistically. Reverberant sound is at least partially

diffuse. Diffuse field is defined as a field that consists of an infinite amount of evenly spaced plane waves from every direction that do not have correlation between them [9]. Energy is not flowing to any direction and thus the intensity is zero. Diffuse field is a theoretical ideal, but it can be used to model reverberation in a room. The simplest description of late reverberation is to measure reverberation time. It is defined as that time required for the sound in a room to decay 60 dB [10]. Reverberation time, and especially a similar description called early decay time (EDT), corresponds to our perception of reverberation in a temporal sense [11]. Reverberation time depends on frequency. There are also other measures to describe reverberation but reverberation time is the simplest and the most used.

2.3 Structure of Ear

The human ear can be divided into three parts: external ear, middle ear and internal ear [12]. The structure of the ear can be seen in Fig. 2.2.

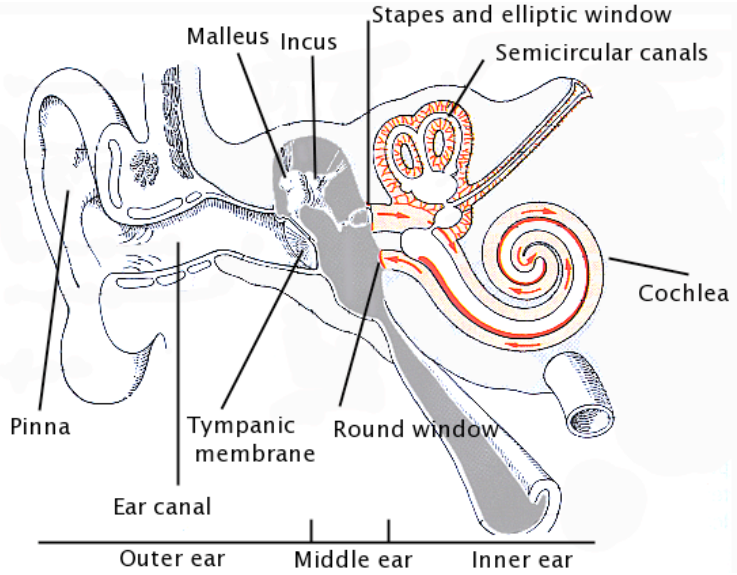

Figure 2.2: The structure of the ear. [13]

The external ear consists of a pinna, ear canal and tympanic membrane. The pinna is the visible part of the ear and it is called the ear in everyday language. It helps in the localization of sound objects. This will be covered more carefully later in this thesis. Sound enters through the ear canal. The tympanic membrane is in the end of it. Pressure fluctuations propagated through the ear canal set the membrane in motion. The external ear is a passive and linear system. It only transfers sound to the middle ear and filters it [14].

The main function of the middle ear is to transmit the sound from the tympanic membrane to the oval window. It performs an impedance matching between air in the external ear and the

fluid of the internal ear. The conduction of sound is made with an ossicular chain which consists of three small bones: the malleus, incus and stapes. The ossicular chain works as a mechanical impedance matcher.

The main part of the internal ear is the cochlea, which is a shell-shaped organ. It is filled with fluid. A straightened version of it can be seen in Fig. 2.3. Straightening the cochlea is impossible in a real life, but parts of it can be seen more clearly this way. The stapes makes the fluid in the cochlea vibrate. This also vibrates the basilar membrane. On the basilar membrane is the organ of Corti, which includes several rows of hair cells. Above the hair cells is the tectorial membrane. Each hair cell has many hairs, which are in contact with the tectorial membrane. When the basilar membrane vibrates there is a difference in motion between the basilar and tectorial membrane. This causes that hairs in the hair cells bend. Bending of the hairs causes hair cells to send impulses to the auditory nerve. The whole system is very complex but as a simplification these impulses are understood as sound by the brain.

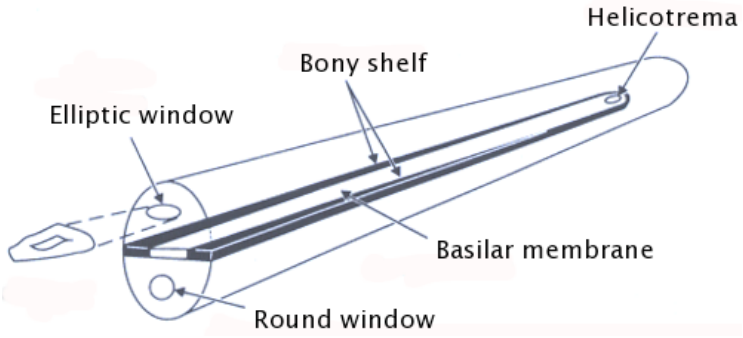

Figure 2.3: Theoretical view of straightened cochlea. [14]

The basilar membrane is a long membrane that has different width, mass and flexibility properties in the different parts of it. That is why different frequencies resonate at different positions. Békésy [15] measured the displacement of the basilar membrane as a function of frequency and noticed that high frequencies cause displacement near the oval window and low frequencies at the other end of the basilar membrane, see Fig. 2.4. This causes that sounds with different frequencies bend different hair cells. Based on this people can perceive the timbre and the pitch of the sound.

2.4 Basic Properties of Hearing

2.4.1 Frequency Range of Hearing

Humans have a frequency range of hearing of about 20 - 20000 Hz. The loudness we perceive is different from the sound pressure level. Loudness is a function of frequency and sound pressure. Hearing is most sensitive at mid frequencies and less sensitive at low and high frequencies.

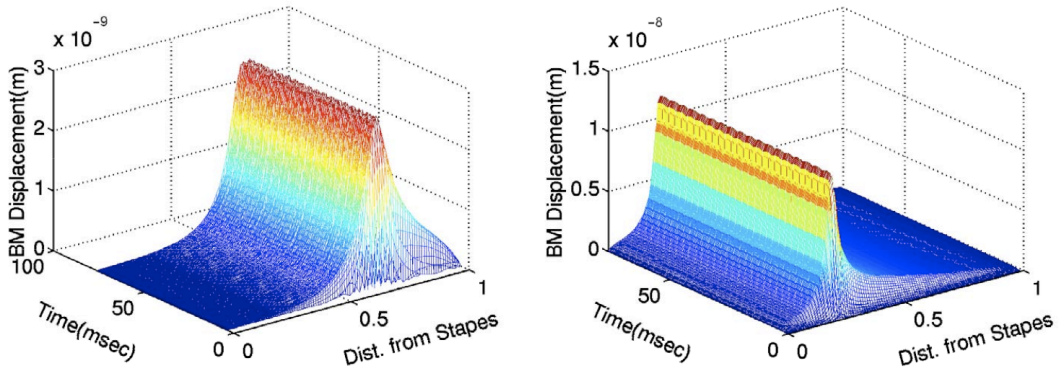

Figure 2.4: Basilar membrane displacement of 600 Hz (left) and 6 kHz (right) 30 dB tones. [16]

Equal-loudness curves can be seen in Fig. 2.5. Lines represent constant loudness over frequency. [4]

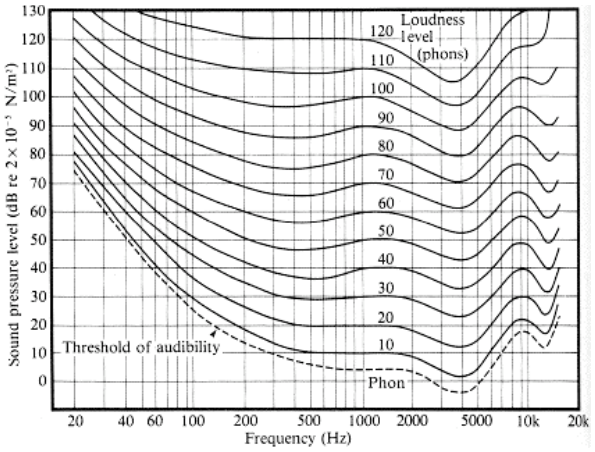

Figure 2.5: Equal loudness curves. [4]

2.4.2 Critical Bands

When two sinusoidal tones are close in frequency, their amplitude envelopes overlap on the basilar membrane. These kind of tones are in the same critical band [4]. Hearing handles broadband sound in a way that partial sounds in the same critical band are analyzed as one entity [14]. According to the Bark-scale, at low frequencies a critical band is about 100 Hz and at mid and high frequencies it is about one third octave.

Critical bands can be measured by using the following method. A test subject listens to one narrowband noise which is a reference sound. He/she also listens to another noise. The subject can change the bandwidth of that noise while sound pressure level is kept as a constant. When

the subject increases the bandwidth of the noise, loudness remains as a constant at first. After a certain value loudness starts to increase. This value tells the width of the critical band at that frequency. Critical bands obtained using this method correspond to the Bark-scale.

A similar way to get critical bands is to use the concept of equivalent rectangular bandwidth (ERB) [17]. It is measured using noise masking around the test signal. Masking noise acts as an acoustical filter. The passband of the acoustical filter can be changed in order to determine how large part of the test signal we can hear [14]. Using this method the width of the ERB-bands in different frequencies can be obtained.

The width of the ERB-band can be calculated from

$$\text{ERB} = 24.7 + 0.108 f_c \tag{2.5}$$

where f_c is the middle frequency of the ERB-band. From Eq. (2.5) an approximate for ERB-band number can be calculated

$$R_{\text{ERB}} = 21.3 \log_{10}(1 + f/228.7\text{Hz}) \tag{2.6}$$

where R_{ERB} is the number of the ERB-band corresponding to frequency f.

The simplest estimate of the critical bands is to use one third octave bands even though it is not very accurate.

2.4.3 Masking

When there are two or more simultaneous tones, one may mask others. This means that the threshold of hearing for other tones has increased. Masking is caused by overlapping displacement of the basilar membrane, see Fig. 2.4. For example if one is in a noisy space, people need to talk louder that one may hear it. If masking tone is narrowband, it masks more tones that are closer to it in frequency. Tones higher in frequency compared to masker tone are masked more effectively than tones lower in frequency. The more intensity masking tone has, the broader the range of masking is. Broadband noise masks all frequencies. Increasing the intensity of the broadband noise increases masking. [4]

2.4.4 Timbre

Timbre is used to denote tone colour or quality of the sound. Timbre is defined by the American National Standards Institute [18] "Timbre is that attribute of auditory sensation in terms of which a listener can judge two sounds similarly presented and having same loudness and pitch as dissimilar." The closest physical analogue is spectrum [14]. However physical spectrum differs clearly from the spectrum calculated by our hearing system and also from the sensation of timbre.

2.4.5 Temporal Accuracy

Typically the auditory system is described to analyze sound in the frequency domain and that temporal properties are less important. Anyhow, temporal properties have some effect if they are significant enough.

Green [19] studied temporal acuity by playing two transient signals that have identical energy spectra but different phase spectra. When transients were shortened, a point was reached where they could not be discriminated. This point corresponds to the fastest variation in the pressure waveform that ear can follow. Green suggested that this value could also be the minimum time constant of the auditory system. He came to conclusion that this value is about 2 ms.

Yost and Moore [20] studied temporal changes in spectrum. They played flat-spectrum stationary noise and noise with varying spectral properties. Listeners were able to discriminate the samples if temporal modulation was less than 10 cycles per second. This suggests that the temporal accuracy for changing spectrum is 100 ms.

Goldberg [21] studied the audibility of temporal decay at low frequencies, between 32 and 200 Hz. He found that the lower the frequency is, the longer the decay must be in order to be heard. At 200 Hz the threshold of decay time was about 0.1 s and at 32 Hz more than 1 s.

As a conclusion it can be said that temporal accuracy depends on the situation. There is not one time constant that would describe the temporal accuracy. Examples presented here give a few views of how hearing works in time.

2.5 Spatial Hearing

People are able to hear the direction of sound in three dimensions. Directional hearing is based on three cues: inter-aural time difference (ITD), inter-aural level difference (ILD) and monaural cues [22].

2.5.1 ITD and ILD

Let us assume that there is a sound source 45 degrees right from the listener in the horizontal plane, see Fig. 2.6. It can be seen that sound comes first to the right ear and after a delay to the left ear. So there is a time difference between ears, called inter-aural time difference. Sound is typically modeled to travel around the head. If the head is assumed to be sphere-shaped, approximate ITD can be calculated from

$$\tau_{\mathrm{ITD}} = \frac{D}{2c}(\varphi + \sin\varphi) \tag{2.7}$$

where D is the diameter of the head, c the speed of sound and φ the angle of the sound source [23]. Maximum ITD is measured when the sound source is in the side. It can be as large as $800\ \mu$s depending on the size of the head. ITD is quite independent of frequency. Finer results are presented for example in [24].

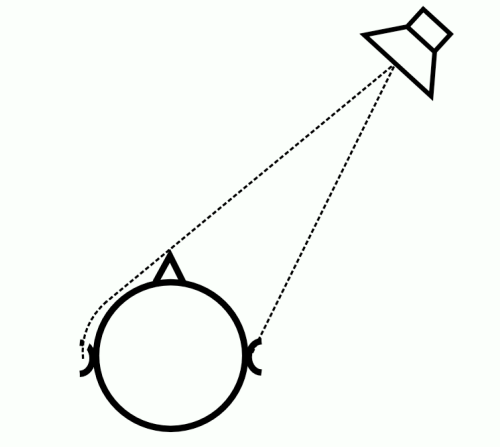

Figure 2.6: Sound source 45 degrees right from the listener.

Sound comes delayed to another ear if the sound source is on the side but the head also attenuates the sound. The sound on the contra lateral ear is attenuated compared to the ipsi lateral ear. This causes inter-aural level difference. ILD is depended on the frequency. At low frequencies the size of the head is small compared to wavelength. Therefore the head causes only minor attenuation. At high frequencies the head is large compared to wavelength and ILD is greater. Also the shape of the outer ears influences ILD at high frequencies.

Traditionally, localization in the left-right direction has been presented as a duplex theory. ILD does not work at low frequencies because there is not enough difference in the sound pressure level between ears. The auditory system is calculating ITD based on phase differences. At high frequencies ITD is larger than the cycle time so there is no unequivocal relationship between the delay and the phase. Based on this it was concluded that the direction of sound is calculated from ITD below 1500 Hz and from ILD above it. Later it has been discovered that ITD also works above 1500 Hz. However the duplex theory usually gives results that are close enough to reality. [22]

2.5.2 Monaural Cues

Using ITD and ILD it can be decided if sound is coming from left or right. Other directions cannot be decided because for example if the sound source is in front of the listener or behind the listener ITDs and ILDs are roughly the same. So called cones of confusion can be calculated. In the cone of confusion ITD and ILD remain the same, see Fig. 2.7. So we need other cues to decide the elevation angle of the sound source.

When sound comes to our ear, the head, body and external ear filter the sound. This is caused by the reflections for example from the ear or shoulders. Some frequencies are boosted and some frequencies are attenuated. This filtering is depended on the elevation angle. Based on this people have learned that certain filtering of sound means that it is coming from certain direction.

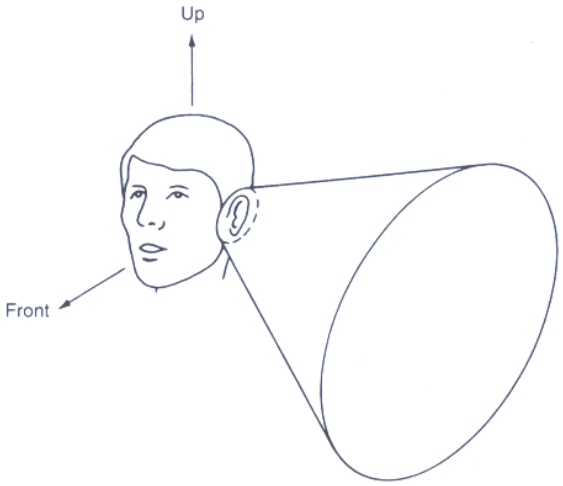

Figure 2.7: Cone of confusion. Inter-aural cues are the same within this cone. [17]

Changes in spectrum in changed elevation angle are called monaural cues.

2.5.3 Interaural Coherence Cues

ITD, ILD and monaural cues determine the direction of the auditory event. However, we are also able to perceive spatial attributes such as reverberation. Interaural coherence cues (IC) are used for the perception of space.

ITD is calculated by the auditory system from phase differences. The first model for calculating ITD was proposed by Jeffress [25]. In it there are delay lines that are connected by several coincidence detectors [22]. Signal from the left ear is summed with the delayed signal from the right ear and vice versa. There are multiple delays, see Fig. 2.8. This model can be considered as calculating cross-correlation between signals from different ears. It is called inter-aural cross-correlation (IACC).

The effect of interaural coherence was studied in [26]. Two signals were reproduced dichotically using headphones. The amount of coherence, k, between signals was changed. k is defined as the maximum absolute value of the normalized cross-correlation function of two signals $x(t)$ and $y(t)$, or

$$k = \max_{\tau} |\Phi_{xy}(\tau)|. \tag{2.8}$$

If $k = 1$ there is one precisely localized, spatially compact auditory event, see Fig. 2.9. When k degreases, the auditory event gains spatial extent, until at about $k = 0.4$. Then it fills the whole head or even the space around it. For smaller k the auditory event starts to split into two components, one at each ear.

In [27] it was stated that the diffuseness of sound transforms into IC cues. Let us consider few examples. In a free space there are only few reflections and coherence between signals to each ear is strong. Sound is perceived to mainly consist of the direct sound. In an echoic room there

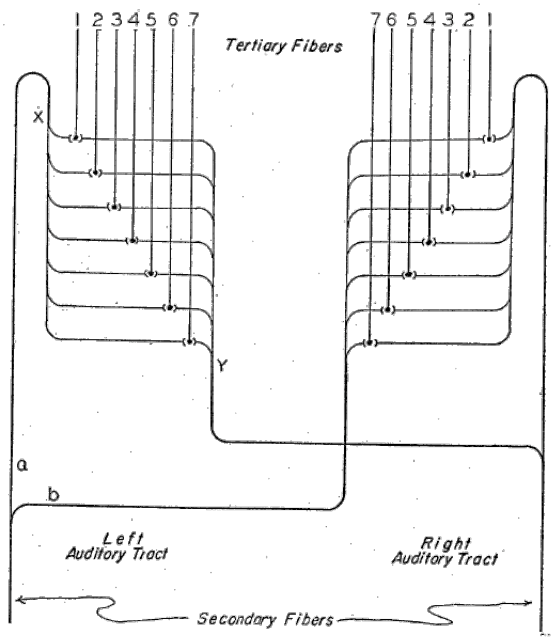

Figure 2.8: Hypothetical mid-brain mechanism for the localization of low-frequency tones [25].

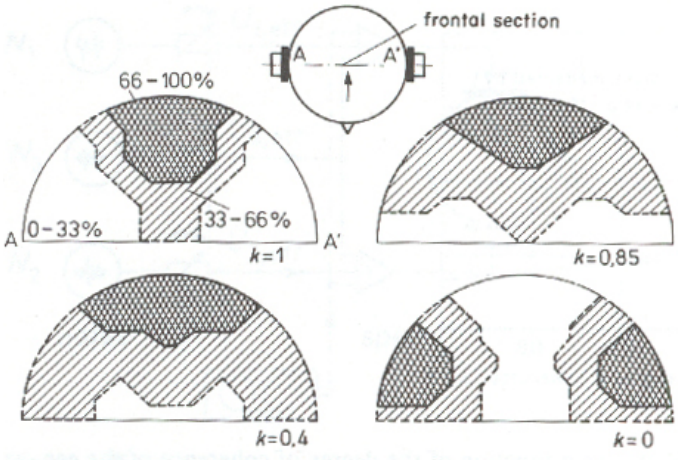

Figure 2.9: Position of auditory events projected onto a frontal section of the upper half of the head for broadband noise of various degrees of interaural coherence presented over headphones. The differently shaded areas indicate regions in which components of auditory events were reported in $0-33\%$, $33-66\%$ and $66-100\%$ of the cases. k stands for the coherence of input signals. [23]

are multiple reflections and sound field is mainly diffuse, so coherence between signals to each ear is weaker and sound is perceived to be reverberant.

There are multiple models about how these interaural coherence cues are calculated. In [27] it was presented that IC cues would be calculated using cross-correlation. For example in [28] it was suggested that auditory source width is increased with fluctuation of ITD. This could indicate that IC cues would be calculated from ITD and ILD fluctuation.

2.5.4 Precedence Effect and Perception of Reflected Sound

Sound in a room consists of a direct sound and reflections. The direct sound comes from the direction of the sound source but the reflections come from the walls and the ceiling and the floor etc. The directions of the reflections are different than the direction of the direct sound. However we are able to hear that for example speech comes from one certain direction instead of multiple directions.

Our hearing mechanism is able to gather reflections of about 50 ms after the direct sound and to give an impression that all this sound is coming from the direction of the direct sound. This is called the precedence effect. The direct sound comes the shortest path so it is always first in our hearing and the directions of the reflections are integrated to it. The precedence effect works best $20 - 30$ ms after the direct sound. A reflection is heard as a discrete echo if it comes $50 - 80$ ms after the direct sound. Reflections less than 1 ms after the direct sound contribute to the direction of the sound. After that reflections do not contribute to the direction in general.

If the level of the reflection is large enough, for example 10 dB louder than the direct sound for 20 ms delayed sound, it can be heard as a discrete echo even if the delay is shorter than 50 ms. Reflections can also add spaciousness to sound or broad image. In Fig. 2.10 it is presented how reflections with different delays and levels contribute to sound. [10]

2.5.5 Distance of Auditory Event

As mentioned before, we are able to hear the direction of sound in three dimensions, but we are also able hear how far the sound source is. In [29] four cues for auditory distance judgement are presented :

1. Sound pressure level (the greater the SPL, the shorter is the judged distance).
2. The amount of reverberation (the greater the ratio of direct to reverberant energy in the received signal, the shorter is the judged distance).
3. Spectral shape of the received signal (the greater the high-frequency content of the stimulus, the shorter is the judged distance).
4. Binaural cues (for sources off midline and closer than about 1 m, the greater the ITD or ILD, the shorter is the judged distance).

All these cues correspond to everyday situations. If a sound source is further away it is more silent. In a room distant sound sources are typically reverberant because the direct sound has

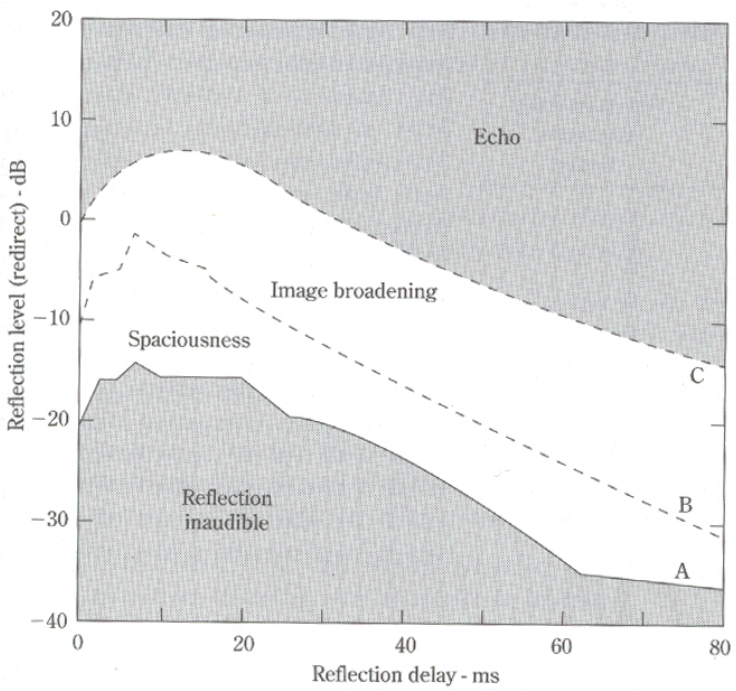

Figure 2.10: The effects of lateral reflections on the perception of direct sound in a simulated stereo arrangement. [10]

attenuated and the predominant energy is in the diffuse field of the room. Also high frequencies attenuate faster than low frequencies. Hence distant sound sources have more energy in low frequencies.

The problem is that these cues are not unambiguous. For example a distant sound source can be really loud and have more energy at high frequencies. It should be judged to be close even though it really is far away. Familiarity of the target helps in the perception of the distance. For example we know how loud approximately people talk and if we know the talker we know even more accurately how loud he/she talks. Based on that information we can decide how far the talker is based on SPL we receive.

Chapter 3

Head-Related Transfer Functions

People have two ears. Using these two ears people are able to hear directions in three dimensions based on principles presented in Section 2.5. Sound pressure at each eardrum includes all the information needed about direction. So if this pressure can be measured and reproduced, the result is directional sound in three dimensions.

According to [30] a head-related transfer function (HRTF) is a transfer function that, for a certain angle of incidence, describes the sound transmission from a free field to a point in the ear canal of a human subject. HRTFs are different for different ears. So basically sound can be filtered using an HRTF for certain direction. If this filtered sound is reproduced into the ear canals the result is that sound appears to be coming from the desired direction. Timbre should be the same as if sound would come from that direction for real.

3.1 Theory of HRTFs

Sound pressure sets the eardrum in motion. This motion is transformed into perception of sound in our hearing system. Knowing the motion of the eardrum includes all the information about the sounds we hear. Getting a microphone to the eardrum can be quite hard or even dangerous and also getting a loudspeaker (a headphone in this case) to the eardrum is at least as hard. That is why HRTFs are typically measured at the entrance of the ear canal.

The direction-dependent part of sound transmission consists of the transmission from a free field source to the sound pressure at the entrance of the ear canal. Sound transmission from the entrance of the ear canal to the eardrum is direction-independent. Thus measuring the HRTF at the entrance of the ear canal includes all spatial information. [30]

HRTFs can be measured at the entrance of the ear canal using two different approaches: blocked ear canal or open ear canal. The ear canal can be blocked using for example an earplug. Then the effect of the ear canal is blocked from the measurement. Using this approach we get sound pressure at the entrance of the ear canal. When HRTF-filtered sound is reproduced, it is likely reproduced using headphones. Then sound is inserted to the entrance of the ear canal or

few centimeters away from it. The effect of the the ear canal is inserted to sound at this stage.

If HRTFs are measured using open ear canal, the effect of the ear canal should be removed from the HRTFs. Otherwise it is present in the listening two times (first time in the measurement and second time during the playback). Using both methods good results can be obtained. Anyhow, in most of the publications the blocked ear canal is supported, for example in [30]. The reason for this is that results obtained using the blocked ear canal are more consistent. A disadvantage of the blocked ear canal measurement is that it needs compensating the difference in the acoustic source impedance when listening using headphones. More about this in Section 3.1.1.

3.1.1 Obtaining HRTFs from Measurements Using Blocked Ear Canals

Measured head-related transfer functions consist of

$$Y_{\text{meas}}(f) = L(f)H_{\text{HRTF}}(f)M_{\text{min}}(f) \tag{3.1}$$

where $Y_{\text{meas}}(f)$ is the Fourier-transform of the output signal of the microphone, $L(f)$ the frequency response of the loudspeaker, $M_{\text{min}}(f)$ the frequency response of the miniature microphone and $H_{\text{HRTF}}(f)$ the actual HRTF, in this case a mapping from the free field to the entrance of the blocked ear canal.

The frequency response of the loudspeaker can be measured using a high-quality microphone

$$Y_{\text{spkr}}(f) = L(f)M_{\text{ref}}(f) \tag{3.2}$$

where $Y(f)_{\text{spkr}}$ is the Fourier-transform of the output signal of the microphone in this measurement and $M_{\text{ref}}(f)$ the frequency response of the high-quality microphone.

Headphones should also be equalized. They can be measured in the same way as HRTFs but having headphones positioned in the same way as in the listening and using them as a source. These responses are often called headphone transfer functions (HPTF or PTF). Then

$$Y_{\text{phone}}(f) = P(f)H_{\text{PTF}}(f)M_{\text{min}}(f) \tag{3.3}$$

where $Y(f)_{\text{phone}}$ is the Fourier-transform of the output signal of the microphone, $P(f)$ the frequency response of the headphones, H_{PTF} the transfer function from the headphones to the entrance of the blocked ear canal.

Listening to a sound sample using headphones causes a signal

$$Y_{\text{eardrum}}(f) = P(f)H_{\text{PTF}}(f)H_{\text{B-O-phone}}(f)H_{\text{earcanal}}(f)X(f) \tag{3.4}$$

where $Y(f)_{\text{eardrum}}$ is sound pressure at the eardrum, $H_{\text{B-O-phone}}$ is a mapping from blocked ear canal to open ear canal when using headphones, $H_{\text{earcanal}}(f)$ the transfer function of the ear canal from the entrance of the ear canal to the eardrum and $X(f)$ the actual signal.

The mapping from the pressure at the entrance of the blocked ear canal to the pressure at the entrance of the open ear canal can be different when listening using headphones than at free field. This is caused by the difference in the acoustic source impedance. This can be compensated using so called pressure division ratios (PDR) [31]. They can be obtained by measuring pressure at the entrance of the blocked and open ear canal using headphones and a loudspeaker. PDR is defined as

$$\text{PDR} = \frac{P_{\text{open-free}}/P_{\text{blocked-free}}}{P_{\text{open-phone}}/P_{\text{blocked-phone}}} = \frac{H_{\text{B-O-free}}(f)}{H_{\text{B-O-phone}}(f)} \tag{3.5}$$

where P_{X} is pressure. Subscript free means that pressure was measured in free air, in this case using a loudspeaker. Subscript phone means that pressure was measured having headphones on the ears, the source was headphones. If PDR is close to unity, the headphone is called an FEC headphone. This means that the headphone has Free-air Equivalent Coupling to the ear. Then no PDR compensation is needed.

If we combine all these and assume that the frequency response of the high-quality microphone used to measure the frequency response of the loudspeaker is 1 we get

$$\begin{aligned}
&\frac{Y_{\text{meas}}(f)Y_{\text{eardrum}}(f)\text{PDR}(f)}{Y_{\text{spkr}}(f)Y_{\text{phone}}(f)} \\
&= \frac{L(f)H_{\text{HRTF}}(f)M_{\text{min}}(f)P(f)H_{\text{PTF}}(f)H_{\text{B-O-phone}}(f)H_{\text{earcanal}}(f)X(f)H_{\text{B-O-free}}(f)}{L(f)M_{\text{ref}}(f)P(f)H_{\text{PTF}}(f)M_{\text{min}}(f)H_{\text{B-O-phone}}(f)} \\
&= H_{\text{HRTF}}(f)H_{\text{B-O-free}}(f)H_{\text{earcanal}}(f)X(f).
\end{aligned} \tag{3.6}$$

This is the same transfer function as if a sound source would really be in that direction. This way perfect reconstruction of the sound pressure in the eardrum can be reproduced, at least in theory. In reality there always is some error and approximation. [32]

3.1.2 Obtaining HRTFs from Measurements Using Open Ear Canals

Using open ear canals HRTFs are gained in a similar way as with closed ear canals. The difference is in the lack of need for PDR compensation and that the ear canal resonance is present in the measurements. Measured head-related transfer functions consist of

$$Y_{\text{meas}}(f) = L(f)H_{\text{HRTF}}(f)R_{\text{earcanal}}(f)M_{\text{min}}(f) \tag{3.7}$$

where $R_{\text{earcanal}}(f)$ is the resonance of the ear canal and $H_{\text{HRTF}}(f)$ the actual HRTF, in this case a mapping from free field to the entrance of the open ear canal. Other variables are the same as with blocked ear canal.

With open ear canal measurement PTF is thought as a transfer function from the headphones to the entrance of the open ear canal. PTFs also have ear canal resonance. The measured response is

$$Y_{\text{phone}}(f) = P(f)H_{\text{PTF}}(f)R_{\text{earcanal}}(f)M_{\text{min}}(f). \tag{3.8}$$

Listening using headphones causes a signal

$$Y_{\text{eardrum}}(f) = P(f)H_{\text{PTF}}(f)H_{\text{earcanal}}(f)X(f). \tag{3.9}$$

Mapping from blocked ear canal to open ear canal $H_{\text{B-O-phone}}$ is not present this time because PTF was defined to be a transfer function to the entrance of the open ear canal. PDR compensation is not needed this time at all because all responses were measured using open ear canal.

If we combine all these and assume that the frequency response of the high-quality microphone used to measure the frequency response of the loudspeaker is 1 we get

$$\frac{Y_{\text{meas}}(f)Y_{\text{eardrum}}(f)}{Y_{\text{spkr}}(f)Y_{\text{phone}}(f)}$$

$$= \frac{L(f)H_{\text{HRTF}}(f)M_{\text{min}}(f)P(f)H_{\text{PTF}}(f)R_{\text{earcanal}}(f)H_{\text{earcanal}}(f)X(f)}{L(f)M_{\text{ref}}(f)P(f)H_{\text{PTF}}(f)R_{\text{earcanal}}(f)M_{\text{min}}(f)} \tag{3.10}$$

$$= H_{\text{HRTF}}(f)H_{\text{earcanal}}(f)X(f).$$

3.2 Measuring HRTFs in Practice

In this section it is explained how HRTFs are measured in practice. The method presented here is not the only way to measure HRTFs, but rather an example how it can be made.

3.2.1 Placement of Microphone

In this project open ear canal measurement was used. The reason for this was the lack of need for a compensation of the difference in impedance as mentioned in Section 3.1.2. An example from the position of the microphone can be seen in Fig. 3.1. The microphone is positioned at the entrance of the ear canal near the tragus. Even the slightest movement of the microphone causes changes in the frequency response that can be heard. This was noticed by the author in informal testing of HRTF-measurement. That is why all measurements must be done at once without removing or moving the microphone between measurements. Also PTFs must be measured without moving the microphones.

A reason for the difference in the response caused by a slight movement of the microphone might be found from resonances in the ear canal. There are several resonances in the ear canal that are in the audio frequencies. A few resonances are depicted in Fig. 3.2 . Sound pressure depends on where in the modal resonance the microphone is. If the microphone is in the anti-node,

Figure 3.1: Microphone is inserted to the entrance of the ear canal.

pressure is equal to the maximum of the resonance. If the microphone is in the node, pressure is equal to zero. Dimensions in these modal resonances are small. Thus a small difference in the position of the microphone causes a difference in the frequency response that can be heard. This should be studied more carefully in the future.

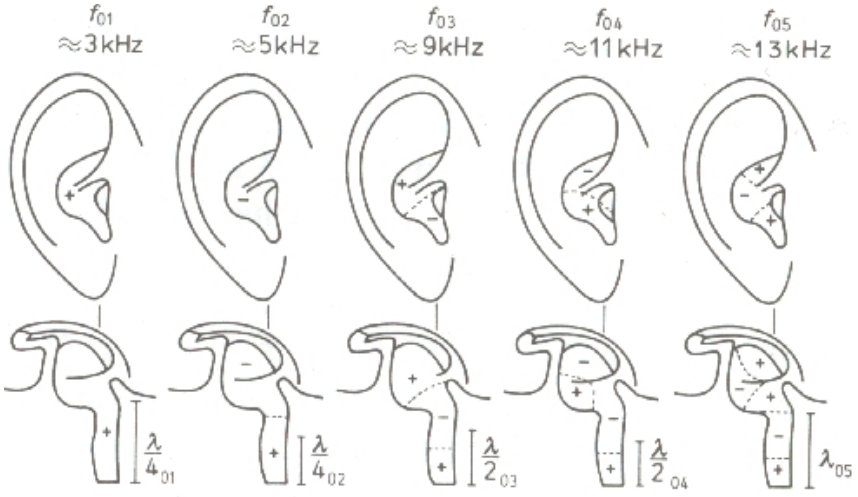

Figure 3.2: Distribution of sound pressure for several natural resonances of a model external ear with a high-impedance termination. The dotted lines denote nodal planes of sound pressure. [23]

These resonances are a problem in the measurement of HRTFs. Theoretically these resonances should cancel out with PTFs because PTFs have the same resonances as HRTFs, but in practice big dips and peaks are hard to be equalized. More about equalization in Section 3.2.3.

3.2.2 Measurement Procedure and Apparatus

Measurements were done in an anechoic chamber. Microphones used were Sennheiser KE4, see Fig. 3.1. The microphones were inserted to the entrance of the ear canal as described in Section 3.2.1. A test subject was sitting in a chair that had a rest for the head, see Fig. 3.3. The subject was looking into one direction all the time while sitting still. A loudspeaker was attached to the end of a boom that turns both in azimuth and elevation angles.

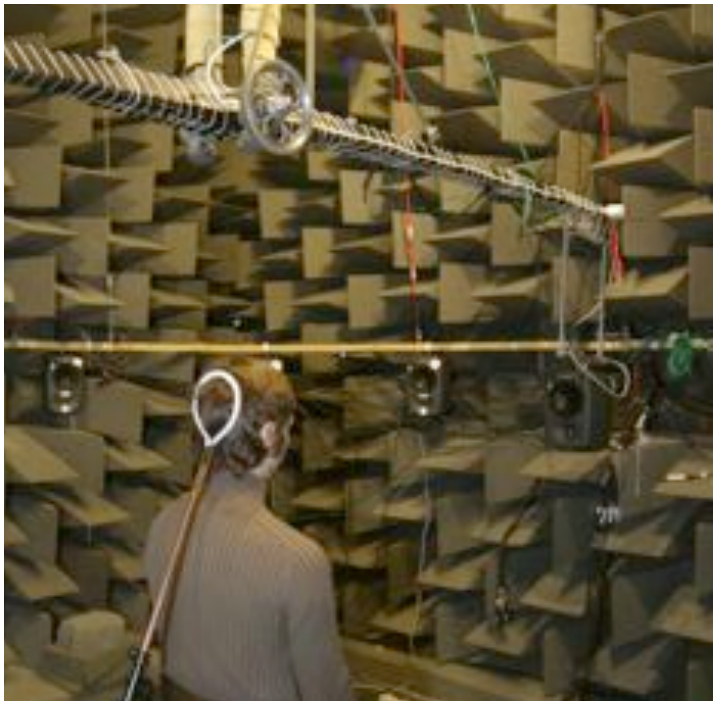

Figure 3.3: Measuring HRTFs. Loudspeaker at the end of the boom is the sound source.

The boom is not completely acoustically transparent and there are also other loudspeakers in the room. Thus the room is not fully anechoic but rather "low-echoic". However differences compared to an anechoic room, which does not have loudspeakers and a boom, in the magnitude frequency response were smaller than 0.5 dB when using third octave smoothing and smaller than 1 dB when using 1/12 octave smoothing, see Fig. 3.4. The impulse response was filtered in octave bands. From these responses it was observed that sound pressure levels of reflections caused by the boom and the loudspeakers were more than 20 dB more silent than the sound pressure level of the direct sound. The loudspeaker used was Genelec 8030A. It is a two-way loudspeaker. In theory this can cause problems because low frequencies come from different direction than high frequencies. This problem was considered to be insignificant because the distance between centers of the drivers is only 0.11 m and exact direction of the sources is not the main focus in this project.

Impulse responses were measured using a method called logarithmic sweep (logsweep). The

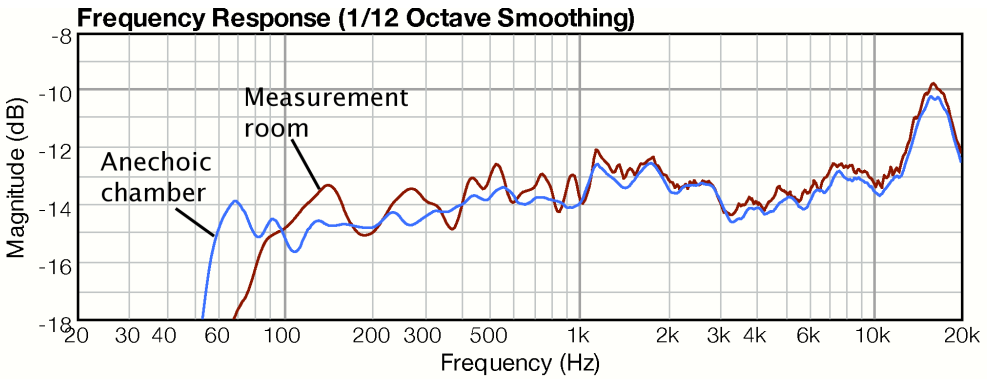

Figure 3.4: Frequency response of a certain loudspeaker measured in an anechoic chamber and in the measurement room used for HRTF measurement. The boom was in the measurement room while measuring. Difference at low frequencies is caused by difference in the room size.

excitation signal is a sine sweep where frequency increases exponentially. The sweep started from 20 Hz and ended at 22 kHz. The sampling frequency was 48 kHz. The length of the sweep was 1 second. These sweeps were reproduced using a loudspeaker and were recorded with microphones that were at the ears. The inverse of the excitation signal was calculated by

$$H_{\text{inverse}}(f) = \frac{1}{H_{\text{original}}(f)} \tag{3.11}$$

where $H_{\text{original}}(f)$ is the frequency response of the original sweep and $H_{\text{inverse}}(f)$ is the inverse frequency response. Because the sweep started from 20 Hz and ended at 22 kHz, sound pressure level is zero outside this band. Thus sound pressure level in the inverse filter is infinity outside the band of the sweep. That is why the inverse signal must be low-pass and high-pass filtered in order not to boost infra- and ultrasounds, see Fig. 3.5. Impulse response is obtained by calculating convolution between a bandpass filtered inverse filter and a measured sweep. The computation was performed in the MATLAB environment. For more information about impulse-response measurement with sweep, see [33], [34].

For the project a large number of HRTFs were needed to be measured. That is why the boom was constantly rotating in azimuth direction while measuring. This ensured faster measurement than if the boom had been stopped for every measurement. The rotation of the boom in azimuth direction causes noise that is only slightly noticeable. The boom and the machine it is connected to and also the loudspeaker are quite heavy so stopping it instantly was not possible. Thus stopping the boom would have taken too much time. The boom was rotating $2-3$ degrees in one second which is the length of the sweep. This causes that high frequencies are measured from other direction than low frequencies. This is similar difference as caused by the two-way structure of the loudspeaker. With this kind of rotating speed, the HRTFs all around the subject in one elevation with a spacing of 6 degrees in azimuth can be measured in two minutes. The

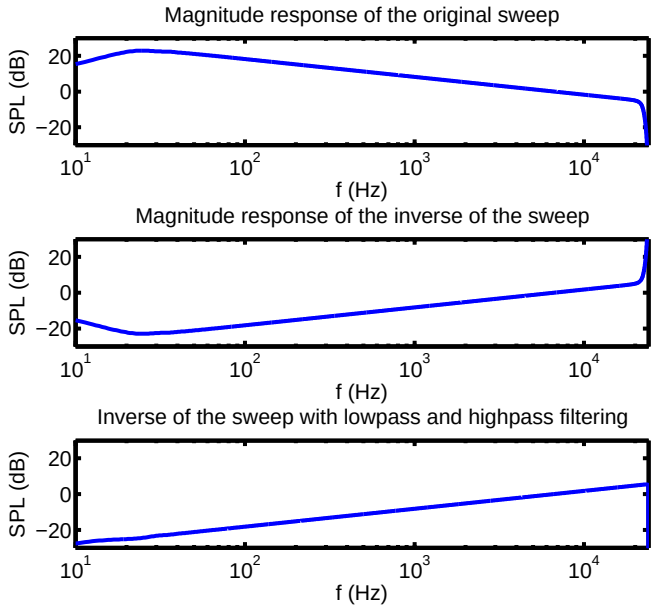

Figure 3.5: The magnitude response of the excitation sweep can be seen in the upper part of the figure. The inverse of the excitation signal is in the middle. The inverse spectrum with lowpass and highpass filtering is in the lower part of the figure.

measurement system was controlled using MAX/MSP software.

3.2.3 Equalizing HRTFs in Practice

There are several ways to do headphone and loudspeaker compensation. The simplest way to do this is to transform signals to the frequency domain and to divide measured HRTF with the frequency response of the loudspeaker and headphones. The problem with this approach especially with headphones is that at high frequencies the response is not consistent between different measurements, dips and peaks in the response may appear at different frequencies.

Let us consider a dip at certain frequency in a PTF. If this response is used as an inverse filter, this frequency is boosted. Now these same headphones are used for listening. It is impossible to position the headphones exactly in the same position. Now the dip is moved to another frequency and a peak may appear at this frequency. So there is a peak in the response of the headphones and also in the correction response. The result is a clearly perceived boost in the frequency response at this frequency. Peaks in the frequency response are more easily heard than dips [35]. That is why especially compensating dips in the frequency response should be done with a great care if at all.

PTFs are typically consistent up to about 7 kHz [22], depending on headphones. At higher frequencies responses vary significantly between different measurements. This upper limit can be observed by measuring several PTFs by taking headphones off and placing them back on the

ears between measurements. The upper limit in frequency for a consistent PTF can be seen by comparing different PTFs.

PTFs can be directly used for the compensation filter up to the frequencies where it is consistent. The frequency response is simply inverted to get the compensation filter. Above the upper limit only peaks in the response are compensated. This means that peaks in the compensation filter are compressed and dips in the frequency response of the compensation filter are left unaltered. A resulting compensation filter can be see in Fig. 3.6.

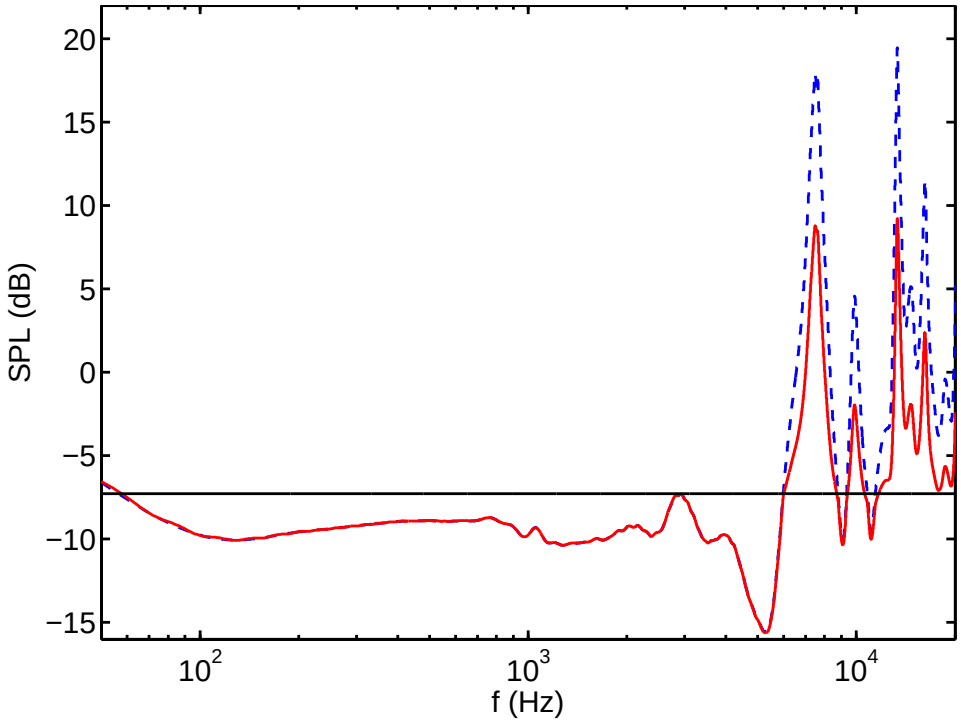

Figure 3.6: Dashed line is direct inversion calculated from measured PTF. In solid line peaks of the response have been compressed. Solid horizontal line is level above which compression starts.

Peaks can be found by finding local maxima or by comparing response to a constant level or by comparing response to a highly smoothed version of the response. For ultimate quality, peaks can be manually found by looking the magnitude response. This is naturally very time-consuming so an automated algorithm was created. By informal testing the best results were gained by comparing the response to a constant level. This level was obtained by calculating the rms-value of the signal in the frequency range where the response is consistent. This was decided to be between 40 Hz and the first large dip in the inverse response. This dip lies typically around 5 kHz. The level is shifted slightly upwards from this rms-value, for example by multiplying the rms-value by 1.25.

Peak compression was implemented by subtracting the constant level from values greater than

it. This value is divided by the compression ratio and the constant level is added again. A good value for the compression ratio is about 3. Peaks were compressed enough that they do not cause coloration. If peaks were compressed more, there would be perceivable dips in the magnitude response when reproducing sound using these compensation filters.

As mentioned before, PTFs change clearly between different positionings of the headphones and the headphones cannot be positioned perfectly. That is why the best possible compensation filter is such that works robustly producing satisfactory compensation. It is not possible to achieve perfect compensation at high frequencies using conventional headphones. The quality of the compensation can be observed by filtering different PTFs by a compensation filter that has been calculated from another PTF. Filtering by different PTFs corresponds to listening through headphones. The result from this filtering should be a magnitude response of unity. An example result of the compensation can be seen in Fig. 3.7. A compensation filter with compressed peaks produces responses that have dips in the response at high frequencies but no peaks or at least smaller peaks. A compensation filter that is a direct inverse of the PTF produces responses that also have peaks in them.

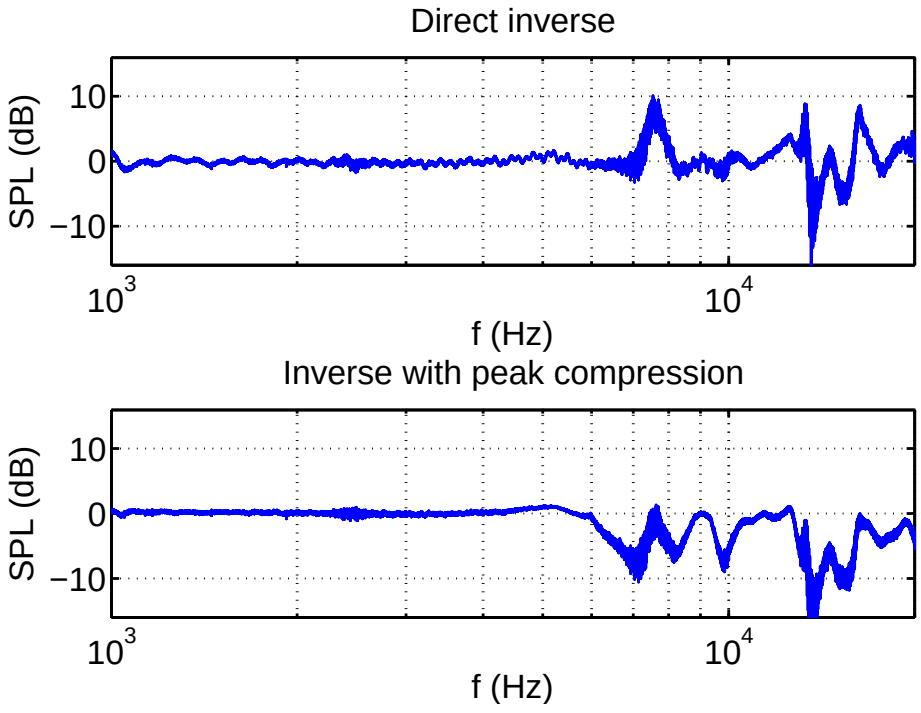

Figure 3.7: The result of the compensation by a headphone compensation filter that has been calculated from a PTF by directly inverting it in the frequency domain can be seen in the upper part of the figure. In the lower part of the figure, the compensation has been made by a filter that has compression of the peaks at high frequencies. These compensation filters correspond to filters presented in Fig. 3.6.

The quality of the headphone compensation was also tested by listening to pink noise that was filtered with different compensation filters. Peaks in the frequency response were clearly audible when using the direct inverse filter. Boosts at certain frequencies were so large that they sounded annoying. When using the filter with compressed peaks the response sounded significantly smoother. Slight coloration could be heard at high frequencies but it was mild compared to the direct inverse filter. Dips in the response were perceivable but not annoying.

Inverse filters can become unstable, because PTFs are typically mixed-phase filters, which causes direct inverses of them to be unstable. Often the phase response is discarded and compensation is designed for the magnitude response [22]. Filtering the response of the loudspeaker out can be done in a similar way, however this was not performed in this work. The reason for this was that the binaural reproduction was to be compared to the loudspeaker reproduction, where the responses of loudspeakers exist. Also the frequency response of the loudspeakers used in the measurement was flat enough, deviation less than 2 dB between 58 Hz and 20 kHz.

The length of actual HRTFs is less than 150 samples at the sampling rate of 48 kHz, which corresponds to about 3 ms or 1 meter of propagation of sound. This length corresponds to a time difference between the direct sound to the nearest ear and reflections from the body of the listener. Headphone compensation added some length to the used impulse response. The result was 512 samples long impulse response. The last 100 samples were windowed with a Hann window to get rid of discontinuations. The rest of the response was discarded. This way the reflections from for example loudspeakers in the room were removed.

An example of a measured HRTF can be seen in Fig. 3.8.

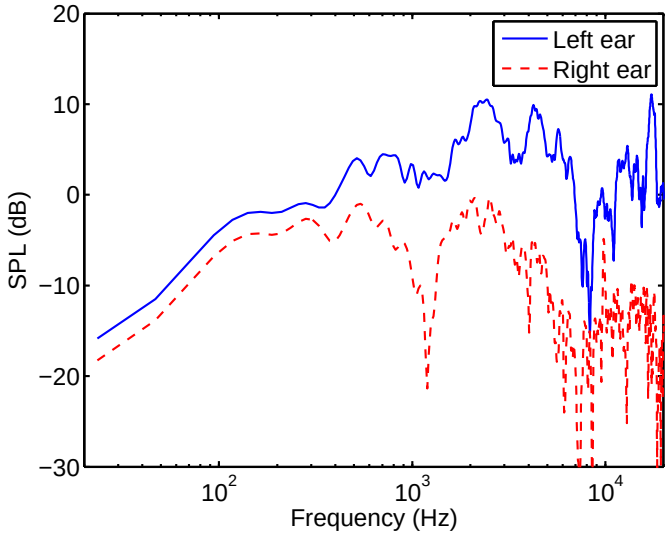

Figure 3.8: HRTFs measured with open ear canals. In the figure, there is a HRTF for both left and right ear. The sound source is 45 degrees to the left in the horizontal plane. PTF correction is added.

3.2.4 Transfer Functions for Binaural Reproduction

As mentioned in the previous sections, measured HRTFs have differences to what an HRTF is defined to be. HRTFs include the response of the loudspeaker, responses were not measure in a completely anechoic chamber and the sound source was constantly moving during the measurement. So strictly speaking the measured responses are not HRTFs. They could be called for example transfer functions for binaural reproduction. However for simplicity these transfer functions are referred to as HRTFs in the rest of the thesis.

3.3 Synthesis of Spatial Sound Using HRTFs

After HRTFs have been equalized, using them is quite straightforward. Sound is reproduced by headphones. Convolution between an input signal and the HRTF for each ear is computed. The signal convolved with the left HRTF is reproduced from the left channel of the headphones and the signal convolved with the right HRTF is reproduced from the right channel of the headphones. The auditory object is perceived to be positioned to the direction from which the HRTFs have been measured, at least in theory.

3.4 Problems with HRTF techniques

When HRTFs are used in binaural reproduction, at least the following problems can occur:

- Auditory objects are not externalized, but instead remain localized inside the head or on the surface of the head.
- Front-back or back-front confusion. Auditory objects that should be in front of the listener are perceived to be behind the listener and vice versa.
- Error in perceived direction of an auditory object.

When listening to a stereo recording using headphones, the auditory object is panned between the ears inside the head which is called lateralization. With HRTFs the auditory object should be outside the head which is called externalization. Especially with frontal sound sources the auditory object does not externalize properly. Instead the auditory object is localized inside the head or on the surface of the head or just outside the head. Let us consider the measuring of the HRTFs and the cues for the perception of distance as presented in Section 2.5.5. If HRTFs are measured over 1 meter away from the sound source in anechoic chamber the only cue for the distance is sound pressure level. If SPL is kept constant, there are no cues for the perception of the distance. HRTFs have no reverberation and if the input signal is anechoic the result is also anechoic. As presented in Section 2.5.5 the lower the amount of reverberation, the shorter the judged distance. Thus it is expected that sound sources are not externalized far away.

The level of externalization between anechoic and echoic HRTFs was compared in [36]. Echoic HRTFs were made with a room modeling program. The result was that echoic HRTFs externalized in 79 % of the cases while anechoic HRTFs externalized only in 40 % of the cases.

Frontal sources are the most problematic with HRTFs. Problems can be the lack of externalization, front-back confusion or that the direction of the auditory object is perceived to be higher in elevation than it should be. A reason for this could be that a frontal sound source is in contradiction with visual input. We cannot see a sound source in front of us so it is hard to believe that there really is something. The effect of head-tracking to front-back and back-front reversal was studied in [36]. With head-tracking reversal happened in 28 % of the cases and without it in 59 % of the cases. More about head-tracking in Section 5.3.

Chapter 4

Directional Audio Coding

The principles of directional audio coding (DirAC) [1] are presented in this chapter. Binaural reproduction is not considered here. More detailed view is presented in Ch. 5 and also how DirAC was implemented binaurally. DirAC is based on a technique called spatial impulse response rendering (SIRR) [6], [37]. SIRR is used to reproduce room impulse responses.

4.1 Basic Idea of DirAC

DirAC is based on an idea that there is no need to reproduce sound pressure field perfectly. It is assumed that it is enough to reproduce sound perfectly only in a sense what humans perceive. As presented in previous sections, the auditory perception differs clearly from what we can measure using a microphone.

DirAC is based on four assumptions about the interaction between sound field properties and perceptual attributes they produce [6], [1]:

1) The direction of sound transforms into ITD, ILD and monaural cues, see Section 2.5.

2) The diffuseness of sound transforms into interaural coherence cues, see Section 2.5.3.

3) Timbre depends on the monaural spectrum together with ITD, ILD and interaural coherence.

4) The direction of arrival, diffuseness and spectrum of sound measured in one position with the temporal and spectral resolution of human hearing determines the auditory spatial image the listener perceives.

The last assumption concludes the idea of DirAC. At one time instant the auditory system is able to decode only one spatial cue within one critical band. Instead of storing a separate audio channel for each loudspeaker, only one audio channel can be stored with information about the direction and the diffuseness of sound at different critical bands. The temporal accuracy should be the same as the temporal accuracy of the corresponding critical band. If audio signals would be stored separately for each loudspeaker, then we would have more precise information about the direction in frequency and time, but we would not be able to perceive it.

DirAC processing consists of two parts: the analysis and the synthesis. In both parts sound is divided in time and in frequency. Processing is performed separately for every block that represents sound at a certain time instant and frequency band. In the analysis phase the direction and the diffuseness of sound are calculated. These are used in the synthesis phase.

In the synthesis phase the sound is divided into two parts: nondiffuse and diffuse sound. The nondiffuse part has certain direction and it can consist for example of direct sound and first few reflections in a room. The diffuse part does not have direction. It can consist for example of diffuse sound in a room. In DirAC diffuse sound can also mean other things. It can be noise that is coming from multiple directions or multiple simultaneous talkers around the microphone. The nondiffuse part is reproduced accurately from the desired direction. The diffuse part is processed so that the direction of it cannot be perceived.

4.2 B-format Audio

B-format microphone signals are used in DirAC. B-format audio consists of four channels: W, X, Y and Z. The W-channel is an omnidirectional channel. X-, Y- and Z-channels have the directional pattern of a dipole. They are orthogonal in space. They correspond to axes in a typical Cartesian coordinate system. The directional patterns of W-, X- and Y-signals can been seen in Fig. 4.1. The dipole signals are scaled by $\sqrt{2}$. There are commercial B-format microphones available such as Soundfield ST350. A low-cost X-Y-direction microphone can be built from four omnidirectional microphones that are placed in a square [38]. In it dipoles are formed by subtracting opposing omnidirectional signals and equalizing the result. The W-channel is formed by summing all four channels.

4.3 DirAC Encoding

Input signals for DirAC encoding are the B-format signals. First the signals are divided into blocks in time and frequency. There are two possible approaches to this: filterbank and short-time Fourier transform (STFT). In STFT certain length blocks of the signal are transformed to frequency domain using fast Fourier transform (FFT). Before calculating FFT, the block of signal is windowed using for example a Hann window. The signal is divided into blocks also in frequency. Division is based on ERB-bands so that one block corresponds to one ERB-band. In the filterbank implementation the signal is first divided into frequency bands. After that the signal can be divided in time. The advantage of the filterbank is that the signal can be divided in time into different size blocks in different frequency bands. Regardless of the method, the signal is divided into blocks that represent the signal at certain time instant and critical band of hearing. This process is depicted in Fig. 4.2.

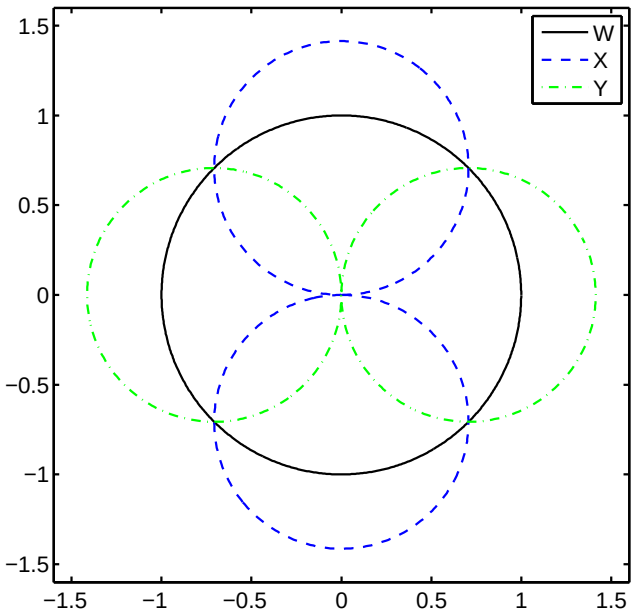

Figure 4.1: The directional patterns of W-, X- and Y-channels of B-format audio.

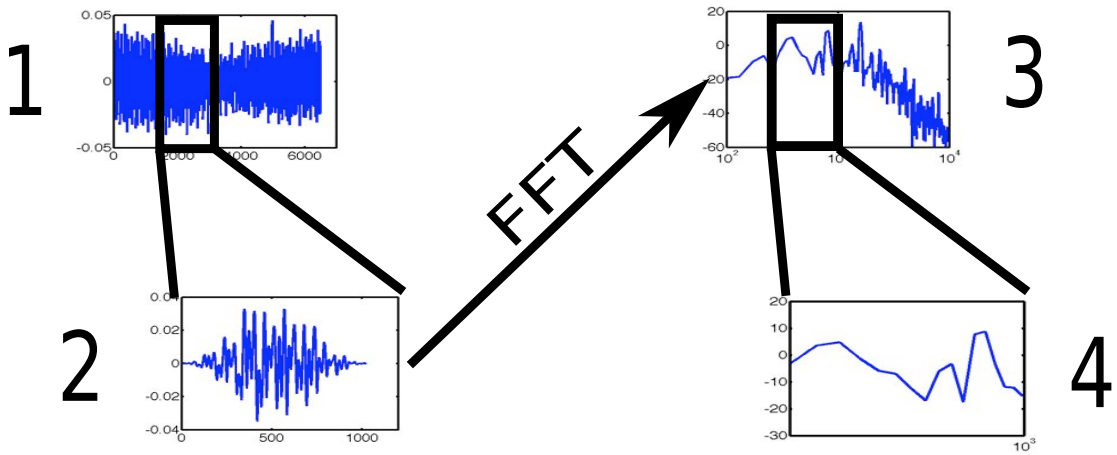

Figure 4.2: Dividing signal into blocks in time and frequency using the STFT-method. In part 1 of the figure there is a continuous time-domain signal. A certain size block of it is cut for analysis and it is windowed using a Hann window. The result can be seen in part 2 of the figure. FFT is calculated from this block, see part 3. Values corresponding to a certain ERB-band are cut for the analysis. In part 4 a block of the signal representing certain time instant and critical band of hearing can be seen.

4.3.1 Direction Analysis

As presented in Eq. (2.2), instantaneous intensity is defined as the product of the sound pressure $p(t)$ and the particle velocity vector $\boldsymbol{u}(t)$. Intensity can be split into two parts: active intensity and reactive intensity. Active intensity corresponds to local net transport of acoustic energy. Reactive intensity corresponds to local oscillatory transport of energy. The time average of it is zero. Active intensity is used to determine the direction of the sound. It can be thought as a time average of instantaneous intensity [6]. The analysis is done in the frequency domain in the STFT-implementation. In the frequency domain, active intensity $\boldsymbol{I}_a$ can be calculated from

$$\boldsymbol{I}_a(\omega) = 2\text{Re}\{P^*(\omega)\boldsymbol{U}(\omega)\} \tag{4.1}$$

where $P(\omega)$ is the Fourier transform of the sound pressure in a time window and $\boldsymbol{U}(\omega)$ of the particle velocity. Re denotes taking real part and * complex conjugate.

Approximates of pressure and particle velocity are needed for intensity calculation. $W(t)$ can be directly used for the pressure $p(t)$ because $W(t)$ is the pressure measured using an omnidirectional microphone. The particle velocity can be calculated from the pressure using acoustic impedance Z which is defined as

$$Z = \frac{p}{u} = \rho c \tag{4.2}$$

for plane waves traveling in positive direction, where ρ is the density of medium and c is the speed of sound [7].

An approximation of the particle velocity can be calculated from the B-format signals using

$$\boldsymbol{u}(t) = \frac{1}{\sqrt{2}Z}\boldsymbol{X}'(t) \tag{4.3}$$

where

$$\boldsymbol{X}'(t) = X(t)\boldsymbol{e}_x + Y(t)\boldsymbol{e}_y + Z(t)\boldsymbol{e}_z. \tag{4.4}$$

$\boldsymbol{e}_x$ represent unit vector in the x-direction Cartesian coordinate axis etc. Dipole signals are divided by $\sqrt{2}$ in Eq. (4.3), because in B-format they are scaled by $\sqrt{2}$, see Section 4.2.

Intensity can be calculated by substituting Eq. (4.3) into Eq. (4.1) and using $W(t)$ for $p(t)$. In practice it is calculated separately for different coordinate axes, for example for x-axis using

$$I_x(\omega) = \frac{\sqrt{2}}{Z}\text{Re}\{W^*(\omega)X(\omega)\}. \tag{4.5}$$

The direction of arrival of the sound can be calculated if the direction of intensity is known. Intensity points to where sound is going to so the direction of the arrival is opposite to this. In DirAC the direction is presented using two angles: the azimuth $\theta(\omega)$ and the elevation $\phi(\omega)$. These can be calculated from

$$\theta(\omega) = \tan^{-1}\left[\frac{-I_y(\omega)}{-I_x(\omega)}\right] \tag{4.6}$$

and

$$\phi(\omega) = \tan^{-1}\left[\frac{-I_z(\omega)}{\sqrt{I_x^2(\omega) + I_y^2(\omega)}}\right]. \tag{4.7}$$

$\tan^{-1}$ gives two results that are separated by π. The correct result must be chosen according to the signs of $-I_x(\omega)$, $-I_y(\omega)$ and $-I_z(\omega)$.

4.3.2 Diffuseness Analysis

Diffuseness can be defined as a proportion of sound energy oscillating locally [6]. The instantaneous energy density can be calculated using Eq. (2.3). Diffuseness is defined as

$$\psi = 1 - \frac{||\langle \boldsymbol{I}(t)\rangle||}{c\langle E(t)\rangle} = 1 - \frac{2Z||\langle p(t)\boldsymbol{u}(t)\rangle||}{\langle p^2(t)\rangle + Z^2\langle \boldsymbol{u}^2(t)\rangle} \tag{4.8}$$

where $||\cdot||$ denotes the norm of the vector and $\langle\cdot\rangle$ denotes time averaging. From Eq. (4.8) can be seen that as simplified diffuseness means the proportion of energy that has no stable direction.

Estimates for pressure and particle velocity are obtained in a same way as in the direction calculation. As well as direction, diffuseness is calculated in the frequency domain. That is why the equation is presented also in the frequency domain

$$\begin{aligned}\psi &= 1 - \frac{2Z||\mathrm{Re}\{P^*(\omega)\boldsymbol{U}(\omega)\}||}{|P(\omega)|^2 + Z^2|\boldsymbol{U}(\omega)|^2} = 1 - \frac{\sqrt{2}||\mathrm{Re}\{W^*(\omega)\boldsymbol{X'}(\omega)\}||}{|W(\omega)|^2 + |\boldsymbol{X'}(\omega)|^2/2} \\ &= 1 - \frac{\sqrt{2}\sqrt{\mathrm{Re}\{W^*(\omega)X(\omega)\}^2 + \mathrm{Re}\{W^*(\omega)Y(\omega)\}^2 + \mathrm{Re}\{W^*(\omega)Z(\omega)\}^2}}{|W(\omega)|^2 + (|X(\omega)|^2 + |Y(\omega)|^2 + |Z(\omega)|^2)/2}.\end{aligned} \tag{4.9}$$

4.3.3 Output of DirAC Encoding

The output of DirAC encoding is an omnidirectional microphone channel $W(t)$ and a side channel. In the virtual microphone version of DirAC also X-, Y- and Z-signals are transmitted. More about the virtual microphone version in Sect 5.2.1. A diagram of DirAC encoding can be seen in Fig. 4.3.

4.4 DirAC Transmission

The side channel includes information about azimuth angle, elevation angle and diffuseness. These parameters are transmitted for every ERB-band for every time domain signal block. There are about 40 ERB-bands. Let us assume that the signal block is 512 samples long. Angles can be presented using really low accuracy because humans cannot perceive the direction of the auditory event very accurately. Localization accuracy is about 3.6 degrees in the front and lower elsewhere [23]. Also diffuseness does not need very accurate presentation. Let us assume that the direction and the diffuseness are presented using 8 bits. This means that for every 512 samples

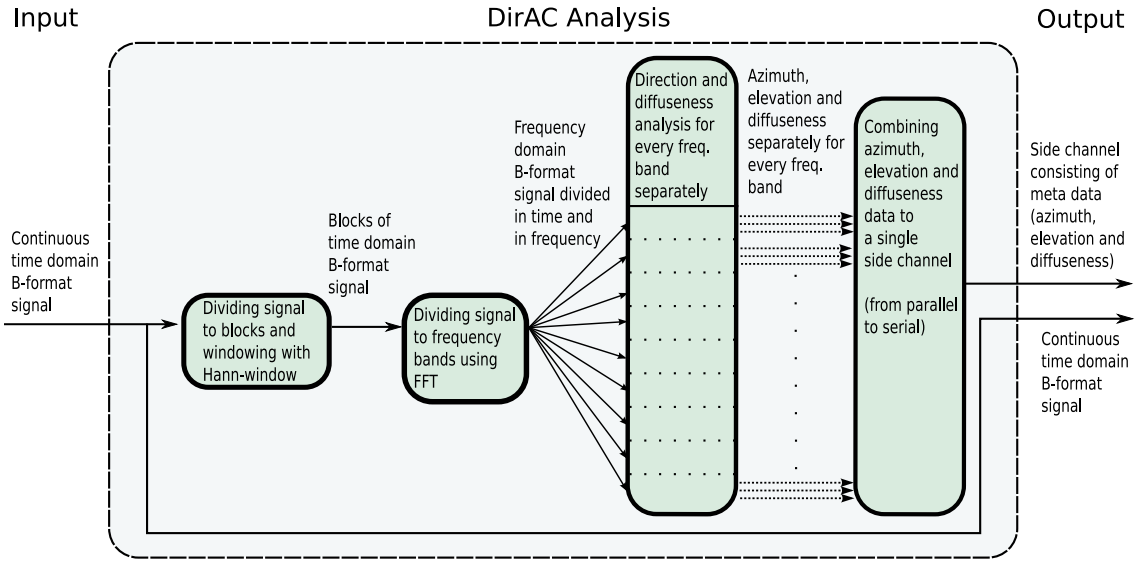

Figure 4.3: A diagram of DirAC encoding. In case of omnidirectional version of DirAC, only omnidirectional signal is in the output.

of audio the number of 8-bit words to be transmitted is 120. From these rough assumptions the result is a bit-rate of 2 bits per sample. Sound signals typically consist of 16-bit words. Compared to that the side channel needs a low bit-rate.

In the omnidirectional version of DirAC the mono audio signal can be compressed. Conventional compression algorithms like mp3 [39] are designed for single channels. They are based on the principle of masking, see Section 2.4.3. In case of DirAC, different frequency bands can come from different directions instead of same direction. If a masking sound is coming from different direction than a masked sound, the masked sound will be more easily perceived. This means that the threshold of hearing for masked sound is lower when using DirAC than when using a mono source. This must be taken into account when compressing a signal. Thus compression of sound is not as efficient as with a mono signal. In the virtual microphone version the compression algorithm should not confuse the phase between channels.

At the end of the transmission the signal must be converted back to an uncompressed signal. Input to DirAC decoding is an uncompressed audio channel or channels and a side channel.

4.5 DirAC Decoding

At the beginning of DirAC decoding a signal is divided in time and frequency. This is done in a same way as in the encoding phase, see Fig. 4.2. The signal is divided into two streams: the diffuse and the nondiffuse stream. This is performed by multiplying the signal by $\sqrt{\psi}$ to obtain the diffuse part and by $\sqrt{1-\psi}$ to obtain the nondiffuse part, see Fig. 4.4. These two parts are

synthesized in a different way.

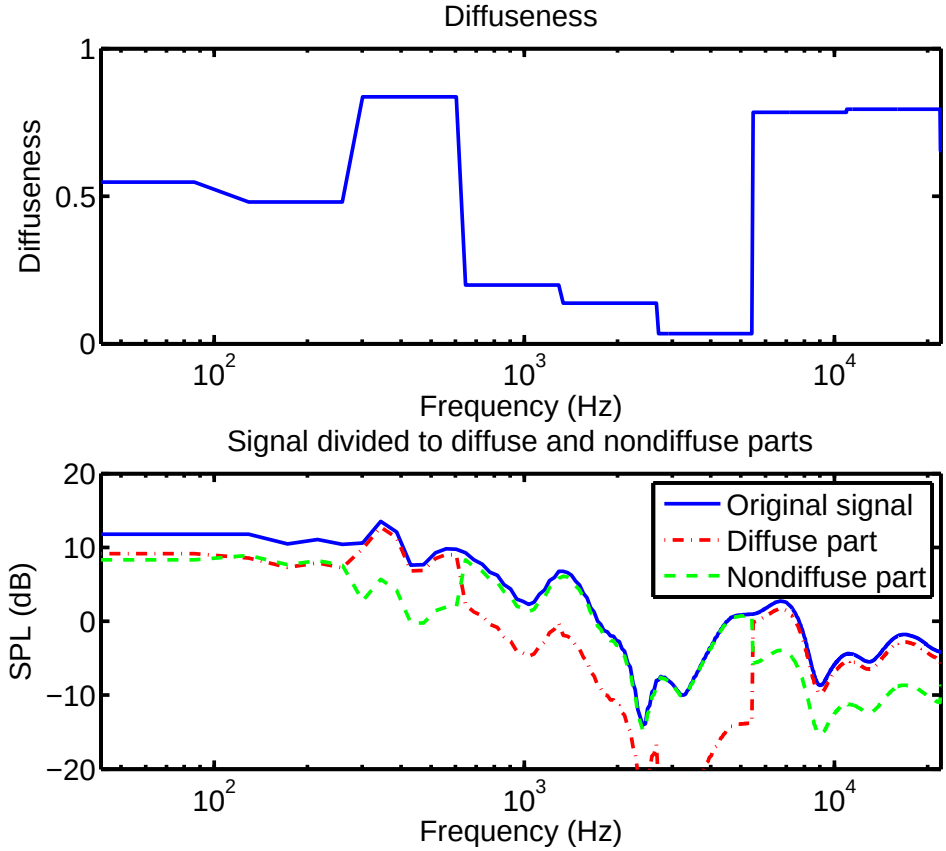

Figure 4.4: The diffuseness of some block of a signal is in the upper part of the figure. In the lower part of the figure the original signal is multiplied by $\sqrt{\psi}$ to get the diffuse part and by $\sqrt{1-\psi}$ to get the nondiffuse part, where ψ is the diffuseness. Multiplication is done in the frequency domain.

4.5.1 Synthesis of Nondiffuse Sound

The nondiffuse part of sound includes sounds that have certain direction. This part of sound is panned to certain direction using vector base amplitude panning (VBAP), see [40]. VBAP uses sets of three loudspeakers to reproduce sound from certain direction, see Fig. 4.5. The same sound is reproduced from all of these three loudspeakers. By adjusting the gains of the loudspeakers, the virtual sound source can be positioned anywhere inside the triangle defined by the loudspeakers. Using this method the nondiffuse part of sound is panned to desired direction at each frequency band. Other techniques could also be used for panning sound. VBAP allows panning in three dimensions in quite a straightforward way, so it is a good choice for panning the nondiffuse sound.

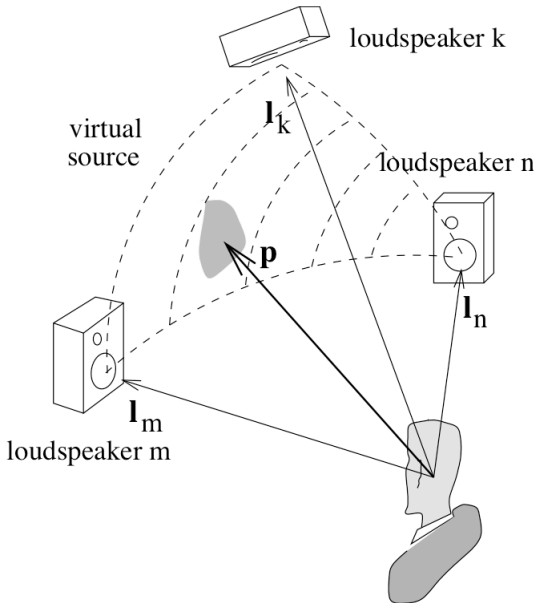

Figure 4.5: Vector base amplitude panning (VBAP). Sound is panned to right direction using sets of three loudspeakers. [41]

4.5.2 Synthesis of Diffuse Sound

The diffuse part of the sound includes sounds that have no stable direction, which means that intensity is oscillating. The aim of the synthesis of the diffuse sound is to create sound that surrounds the listener and it has no distinct direction [6]. The timbre of sound should not be affected.

The simplest way to create diffuse sound is to add the diffuse part of the sound to every loudspeaker. Using this method diffuse sound is coherent in every loudspeaker, so it is positioned to the direction where the closest loudspeaker is. The sound from different loudspeakers comes at different time instants. This causes comb filtering, see Fig. 4.6. If the listener moves his/her head, delays change and also notches in the frequency response move. This is heard as an effect similar to flanger or phaser. Obviously this is not desired.

The solution is to decorrelate the diffuse sound signals to every loudspeaker. There are many ways to do this and they are discussed in more detail in Section 5.2.4. Basically the phase at each loudspeaker must be randomized. This can be done for example by filtering the signal with a random noise burst that has a magnitude response of unity. Noise must be different for every loudspeaker. The result is a signal with the same magnitude response as the original signal but with a random phase response, see Fig. 4.7. Phase responses of the signals for each loudspeaker are random, which means that they are decorrelated. Another solution is to use random delays for different frequency bands and different loudspeakers or randomizing the phase in the frequency domain.

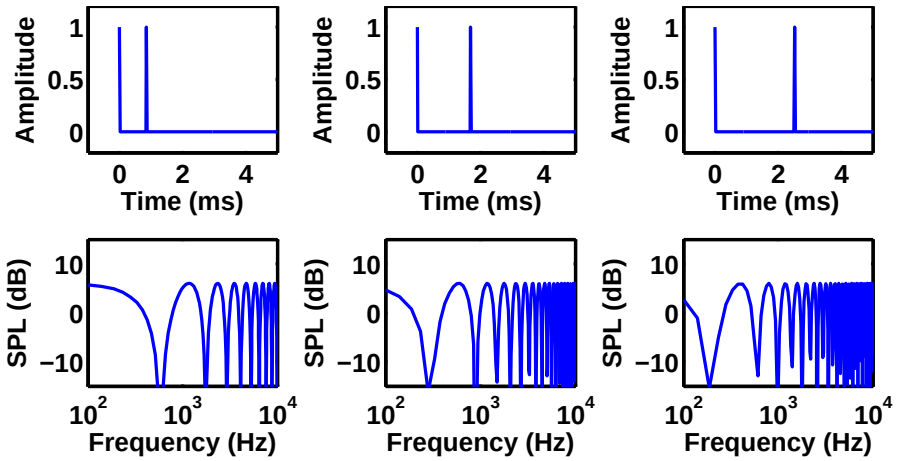

Figure 4.6: Comb filters. The upper part of the figure shows signals that have impulse at time instant zero and another impulse after a delay. The lower part of the figure shows Fourier transforms of the corresponding signals. It can be seen that a delayed impulse causes a comb pattern to the frequency response and notches move in the frequency as a function of the delay.

The result of playing decorrelated signals from multiple loudspeakers around the listener is that sound has no prominent direction and sound surrounds the listener. Moving slightly does not position sound to the nearest loudspeaker.

4.5.3 Output of DirAC Decoding

In the synthesis nondiffuse and diffuse signals are created separately, and signals are also created separately for every frequency band and loudspeaker resulting in a multitude of narrow-band signals. At the last phase of DirAC decoding time domain signals are created for every loudspeaker by summing the corresponding frequency band signals. The nondiffuse and the diffuse signals are also summed together for each loudspeaker. These signals are transformed into the time domain. The synthesis process is depicted in Fig. 4.8.

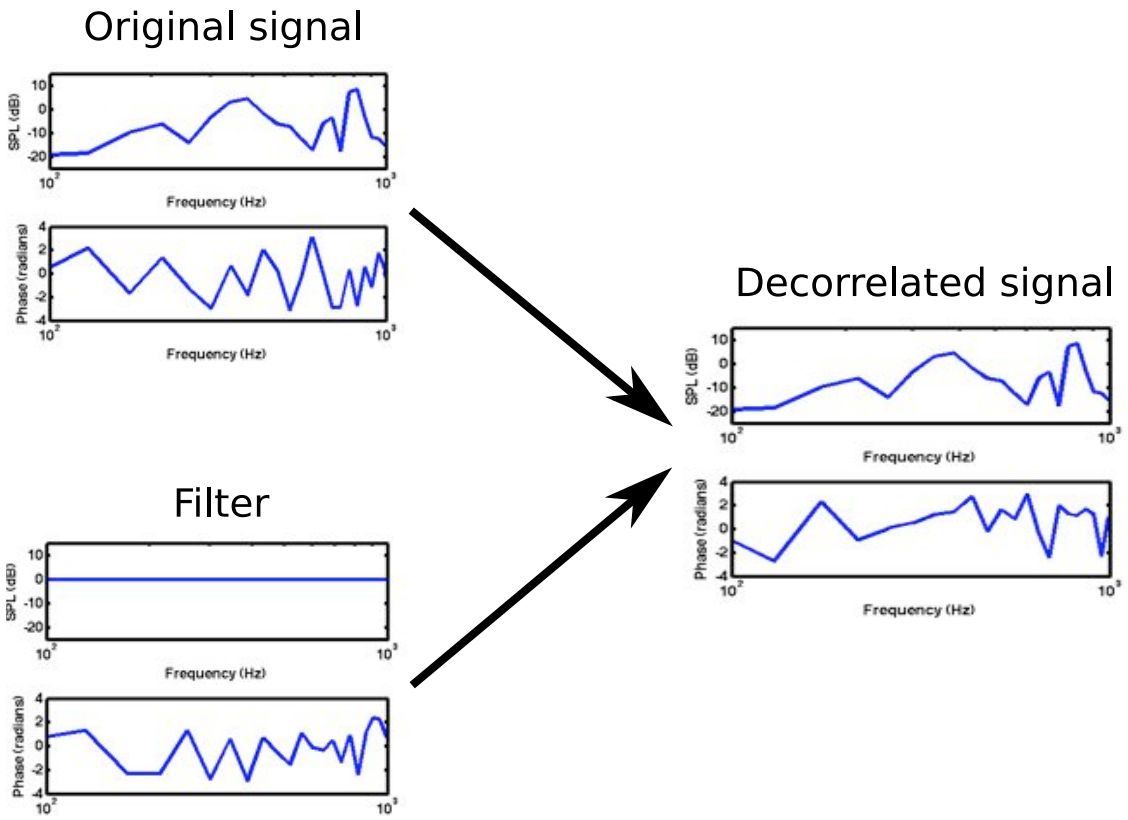

Figure 4.7: Decorrelating a signal by multiplying it by noise that has a magnitude response of unity. Signals are presented in the frequency domain. The upper parts of the figures are magnitude responses and the lower parts are phase responses. The result is a signal that has the same magnitude response as the original signal but a random phase response.

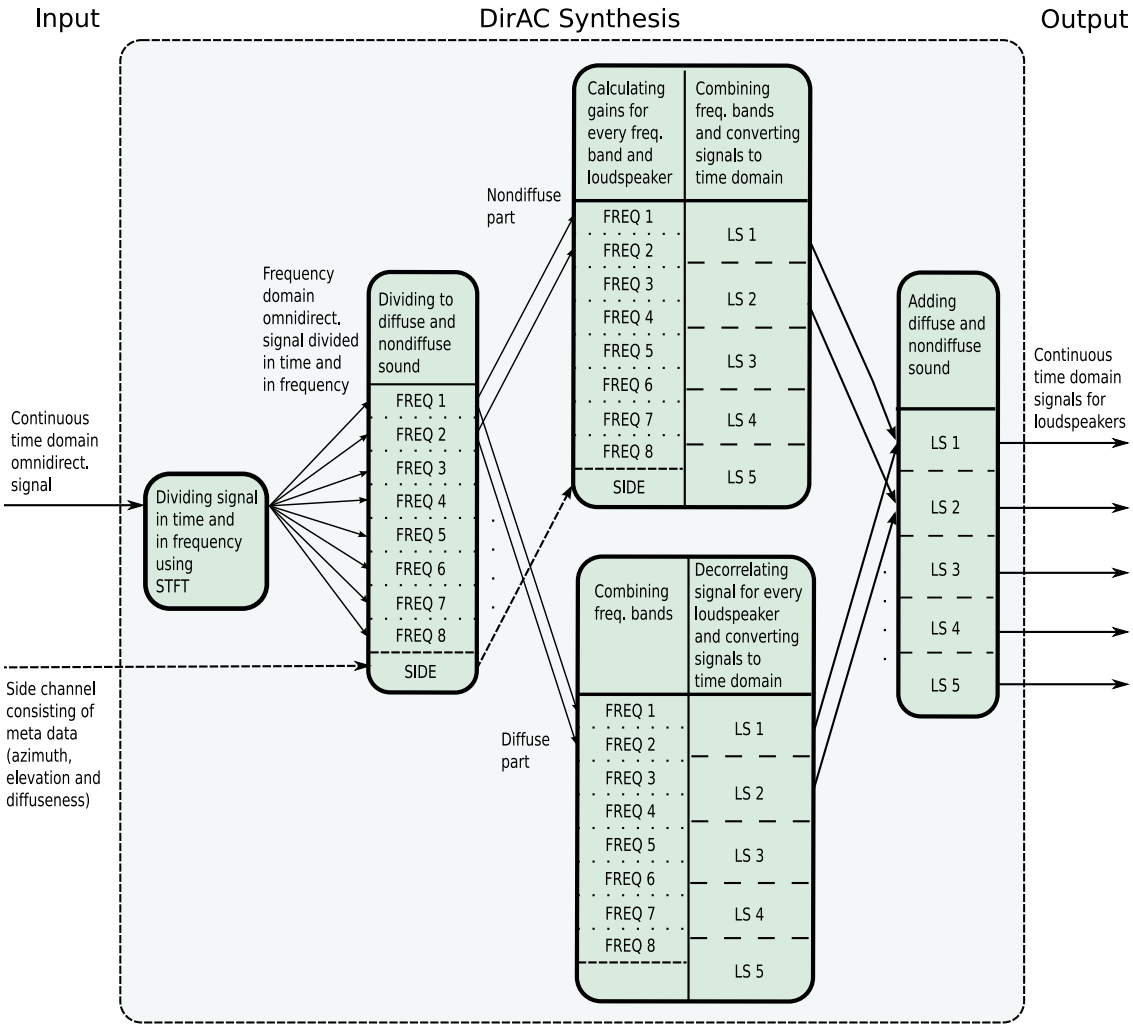

Figure 4.8: A diagram of DirAC decoding. FREQ X stands for frequency band X, LS X stands for loudspeaker X and SIDE stands for the side channel. The omnidirectional version of DirAC is presented in this figure. In this figure there are only 8 frequency bands in order to make the figure less complex. Normally there are more frequency bands.

Chapter 5

Binaural Reproduction for DirAC

This chapter describes how binaural reproduction for DirAC was implemented. It also includes more precise information about the encoding and decoding phases; how they were done in case of binaural reproduction. The principle of binaural reproduction is similar to the loudspeaker reproduction. For example analysis is done in a same way and also synthesis uses same principles as loudspeaker reproduction. Differences and additions to loudspeaker reproduction are explained here.

Binaural reproduction for DirAC was implemented as a real-time application. The application was implemented as an external object to MAX/MSP software using C programming language.

5.1 DirAC Analysis

The STFT approach was selected for this project. The reason for this was that it is computationally more efficient than a filterbank-based implementation. The block size is 1024 samples with a hop size of 512 samples. The blocks are overlapping, see Fig. 5.1. The block is windowed with a Hann window. A windowed 1024 samples long block of signal is transformed into the frequency domain using FFT. Frequency resolution f_{res} can be calculated from

$$f_{\mathrm{res}} = \frac{f_{\mathrm{s}}}{N} \tag{5.1}$$

where f_{s} is the sampling frequency and N is the number of samples in a STFT block. If the window size is too short, the bandwidth of one sample is wider than the bandwidth of the critical bands of hearing at low frequencies. Then the frequency resolution of the analysis would be worse than the frequency resolution of spatial hearing.

The intensity vector inside one block of signal is averaged in time, so the analyzed direction is the average direction of sound and fluctuation causes sound to be analyzed more diffuse. This should be done with temporal accuracy of hearing. If the window size is too long, the analysis is done with worse temporal accuracy than the temporal accuracy of hearing, which causes erroneous direction and diffuseness. Temporal accuracy in the perception of direction

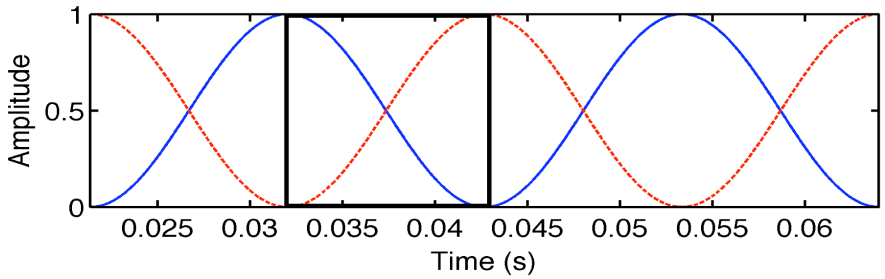

Figure 5.1: A signal is windowed with overlapping Hann windows. Inside the box is the signal at certain time instant. Adding two windows inside the box produces constant amplitude.

is better at high frequencies than at low frequencies. Hence the length of the window size is a compromise between the quality of low and high frequencies.

The sampling frequency was 48 kHz so the frequency resolution was 47 Hz. ERB-band is about 25 Hz at the lowest frequencies, so the frequency resolution of the analysis is almost as good as the frequency resolution of spatial hearing. The length of one block is about 20 ms. At high frequencies, temporal accuracy in the perception of direction is better, thus artifacts can emerge. However, the window size of 1024 samples was considered to be a good compromise.

For real-life recordings or simulated spaces diffuseness and direction change really rapidly, see Fig. 5.2. This is caused by reflections and reverberation. If this would be synthesized, noticeable clicks and artifacts would be produced. Let us assume that there is rapid changing in the direction and the nondiffuse sound is positioned to certain direction only for one time frame of synthesis. The time frame is about 20 ms, so from the loudspeaker corresponding to this direction would be reproduced only a short burst of sound, which would be perceived as a click. Thus diffuseness and the direction must be smoothed. This can be done in the encoding or the decoding phase. The diffuseness was slowed down with a first order IIR filter that had a time constant of 50 ms. Slowing down was applied to intensity and energy vectors defined in Eq. (4.5) and (2.3). Slowing down the direction was not done in the encoding phase. It was done in the synthesis phase.

In order to reduce computational complexity and the size of the side channel, the bandwidth of one frequency band can be larger than one ERB-band. With binaural reproduction the bandwidth can be two ERB-bands without noticing a difference in most of the signals. Even with the bandwidth of four ERB-bands reproduction is acceptable. After about eight ERB-bands the reproduction starts to have clear and significant artifacts. The artifacts are for example that sound sources are perceived to be only in one direction at a time or that sound sources are not stable but instead move towards each other.

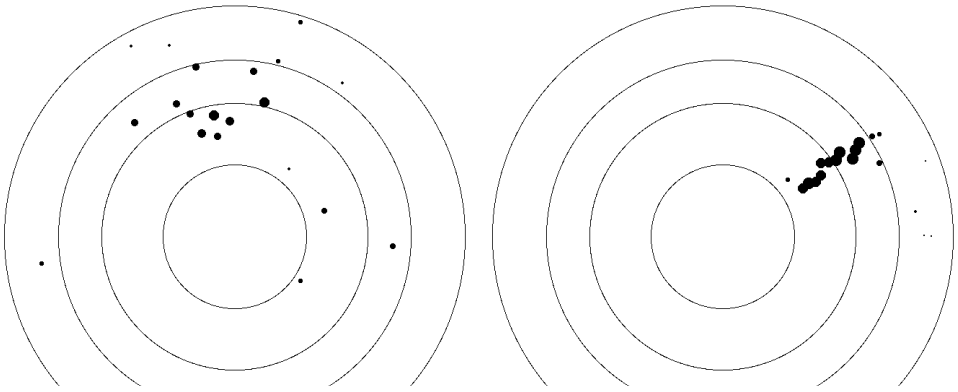

Figure 5.2: A still picture of DirAC analysis. Dots represent analyzed direction of some frequency band. Near the center of the figure are lower frequencies and near the outer circumference are higher frequencies. The larger the dot is the smaller the diffuseness of that frequency band is. The horizontal plane is presented in the figure. Up means in front of listener, right means on the right side of the listener etc. In the left figure the sound source should be right in front of the listener. The sample was recorded in a typical room. In the right figure the sound source should be about 45 degrees to the right. The sample was recorded in an anechoic chamber. The input signal was speech in both cases.

5.2 DirAC Synthesis

Input signals are transformed from the time domain to the frequency domain in a similar way as in the analysis phase. The only difference is that windowing is done with a square-root Hann window. In the frequency domain different frequency bands are multiplied by different gains, due to for example dividing to diffuse and nondiffuse parts, see Fig. 4.4. Multiplying different frequencies with different scalars in the frequency domain corresponds to convolution in the time domain. The diffuseness has steep transitions in the frequency domain, so if the frequency domain gains of the diffuseness in Fig. 4.4 would be transformed into the time domain, the temporal structure of it would be broad. The length of the result of convolution is the sum of the lengths of the input signals minus one. The input signal is not zero-padded, so due to circularity of FFT, the end of the resulting signal is aliased to the beginning of it. This causes clearly noticeable clicks and artifacts. By windowing the signal with a square-root Hann window after processing in the frequency domain, these artifacts are muted to sound pressure levels that cannot be perceived.

Windowing a signal twice with the square-root Hann window corresponds to windowing the signal once with a Hann window. Windowing must be performed in a way that signal is not changed and perfect reconstruction is attained. The Hann window is symmetrical, so if the first half and the last half of it are summed, a constant level of unity is attained. In DirAC, signals are windowed with overlapping Hann windows, see Fig. 5.1.

5.2.1 Virtual Microphone Version of DirAC

The virtual microphone version of DirAC differs from the omnidirectional version of DirAC in transmitted signals. In the omnidirectional version the input signal for every virtual loudspeaker is the same omnidirectional signal. The resulting virtual loudspeaker signal is created from the omnidirectional signal by multiplying and filtering it. The virtual microphone version uses virtual cardioids for input instead of the omnidirectional signal. These cardioids are created for every virtual loudspeaker so the input signal for every virtual loudspeaker is different. A virtual cardioid is formed from a dipole and an omni-directional signal by summing them and dividing by two, see Fig. 5.3. The positive axis of the dipole must point to same direction as desired direction for the virtual cardioid. Desired dipole signal can be created as a composition of dipoles in x-, y- and z-directions

$$D(t) = \cos(\theta)\cos(\phi)X(t) + \sin(\theta)\cos(\phi)Y(t) + \sin(\phi)Z(t) \tag{5.2}$$

where $D(t)$ is the resulting dipole, $X(t), Y(t), Z(t)$ are input dipoles corresponding to Cartesian coordinate axes, θ is the angle in azimuth direction and ϕ is the angle in elevation direction. These virtual cardioids are calculated for every virtual loudspeaker. The direction of the virtual cardioid is same as the direction of the corresponding virtual loudspeaker.

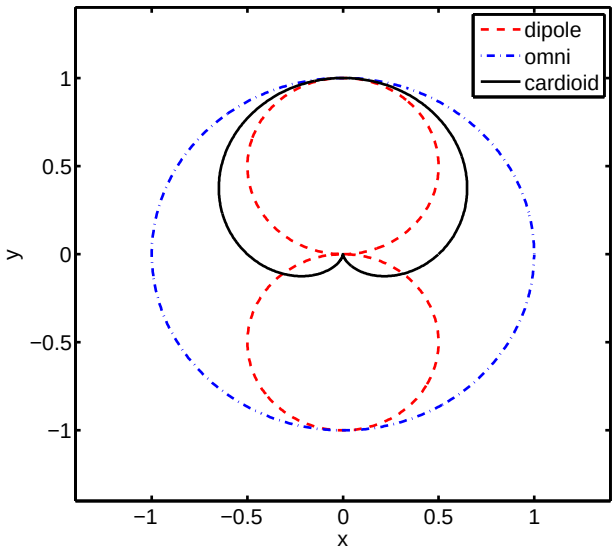

Figure 5.3: A cardioid is created by summing a dipole and an omni-directional signal and dividing by two. The lower lobe of the dipole in the figure has opposite phase than the upper lobe of it and the omni-directional signal.

Dipoles can also be used instead of virtual cardioids. An advantage of a dipole is that the directional pattern of the main lobe is narrower than with virtual cardioids. Thus signals for adjacent loudspeakers are more incoherent. A problem with dipoles is that the lobe to the oppo-

site direction is as large as the lobe to the desired direction. The directional pattern can also be something between a cardioid and a dipole, for example a supercardioid.

In the virtual microphone version the perceived directions of point sources are more stabile than in the omnidirectional version. Let us assume that there is one point source at 0 degrees in azimuth in the horizontal plane. Let us also assume that there are errors in the analysis for example due to another simultaneous source. The direction of the source fluctuates between 0 and 45 degrees. If virtual cardioids are used, reproducing sound from 45 degrees in azimuth uses the virtual cardioid of that direction. Since the sound source is in 0 degrees, sound is attenuated in the virtual cardioid pointing to 45 degrees. As a result, sound is reproduced more silent from other directions than the right direction. With minor errors in analyzed direction, reproduction is perceived to be more stabile than with the omnidirectional signal. If errors are major, the loudness of sound is perceived to vary in the virtual microphone version. In the omnidirectional version sound would not attenuate but instead it would be reproduced from a wrong direction. However, in most of the cases virtual cardioids produce better results than the omnidirectional version, because attenuation is not perceived and stability of the point sources is slightly better.

The major advantage of virtual microphones is in the synthesis of diffuse sound. This is covered in more detail in Section 5.2.4.

5.2.2 Synthesis of Nondiffuse Sound

The simplest way to reproduce nondiffuse sound is to use amplitude panning between two channels of headphones. It is computationally a really efficient way to position auditory objects. Amplitude panned sound reproduced using headphones generally gives the impression that the sound appears to come from inside the head. Sound images move along the axis between the ears. This is called the lateralization of the images [29]. This is the main problem using this approach, since sources can be panned only in the left-right dimension. Also, auditory events are inside the head, which does not correspond to real life situations.

Another approach is to use HRTF technique for nondiffuse sound. Using HRTFs, all possible directions in three dimensions can be synthesized. This was chosen as a reproduction method for the nondiffuse sound in this project. Synthesizing sound with HRTFs is computationally heavy. A problem with HRTFs is that it is impossible to measure them for every possible angle with infinite spatial resolution. HRTFs are measured with some practical resolution, for example 15 degrees spacing, and all other HRTFs are interpolated from them.

Third option would be to use some simplified model of HRTFs. One such model was presented in [42]. The effect of the head and the torso were modeled in it using a spherical head and torso. The HRTFs of this model are approximated by two time delays and two first-order IIRs.

Virtual Loudspeakers

The simplest way to interpolate HRTFs is to handle them as virtual loudspeakers, see Fig. 5.4. For every loudspeaker there is one pair of HRTFs. Output signals are created for loudspeakers and then filtered with corresponding HRTFs. Filtered outputs for every loudspeaker are summed to get signals for left and right ears. The advantage with this method is that it works well also with virtual cardioids in the virtual microphone version of DirAC, because the input signals to HRTF filtering are loudspeaker signals. In the virtual microphone version the virtual cardioids are created to correspond to the direction of virtual loudspeakers. These virtual cardioids are processed to get the loudspeakers signals.

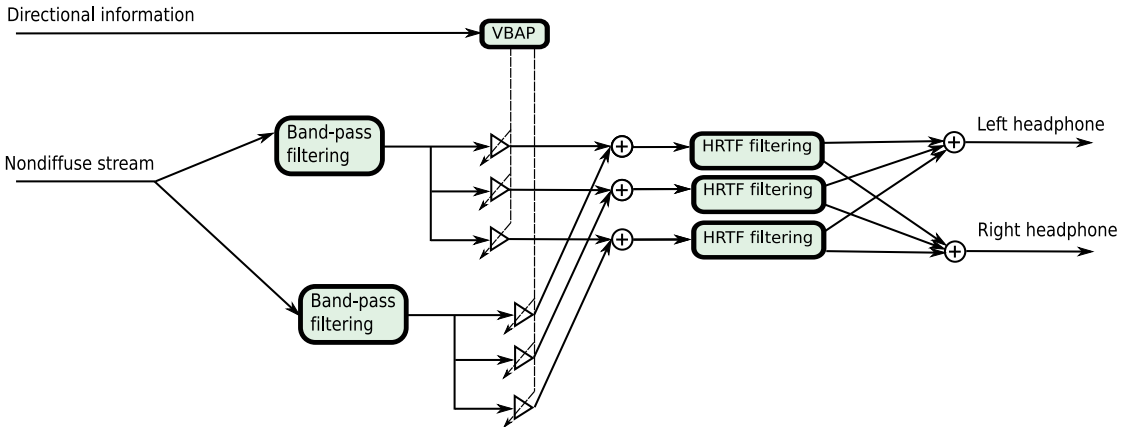

Figure 5.4: Synthesizing nondiffuse sound using virtual loudspeakers. In this figure there are only two frequency bands and three loudspeakers. Typically there are more frequency bands and loudspeakers.

The loudspeaker signals are created with amplitude panning. VBAP is used for amplitude panning. This corresponds to linear interpolation of impulse responses with three impulse responses. In the time domain this can cause theoretical problems, because impulses can arrive in different time instants due to different delays in different HRTFs. As a result there can be three main impulses instead of one. However, the directional resolution used was accurate enough that this was not considered to be a problem. The practical functionality was considered to be more important. Interpolating HRTFs with linear interpolation was studied in [43]. Interpolation is reported to cause for example localization inaccuracy and front-back confusions. The result in this study was that localization accuracy is largely unaffected by the interpolation, even for large interpolation intervals. Linear interpolation was also studied in [44]. In it interpolated magnitude frequency responses of HRTFs were compared to non-interpolated responses. From the results it can be said that interpolation errors are larger at the contralateral than at the ipsilateral side.

As presented in Section 5.1, direction and diffuseness change so rapidly that it causes clicks and artifacts. That is why this change must be averaged. The diffuseness was slowed down in the

analysis phase. The direction was not slowed down in the analysis phase because it would cause sluggishness to rapid changes of direction. This was studied more carefully in [1]. Changes of direction were slowed down using gain averaging. The idea is that the gain of every loudspeaker is smoothed separately. This was implemented using first order IIR filters. Time constants of IIR filters were selected separately for each frequency band. The time constant was chosen to correspond the cycle time of the center frequency of the band. This way averaging was done within the temporal accuracy of hearing. At lower frequencies the accuracy is worse and at higher frequencies it is better. The time constant was found to be optimal at $50 - 100$ times the cycle time. Artifacts were removed but the change of direction was not perceived to be slower than without averaging in most of the cases.

If the gains were directly smoothed, every calculated direction would have the same impact to the resulting direction. Then really silent sounds or sounds that have really high diffuseness could affect direction in a same way as sounds that are loud and nondiffuse. Hence the gains are weighted based on the energy and the diffuseness. The more energy and lower diffuseness the signal has, the more it contributes to the resulting direction. As a result, spatial images are more stable. This energy- and diffuseness-weighted averaging can also cause errors. Let us consider that there are two simultaneous speakers positioned to -45 and $+45$ degrees in azimuth in the horizontal plane. If the speakers are equally loud, both are positioned to correct directions. But, if another speaker is 20 dB louder, every now and then the more quiet speaker is panned to the same direction as the louder speaker. However, in most of the cases the energy weighting produces better results, so it was used.

Interpolating HRTFs

Another approach is to create one pair of transfer functions that position different frequency bands to different directions, see Fig. 5.5. Each frequency band might have different direction so for each frequency band HRTFs must be separately interpolated based on the direction of that frequency band. HRTFs for different frequency bands are bandpass-filtered to get the transfer function of only that band. These bandpass-filtered HRTFs are summed together to get the resulting HRTF. In these HRTFs different frequencies are heard from different directions.

An advantage of this approach is that HRTFs can be interpolated using any method desired. One simple method is to make a minimum-phase filter from the HRTF. This can be done by taking real part of the cepstrum of the HRTF. These minimum-phase filters are linearly interpolated. Correct ITD is added to correspond the ITD of the HRTF of the desired direction. This can be done by calculating the average from the impulse responses or by using Eq. (2.7). More sophisticated interpolation techniques are presented in literature, for example in [45].

In this approach the input signal is filtered with only one pair of HRTFs. With virtual loudspeakers approach the signal is filtered with every HRTF corresponding to virtual loudspeakers. Thus this method is computationally more efficient. Also directional accuracy can be better because of the better interpolation technique.

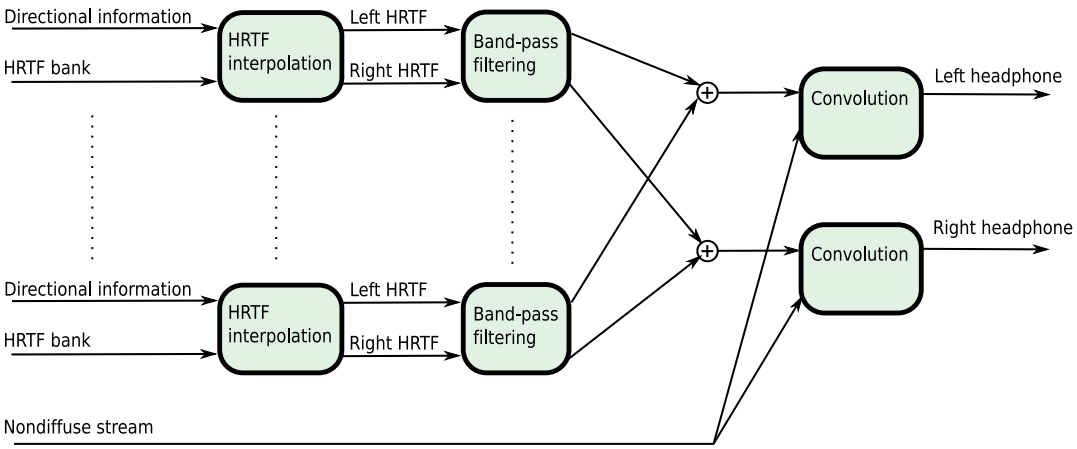

Figure 5.5: Synthesizing nondiffuse sound using interpolation of HRTFs. The interpolation of HRTFs is performed according to the direction of corresponding frequency band.

A problem is in creating a single HRTF from HRTFs of all frequency bands. Since adjacent frequency bands can point to opposite directions, the resulting HRTF can have really steep transitions in the magnitude response between different frequency bands. A corresponding situation can be seen in Fig. 4.4. These steep transitions make the impulse response long. The impulse response must be windowed to a reasonable length or transitions must be smoothed.

With the omnidirectional version this method works quite simply because the input signal is same for all directions. With the virtual mircophone version the input signal is different for different directions. The input signal for HRTF filtering must be constructed from different virtual cardioids. This is done in a similar way as constructing HRTFs in this method. Virtual cardioids are created to correspond to the direction of each frequency band. These virtual cardioid signals are bandpass-filtered to get the signal at that frequency band. These bandpass-filtered virtual cardioids are summed together to get the input signal for HRTF filtering. This signal has also steep transitions in the frequency domain. Windowing in the time domain must be done for example with a Hann window.

Comparing Methods

Virtual loudspeakers was chosen as a reproduction method for the non-diffuse sound. With it timbre of sound was good and no artifacts were present. The directional accuracy cannot be optimal because no advanced interpolation technique can be used. With the interpolating HRTFs method, better accuracy could be attained but problems with artifacts should be solved first. This is left for future study.

5.2.3 Synthesis of Diffuse Sound

Binaural reproduction has two outputs, signals to left and right ears. So the simplest way to create diffuse sound is to add decorrelated signals to each ear, see Fig. 5.6. These two signals cannot be used directly for diffuse sound, because the spectral shape of these signals is different from the spectral shape of the sound pressure at the ear canal caused by diffuse sound. The magnitude response of these two signals must be made similar as with HRTFs, see Fig. 3.8. This can be done by filtering with diffuse field HRTFs. The diffuse field HRTFs correspond to HRTFs measured in diffuse field. It does not position signal to any direction but makes timbre of sound similar as with HRTFs. The diffuse field HRTFs can be measured in an echoic room or by averaging HRTFs from different directions.

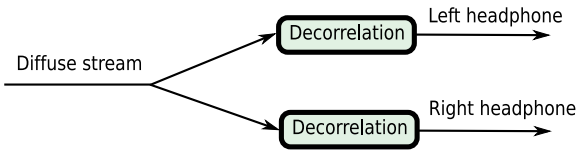

Figure 5.6: Synthesizing diffuse sound by decorrelating diffuse stream to the left and the right headphone.

Reproducing decorrelated signals with headphones was studied in [23]. The degree of coherence between signals to left and right ear was 0. Subjects were asked to describe the direction of an auditory event. The result was that two spatially separated auditory events appear, one at each ear, see Fig. 2.9. In case of DirAC, the results are similar. Decorrelated signals do not interfere direction perception of nondiffuse sound. Diffuse sound sounds pleasant and a feeling of spaciousness is reproduced. However, diffuse sound still is perceived to come from two distinct positions, which are the left headphone and the right headphone. Diffuse sound should be coming from every possible direction and it should surround the listener. So using this method the feeling that the listener would be in a certain space is not reproduced. By reproducing only partially decorralated signals with a coherence of 0.4, sound is perceived to fill the whole head, but the sound does not appear to be very diffuse but instead is perceived to be colored.

Another way is to create signals that are decorrelated and filter them with HRTFs, see Fig. 5.7. HRTFs should be selected so that they cover the whole sphere around the listener. This way diffuse sound does not appear to come from any distinct direction but rather from everywhere around the listener. This way a feeling of being in another space is reproduced.

The directions of HRTFs should be selected evenly from all directions. An adequate number of HRTFs used for diffuse sound is about $12 - 20$ in three dimensions. If fewer HRTFs are used, the diffuse sound can be localized to the directions of the HRTFs. The sensation in that case is that sound is coming from everywhere around the listener, but from distinct point sources. Also there must be HRTFs from every direction. Let us consider if all used HRTFs are measured in

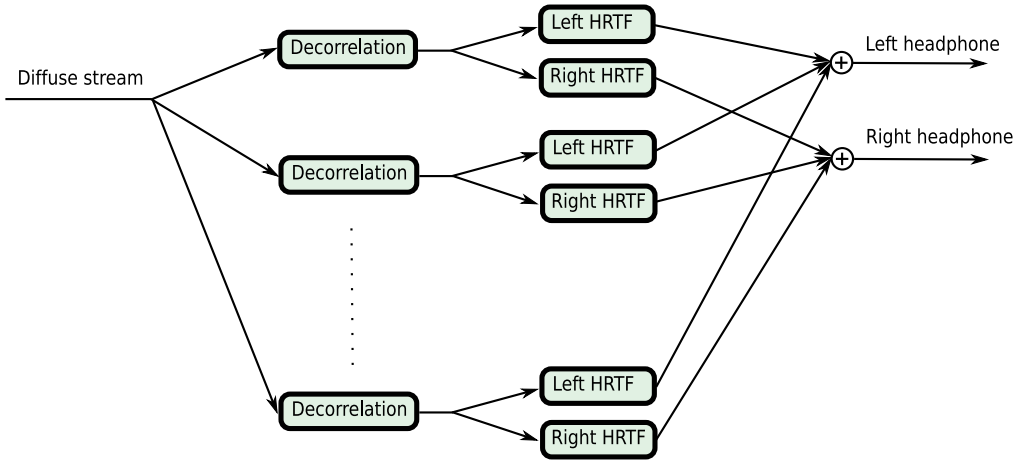

Figure 5.7: Synthesizing diffuse sound by creating decorrelated signals and filtering them with a few HRTFs. The signal is the same in the left and the right HRTF.

the horizontal plane. The direction of the diffuse sound cannot be localized to any direction, but it can be perceived that the diffuse sound is coming only from the horizontal plane.

Let us consider another example where the HRTFs are from three elevations, -50, 0 and $+50$ degrees. With some samples it can be perceived that there are three layers where diffuse sound is coming from. This happens especially with samples that have stationary diffuse sound. A good example of this is a sample recorded in a subway station while a subway train is coming to the station. The diffuse sound is slowly changing humming and noise coming from the train. With this kind of samples it can perceived if HRTFs are spatially too far from each other. If the diffuse sound is rapidly changing, for example a sample measured in a large crowd and consists of several people talking, it is much harder to localize diffuse sound to any distinct direction.

Using maximum possible number of HRTFs for diffuse sound is not feasible. Based on informal listening no better quality was gained after $12 - 20$ HRTFs. Decorrelating signals for each pair of HRTFs is computationally complex so minimum possible number of HRTFs is used.

Third method is similar to the previous one. HRTFs are used in the same way, but signals to the left and the right HRTF of certain direction are incoherent, see Fig. 5.8. In the previous method signals to the left and the right HRTF of a certain direction are always the same. The advantage of the third method is that even though signals would not be properly decorrelated, the resulting diffuse sound is perceived to be more diffuse than with normal usage of HRTFs. The disadvantage is that most of the sound seems to be coming from headphones, similarly as in the first case. Some other directions are also perceived. Perception is something between the first two methods.

In informal testing the best results were obtained using the second method. It worked satisfactoryly in all circumstances.

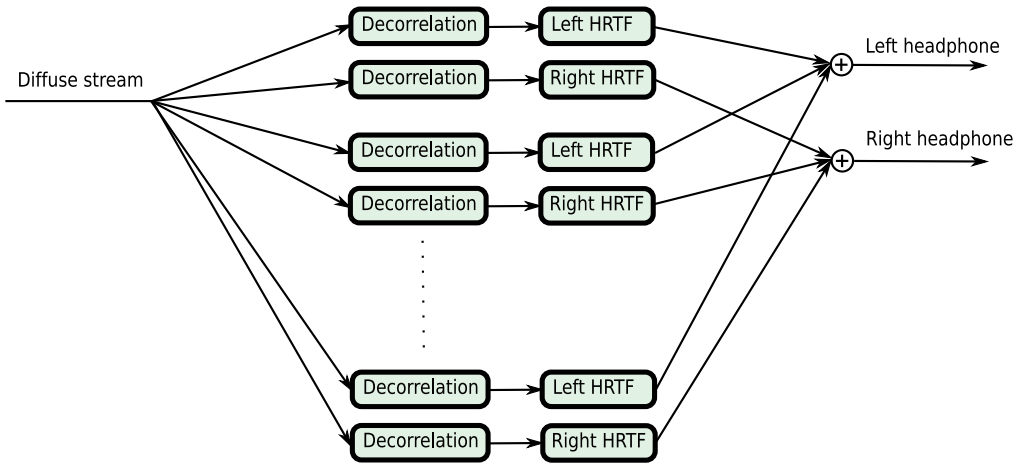

Figure 5.8: Synthesizing diffuse sound by creating decorrelated signals and filtering them with a few HRTFs. The signal is differently decorrelated for the left and the right HRTF.

5.2.4 Decorrelation Method

In the previous section it was explained, how the synthesis of diffuse sound is performed, but the decorrelation method was not explained. Decorrelation is discussed in this section. If correlated signals were used as diffuse signals and positioned using HRTFs, the result would be an auditory event inside the head. That is why proper decorrelation is needed. Decorrelation is not necessarily done in a mathematical sense, but instead in a psychoacoustical sense.

The simplest way to decorrelate signals is to convolve them with a noise burst. The noise should have a magnitude response of unity but a random phase response. The noise can have constant amplitude in the time domain or it can be exponentially decaying. The length of the noise must be short enough that it does not make impulse-like sounds perceptually longer. On the other hand a short noise burst does not decorrelate low frequencies properly. An advantage of exponentially decaying noise is that it can be longer without making impulse-like sounds to be perceived longer. A disadvantage of it is that because of the shape of the burst, low frequencies are not really random. The phase of low frequencies is 0 or 180 degrees. The more steep the decaying is, the less random is the phase. With a long burst sufficient decorrelation was obtained, but artifacts emerged and with a shorter noise burst signals were not decorrelated.

Better results were obtained using random delays, see for example [46]. The diffuse sound is divided into frequency bands that correspond to the critical bands, similarly as with DirAC processing. Every frequency band is given a random delay. The frequency bands are summed together. The result is a signal that has random delays at every frequency band but the magnitude response has not been changed. Let us consider reproduction of these decorrelated signals from multiple (virtual) loudspeakers. If we consider one frequency band, sound is positioned to a loudspeaker that has the shortest delay at that frequency band due to the precedence effect, see

Section 2.5.4. Every frequency band is panned to a random loudspeaker. As a result the direction of a broadband signal cannot be determined but instead sound is perceived as diffuse. The lower limit of delays should not be zero but rather a few milliseconds. This way the diffuse sound does not interfere the perception of direction. The nondiffuse sound comes always first to ears and thus determines the direction of sound.

At the borders of two frequency bands there can be dips in the magnitude response. This is caused by the difference in phase in that frequency and thus destructive interference. Let us consider only one frequency band. When differently decorrelated signals are reproduced from multiple loudspeaker and measured in one point, the result is similar as randomly delayed signals would be summed together. This corresponds to having an FIR filter that has randomly spaced impulses. The magnitude response of that kind of filter is not flat. However, fluctuation in the frequency response is dense and in most of the critical bands peaks and dips fall in the same critical band. Hence the result is not perceived to be colored. Also delays are slightly different between different ears. Thus filtering to different ears is different. This also makes sound less colored. This was tested by playing same noise from two directions using loudspeakers. Another noise had delay in it. When listening by muffling another ear, sound was perceived to be more colored than when listening normally using both ears.

The delays cannot be arbitrary long. If they are within eligible boundaries they do not contribute to the perception of direction or spaciousness. If the upper limits of delays are increased, sound is perceived to become more spacious. If delays are lengthened even more, distinct echoes are heard. This is in line with the precedence effect. Also it was noticed that low frequency bands can have more delay than high frequencies. On the other hand, if delays are limited to be too short, the signal is not decorrelated. So the upper limit of the delays is a compromise between time domain artifacts and decorrelation.

At low frequencies, below 1500 Hz, the perception of direction is based on the inter-aural phase differences, see Section 2.5.1. Thus making the phase random makes signal decorrelated. The fluctuation range of the delay must be long enough that all phases are possible. Thus the range is determined by the longest cycle time in that frequency band. Above 1500 Hz perception of direction is based on level differences. Making phase random does not cause signal to be perceived decorrelated. This can be tested by playing noise signals that have opposite phase using headphones. At low frequencies there are two auditory objects, one at each ear, but at high frequencies one auditory object is positioned in the middle of the head. The delays must be long enough that there is fluctuation in the ILD and the direction of sound cannot be perceived. Formal testing of this is left for future study.

By informal testing, boundaries were found for delays at low and high frequencies. At low frequencies the fluctuation range of delays should be about 80 ms and at high frequencies the fluctuation range of delays should be about 30 ms. For mid frequencies values are interpolated. Using these delays, no artifacts were present. Decorrelation is not perfect, but it was found adequate.

Random delays at different frequency bands can be implemented as an FIR filter. The filter consists of FIR band-pass filters. Passbands of the filters correspond to critical bands of hearing. These filters are given a random delay and then summed together. Decorrelation can be added to a signal by simply convolving the filter and the signal.

Using virtual cardioids improves decorrelation. Let us consider two virtual cardioids pointing to opposite directions in a free field. Sounds arriving from the different directions do not have high coherence in typical situations. For example in rooms most of the sound is from reverberation, at least in the far field. The reflections measured in a single point are not most likely in phase. Thus signal is decorrelated. Only if a sound sample would be recorded in an anechoic room using two loudspeakers that are in the opposite directions at the same distance playing same sound signals that are identical in phase, the signals would have correlation between them. However this case is really rare, and it can be said that signals from different directions are not correlated.

As can be seen from Fig. 5.3, the directional pattern of a cardioid is quite wide. Two virtual cardioids are always overlapping. Least correlation is in the cardioids pointing to opposite directions. The closer to same direction cardioids are pointing, the more correlated they become.

In DirAC, diffuse sound is reproduced from $12 - 20$ directions in three dimensions. Hence distance between virtual loudspeakers is about 60 degrees. The cardioids are only partly overlapping, see Fig. 5.9. Thus signal is partly decorrelated. Decorrelation algorithm does not need to decorrelate signals completely since they are already partly decorrelated. The length of the delays can be shortened and artifacts in the time domain can be avoided. Using dipoles instead of cardioids, even better decorrelation can be obtained. The directional pattern of a dipole is narrower than the directional pattern of a cardioid. Thus adjacent dipoles are less correlated and better decorrelation is obtained. The dipole has a large back lobe that can cause problems with some samples. Most robust results were gained using a supercardioid. It has a slightly narrower main lobe than a cardioid but smaller back lobe than a dipole.

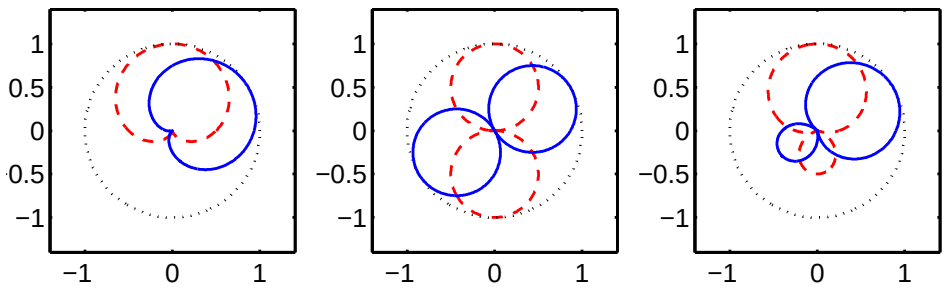

Figure 5.9: Overlapping directional patterns pointing to 0 degrees and 60 degrees. On the left cardioids, in the middle dipoles and on the right supercardioids.

5.3 Head Tracking

Head tracking means tracking the position of the head. In this thesis only the orientation of the head is of interest, not the absolute position. When using headphones, by head tracking an auditory event can be positioned to the same direction even though a listener moves his/her head. Let us consider that there is an auditory event right in front of the listener, see Fig. 5.10. If the listener moves his/her head, the auditory event moves with the movement of the head. This does not correspond to a situation in normal life where an auditory event does not move with the head movement. With head tracking the position of the auditory event can be kept as constant. This way more realistic reproduction of spatial sound can be achieved. Also head tracking reduces front-back confusions as mentioned in Section 3.4.

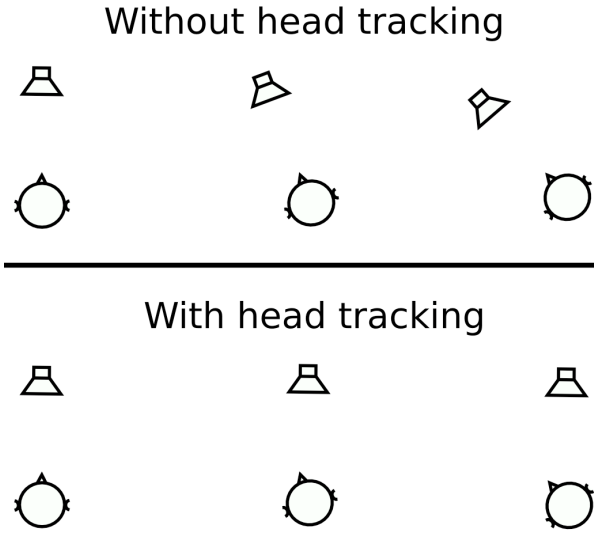

Figure 5.10: The effect of head tracking when listening with headphones. A loudspeaker represents the direction of the perceived auditory event. Without head tracking the auditory event moves with the movement of the head. With head tracking the direction of the auditory event can be kept as constant.

5.3.1 Angles to Specify Orientation of Head

The orientation of the head can be specified using three angles: azimuth, elevation and tilt, see Fig. 5.11. Azimuth corresponds to rotating the head in azimuth direction, looking to left and right. Elevation corresponds to rotating the head in elevation direction, looking up and down. When moving the head in tilt direction, the face is facing to same direction but the head is rotating. When rotating 90 degrees to right in tilt direction, left ear is pointing up and right ear is pointing down.

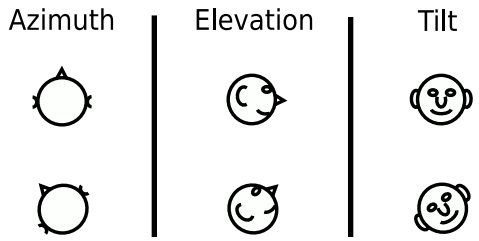

Figure 5.11: Moving the head in different directions.

5.3.2 Updating Direction Calculation in DirAC to Match Head Movement

In DirAC the direction of sound is determined by the intensity vector. The nondiffuse sound is synthesized based on direction information calculated in the analysis phase. Let us assume that a listener is looking to 0 degrees in azimuth and an auditory event is positioned to 0 degrees in azimuth. The listener turns his/her head to -45 in azimuth. Relative to the listener, the auditory event is now in $+45$ in azimuth in order to stay in the same place relative to world. Hence, the directional information is corrected by shifting all calculated azimuth angles by $+45$ degrees. New azimuth is calculated by

$$\theta_2 = \theta_1 - \alpha \tag{5.3}$$

where θ_2 is new azimuth, θ_1 is old azimuth and α is head azimuth.

Head elevation and tilt are updated in a similar way. However changing them changes both azimuth and elevation. That is why calculation of new angles is done in the Cartesian coordinate system. Cartesian coordinates are calculated from spherical coordinates by

$$\begin{aligned} x &= \cos(\theta)\cos(\phi) \\ y &= \sin(\theta)\cos(\phi) \\ z &= \sin(\phi) \end{aligned} \tag{5.4}$$

where θ is azimuth and ϕ is elevation.

The effect of head elevation changes are calculated by

$$\begin{aligned} x_2 &= \cos\Big(-\beta + \arctan\Big(\frac{z_1}{x_1}\Big)\Big) \cdot \sqrt{1 - y_1^2} \\ y_2 &= y_1 \\ z_2 &= \cos\Big(-\frac{\pi}{2} - \beta + \arctan\Big(\frac{z_1}{x_1}\Big)\Big) \cdot \sqrt{1 - y_1^2} \end{aligned} \tag{5.5}$$

where β is head elevation. The subindices 2 mean the new coordinates and the subindices 1 mean the old coordinates. The effect of head tilt changes are calculated by

$$\begin{aligned} x_2 &= x_1 \\ y_2 &= \cos\Big(-\gamma + \arctan\Big(\frac{z_1}{y_1}\Big)\Big) \cdot \sqrt{1 - x_1^2} \\ z_2 &= \cos\Big(-\frac{\pi}{2} - \gamma + \arctan\Big(\frac{z_1}{y_1}\Big)\Big) \cdot \sqrt{1 - x_1^2} \end{aligned} \tag{5.6}$$

where γ is head tilt. Cartesian coordinates are changed back to spherical coordinates by

$$\begin{aligned} \theta &= \arctan\left(\frac{y}{x}\right) \\ \phi &= \arctan\left(\frac{z}{\sqrt{x^2+y^2}}\right). \end{aligned} \tag{5.7}$$

In the equations above, function $\arctan$ gives two answer. The correct one must be chosen according to the sign of the numerator and the denominator.

In the virtual microphone implementation also virtual cardioids must be created according to the head position. This can be done by creating new X-, Y- and Z-dipoles that correspond to the orientation of the head. Virtual cardioids can be calculated from these dipoles. Equations for calculating new dipoles are presented in Appendix A.

The diffuse sound does not need changes due to head tracking. Angles from where it is reproduced change relative to the world, but it is reproduced from every direction so changing the orientation of the head does not affect it.

5.3.3 Implementation of Head Tracking

Head tracking can be implemented using any desired method. The only requirement is that the head orientation is given using azimuth, elevation and tilt.

In this project head tracking was implemented using sound intensity calculation. Let us first consider head tracking only in the azimuth direction. The idea is similar as in the direction analysis of DirAC. A small-size B-format microphone is attached to the head. This kind of microphone can be composed from four omnidirectional microphones as presented in Section 4.2. A sound source is placed into a fixed location reproducing continuous signal, for example noise. The direction of this sound source relative to the head can be calculated from B-format signals using methods presented in Section 4.3.1. This is done exactly in the same way as in the DirAC analysis. Let us assume that the sound source is right in front of the listener. If the sound source is analyzed to be 45 degrees to the left, the listener has rotated his/her head 45 degrees to the right. Angles in azimuth direction are calculated as simply as that.

Of course playing noise from loudspeakers while listening to something using headphones is not a good idea. Headphones typically do not attenuate background noise enough so that it would not be heard. A solution is to use ultrasounds for head tracking. A sound source can reproduce for example high-pass filtered noise which has a cut-off frequency of about 18 kHz. Normally adults are not able to hear anything above 17 kHz even though theoretically humans can hear up to 20 kHz. This cut-off frequency must be selected carefully so that listeners are not able to hear this high-pass filtered noise. The direction is calculated from the reproduced band, for example 18 - 20 kHz. The upper limit for the band is the bandwidth of the loudspeaker and the microphone.

Also elevation and tilt are needed to be tracked for better quality. The B-format microphone must work in three dimensions in this case. A three-dimensional B-format microphone can be

created in a similar way as in Section 4.2, but six omnidirectional microphones are needed. Two opposing microphones are attached to create a dipole in the direction of the z-coordinate. Two sound sources are needed instead of one to calculate azimuth, elevation and tilt. These two sound sources should be positioned for example to 0 degrees and 90 degrees in the horizontal plane.

These two sound sources must reproduce sound at different frequencies in order not to confuse direction analysis of each other. For example reproducing band-pass filtered noise between 18 and 19 kHz from one loudspeaker and between 19 and 20 kHz from another. The analysis of the direction is done separately for each band. The result is directions of two sound sources relative to the head. Positions of these sound sources are known so they act as anchor points. Head orientation is calculated from these angles.

A direct analytic solution was not found, hence a heuristic solution was used. Calculation of head orientation was done in a Cartesian coordinate system. Cartesian coordinates of the directions of the anchor points are calculated using Eq. (5.4). Two other points are added to the horizontal plane to 0 degrees in azimuth and to 90 degrees in azimuth. These two points correspond to the orientation of the head, let us call them orientation points. The position of these four points are locked relative to each other. These points are rotated to get the first anchor point to 0 degrees in the horizontal plane and the second anchor point to 90 degrees in the horizontal plane, positions of the anchor points when looking straight ahead. The orientation points are rotated with them. The result is that orientation point one points to the direction of where the head is facing and second orientation point points to side of the head. Azimuth and elevation are calculated from the direction of the first orientation point and tilt is calculated from the second orientation point.

The calculated orientation of the head can have rapidly changing fluctuation. This is caused from non-ideal microphones. Calculated head azimuth, elevation and tilt can be smoothed to get rid of this fluctuation. A first order IIR filter was used with a time constant of about 100 ms. It reduced fluctuation and tracking responded to the movement of the head without significant sluggishness. With a longer time constant there was a lag in the orientation and with a shorter time constant there was too much fluctuation.

Head tracking using this method seemed to work with binaural reproduction of DirAC. Directional accuracy was in the region of 3 - 10 degrees and fluctuation was a few degrees, which is far from perfect. However this seemed to be enough in most cases. Typically there is some movement in real-life recordings, so a small change in the direction is not perceived. Also analysis of real-life recordings has significant fluctuation as was seen in Fig. 5.2. This fluctuation is smoothed in the synthesis phase. The smoothing of the direction caused some sluggishness to the direction. However typically people do not rotate their head very fast, so this sluggishness did not cause significant problems.

Chapter 6

Results

The results of this project are presented in this chapter. Formal listening tests were not carried out. The reason for this was that listening tests were wanted to be carried out with a high quality head tracking system. This kind of system was not available at the moment of writing this thesis, but it was under consideration to be purchased. Hence formal listening tests were postponed to the future.

6.1 Implemented Systems

The thesis work consists of three contributions:

- design and construction of HRTF measurement system
- design and implementation of binaural reproduction for DirAC
- design and testing of DirAC-based head-tracking system.

6.2 Subjective Quality of Reproduction

Results presented here are based on subjective considerations by author and other people working with Directional Audio Coding. The quality of binaural reproduction of DirAC was evaluated using two kind of samples.

The first group is real-life recordings. These were recorded using SoundField ST350 B-format microphone. The recordings include for example ice-hockey game on arena, floorball match, traffic in downtown Helsinki, a grocery store, a marketplace, a cafe, inside a train, a subway station and a school. A problem with these samples is that there is no reference sample to which binaural reproduction of DirAC could be compared. Auditory memory is quite short so comparing these samples to reality is quite hard. It cannot be said if binaural reproduction of DirAC is different from how these events really sounded. However these situations are quite familiar to everyone of us, so we have a strong intuition how these kind situations might sound.

Another type of samples was simulated spaces. These simulated spaces were created using the image source method, see [47]. Reflections were modeled up to fifth order. Late reverberation was modeled using a decaying noise. B-format signals where calculated from simulated spaces. These signals were used with DirAC. The advantage of simulated spaces is that a virtual reference can be created. Signals were created for 21 loudspeakers. This was done by dividing the space into sections according to the loudspeakers and sound only from that section was reproduced using the corresponding loudspeaker. In real life, creating such a reference is not possible since there are no microphones that would have so narrow directional pattern. The signals to the loudspeakers were filtered with HRTFs that were measured from the directions of the loudspeakers. Comparison between this virtual binaural reference and binaural reproduction by DirAC was done in real-time. There were three spaces modeled, a small room, a medium size auditorium and a large concert hall. Singing, a symphony orchestra, snare drum playing and rock music were reproduced in these spaces. Rock music was created by reproducing a stereo recording using two loudspeakers that were positioned into the virtual space.

Binaural reproduction was also compared to the loudspeaker reproduction of DirAC and the loudspeaker reproduction of the virtual reference.

The overall result was that the virtual microphone version of binaural reproduction of DirAC worked well compared to the virtual binaural reference. A feeling of being in another space was reproduced well and the space was perceived to be similar as in the binaural reference. In singing samples it was almost impossible to notice any differences. In symphony orchestra sample the timbre was slightly different in some parts. In rock music sample in a small room there was a slight difference in the direction of certain instruments. The snare drum was the most critical sample. The perception of space and timbre were clearly different from the reference in this sample. Real life recordings sounded realistic. Ambient sounds and noise sounded as they could be in real life.

The omnidirectional version of DirAC worked slightly worse than the virtual microphone version. Space did not sound so natural as with the virtual microphone version. It sounded slightly colored, which could be described as metallic. This was noticed especially in real life recordings that had lots of ambient noise, like for example in marketplace, traffic and cafe samples. With the virtual microphone version this ambient noise sounded very pleasant and it surrounded the listener very nicely. With the omnidirectional version this noise did not sound very real or pleasant. Also with the virtual microphone version sound was clearer and brighter.

However, this was expected, since also in the loudspeaker version of DirAC the virtual microphone version works better than the omnidirectional version, see [1]. Problems mentioned with the virtual microphone version were also present in the omnidirectional version.

The main problem with binaural reproduction is externalization of sound sources. Sound sources are not perceived to be very distant. Especially frontal sources are not properly externalized and front-back confusion was reported occasionally. These problems are the same as typically with HRTFs. Externalization with binaural DirAC was not worse than with binaural

reference. Some people even said that externalization worked sometimes better with binaural reproduction of DirAC. Reasons for this are not known.

Even though the externalization did not work very well, the sensation of "being there" was reproduced well. The head tracking was tested only briefly, but using it seemed to enhance this sensation.

6.3 Using Binaural Reproduction of DirAC in Tele-operation of Remote Vehicles

DirAC encoding and decoding were implemented as real-time MAX/MSP external objects, which are also in use in the Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI). They are intending to use the programs in tele-operation of a remote vehicle application. In that project the idea was to be able to control a vehicle remotely. An electric wheelchair was used as a vehicle in this case. The reason for this was that electric wheelchairs are easily obtainable and fairly priced. This wheelchair was remotely controlled. For visual input it had two video cameras. Using them stereoscopic vision was obtained. Sound was recorded using a B-format microphone. Sound was DirAC-encoded and reproduced using binaural reproduction of DirAC. Sound and video image were transmitted using a WiFi connection to a control room, where the operator of the vehicle was able to perceive the environment.

An advantage of DirAC is that only one audio channel and a low bit rate side channel consisting of metadata needs to be transmitted. However, the directions of sound sources can be reproduced. Let us consider driving a car. Visual input is naturally the most important sense, but sound input is also very important. We can see only to one direction at a time but we can listen to every direction. Decisions are made using visual input, but hearing can be considered as a warning system.

Chapter 7

Conclusions and Future Work

7.1 Summary and Conclusions

A system for measuring head related transfer functions (HRTF) was constructed. The main part of the system is an acoustically transparent boom that rotates both in azimuth and in elevation. The system is operated using a MAX/MSP software. Using this system, HRTFs were measured from a few persons. The theory of HRTFs was viewed, and methods to obtain HRTFs from these measurements were presented.

Binaural reproduction for Directional Audio Coding (DirAC) was implemented. It is based on the loudspeaker version of DirAC. The principles of the loudspeaker version of DirAC were explained. The binaural reproduction uses HRTFs that were measured using the measuring system presented in this thesis.

In DirAC, sound is captured with multiple microphones. Each microphone channel is divided into frequency bands. The analysis and the synthesis of sound is performed separately for every frequency band. The diffuseness and the direction of sound is computed in the analysis phase, which are used in the synthesis. In the synthesis phase, sound is divided into nondiffuse and diffuse streams. The nondiffuse sound in the binaural reproduction of DirAC was reproduced as point-like sources. They were implemented using virtual loudspeakers. Signals to the virtual loudspeakers were created using vector base amplitude panning (VBAP). These signals were filtered using corresponding HRTFs.

The aim of the diffuse sound is to produce a perception of surrounding sound lacking prominent direction. The diffuse sound was produced by filtering differently decorrelated versions of the signal by a few HRTFs. HRTFs were chosen so that they covered the whole sphere around the listener. The signals were decorrelated by adding random delays to the different frequency bands.

A head tracking system was composed in this study, which used the directional analysis of DirAC. A sound source is transmitting ultrasonic noise, which acts as an acoustic beacon. A B-format microphone is attached to the head of a listener. The direction of the beacon is computed

from the B-format signals, from which the orientation of the head can be derived. The head tracking system was briefly tested with binaural DirAC. The head tracking seemed to have some fluctuation in the calculated direction, but it seemed to work sufficiently well with DirAC.

In informal tests, the binaural reproduction of DirAC was found to provide good spatial and overall quality. A sensation of being in some other space was reproduced, and the directions of sound sources were perceived naturally. Problems included that sound sources were not properly externalized and front-back confusions were occasionally perceived. The same problems are present in the HRTF techniques in general.

7.2 Future Work

Formal listening tests should be arranged. The quality of the reproduction would be found out. Also ideas how to make the reproduction better could be obtained.

The synthesis of nondiffuse sound should be studied more carefully. As was presented in the thesis, better interpolation techniques could be used. Also composition of HRTFs that reproduce sound at different frequency bands from different directions could be investigated. The decorrelation method used at the moment decorrelates signals properly, but it would be better if it would be shorter in the time domain with at least similar amount of decorrelation. Temporal smearing would be decreased and decorrelation would be computationally less complex.

Binaural reproduction of DirAC could be tested with visual input. This could be done for example by using three-dimensional visual displays. This way the interaction between visual and sound inputs could be studied. The effect of using head tracking should be studied also.

Bibliography

[1] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, pp. 503–516, June 2007.

[2] "The UW-L symphony orchestra at Annett Recital Hall." University of Wisconsin - La Crosse - Music Department, http://www.uwlax.edu/Music/orchestra.jpg, April 2008.

[3] "Soundfield ST350 portable microphone system." Soundfield, http://www.soundfield.com/products/st350.php, April 2008.

[4] T. D. Rossing, F. R. Moore, and P. A. Wheeler, *The Science of Sound*. Addison Wesley, 2002.

[5] F. J. Fahy, *Sound Intensity*. Elsevier Science Publishers Ltd, 1989.

[6] J. Merimaa and V. Pulkki, "Spatial impulse response rendering 1: Analysis and synthesis," *J. Audio Eng. Soc.*, vol. 53, pp. 1115–1127, December 2005.

[7] L. E. Kinsler and A. R. Frey, *Fundamentals of Acoustics*. John Wiley and Sons, Inc., 1950.

[8] M. Karjalainen and H. Järveläinen, "More about this reverberation science: Perceptually good late reverberation," *AES 111th Convention*, New York, U.S.A., September 2001.

[9] T. Lahti, *Akustinen mittaustekniikka*. Libella Oy, 1995.

[10] F. A. Everest, *Master Handbook of Acoustics*. McGraw-Hill Book Company, 2001.

[11] M. Barron, *Auditorium Acoustics and Architectural Design*. E and F Spon, London, 1998.

[12] W. L. Gulick, G. A. Gescheider, and R. D. Frisina, *Hearing - Physiological Acoustics, Neural Coding, and Psychoacoustics*. Oxford University Press, 1989.

[13] W. Nienstadt, O. Hänninen, A. Arstila, and S.-E. Björkqvist, *Ihmisen fysiologia ja anatomia*. WSOY, 1997.

[14] M. Karjalainen, *Kommunikaatioakustiikka*. Libella Oy, 1999.

[15] G. von Bekesy, *Experiments in Hearing*. McGraw-Hill Book Company, 1960.

[16] L. Golipour and S. Gazor, "Evaluating the basilar membrane displacement for speech stimulus: A computational algorithm," *IEEE CCECE/CCGEI*, Saskatoon, Canada, May 2005.

[17] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 1982.

[18] A. N. S. Institute, "USA standard acoustical terminology," 1960.

[19] D. M. Green, "Temporal acuity as a function of frequency," *J. Acoust. Soc. Am.*, vol. 54, no. 2, pp. 373–379, 1973.

[20] W. A. Yost and M. J. Moore, "Temporal changes in a complex spectral profile," *J. Acoust. Soc. Am.*, vol. 81, pp. 1896–1905, June 1987.

[21] A. Goldberg, "Measuring the threshold of audibility of temporal decays," *AES 120th Convention*, Paris, France, 2006.

[22] J. Blauert, ed., *Communications Acoustics*. Springer, 2005.

[23] J. Blauert, *Spatial Hearing*. The MIT Press, 1983.

[24] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *J. Acoust. Soc. Am.*, vol. 82, pp. 157–167, July 1977.

[25] L. A. Jeffress, "A place theory of sound localization," *Journal of Comparative Physiology and Psychology*, vol. 41, no. 1, pp. 35–39, 1948.

[26] J. Blauert, "Spatial mapping of intracranial auditory events for various degrees of interaural cohorence," *J. Acoust. Soc. Am.*, vol. 79, pp. 806–813, March 1986.

[27] J. Merimaa, *Analysis, Synthesis, and Perception of Spatial Sound - Binaural Localization Modelling and Multichannel Loudspeaker Reproduction*. PhD thesis, TKK, 2006.

[28] R. Mason, F. Rumsey, and B. de Bruyn, "An investigation of interaural time difference fluctuations, part 1: the subjective spatial effect of fluctiations delivered over headphones," *AES 110th Convention*, Amsterdam, The Netherlands, May 2001.

[29] B. C. J. Moore, *Hearing*. Academic Press, 1995.

[30] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *J. Audio Eng. Soc.*, vol. 43, pp. 300–321, May 1995.

[31] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Transfer characteristics of headphones measured on human ears," *J. Audio Eng. Soc.*, vol. 43, pp. 203–217, January 1995.

[32] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992.

[33] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," *J. Audio Eng. Soc*, vol. 49, pp. 443–471, June 2001.

[34] A. Farina, "Advancements in impulse response measurements by sine sweeps," *AES 122th Convention*, Vienna, Austria, 2007.

[35] F. E. Toole and S. E. Olive, "The modification of timbre by resonances: Perception and measurement," *J. Audio Eng. Soc.*, vol. 36, pp. 122–142, March 1988.

[36] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head-tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *AES 108th Convention*, Paris, France, 2000.

[37] V. Pulkki and J. Merimaa, "Spatial impulse response rendering 2: Reproduction of diffuse sound and listening tests," *J. Audio Eng. Soc.*, vol. 54, pp. 3–20, January/February 2006.

[38] J. Ahonen, V. Pulkki, and T. Lokki, "Teleconference application and B-format microphone array for directional audio coding," *AES 30th International Conference*, Saariselkä, Finland, 2007.

[39] K. Brandenburg, "MP3 and AAC explained," *AES 17th International Conference*, Florence, Italy, 1999.

[40] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, pp. 456–466, June 1997.

[41] V. Pulkki, *Spatial Sound Generation and Perception by Amplitude Panning Techniques*. PhD thesis, TKK, 2001.

[42] V. R. Algazi, R. O. Duda, and D. M. Thomson, "The use of head-and-torso models for improved spatial sound synthesis," *AES 113th Convention*, Los Angeles, U.S.A., October 2002.

[43] E. M. Wenzel and S. H. Foster, "Perceptual consequences of interpolating head-related transfer functions during spatial synthesis," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, U.S.A., 1993.

[44] F. Christensen, H. Møller, P. Minnaar, J. Plogsties, and S. K. Olesen, "Interpolating between head-related transfer functions measured with low directional resolution," *AES 107th Convention*, New York, U.S.A., 1999.

[45] F. Keyrouz and K. Diepold, "Efficient state-space rational interpolation of HRTFs," *AES 28th International Conference*, Piteå, Sweden, 2006.

[46] M. Bouéri and C. Kyirakakis, "Audio signal decorrelation based on a critical band approach," *AES 117th Convention*, San Francisco, U.S.A., 2004.

[47] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *J. Audio Eng. Soc.*, vol. 47, pp. 675–705, September 1999.

Appendix A

Equations for Dynamic Computation of Dipole Signals for Head Tracking

When using head tracking, dipoles must be updated to point to the right directions based on the orientation of head. New dipoles can be calculated using

$$X_2(t) = \cos(\alpha)\cos(\beta)X_1(t) + \sin(\alpha)\cos(\beta)Y_1(t) + \sin(\beta)Z_1(t) \tag{A.1}$$

$$\begin{aligned} Y_2(t) = &- \Big[\sin(\alpha)\cos(\gamma) + \cos(\alpha)\sin(\beta)\sin(\gamma)\Big]X_1(t) \\ &+ \Big[\cos(\alpha)\cos(\gamma) - \sin(\alpha)\sin(\beta)\sin(\gamma)\Big]Y_1(t) \\ &+ \Big[\cos(\beta)\sin(\gamma)\Big]Z_1(t) \end{aligned} \tag{A.2}$$

$$\begin{aligned} Z_2(t) = &+ \Big[\sin(\alpha)\sin(\gamma) - \cos(\alpha)\sin(\beta)\cos(\gamma)\Big]X_1(t) \\ &- \Big[\cos(\alpha)\sin(\gamma) + \sin(\alpha)\sin(\beta)\cos(\gamma)\Big]Y_1(t) \\ &+ \Big[\cos(\beta)\cos(\gamma)\Big]Z_1(t) \end{aligned} \tag{A.3}$$

where α is azimuth angle of the head, β elevation and γ tilt. Subindex 2 means new dipoles and subindex 1 means old dipoles.