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

Juha Vilkamo

Spatial Sound Reproduction with Frequency Band Processing of B-format Audio Signals

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

Espoo, May 28, 2008

Supervisor: Professor Matti Karjalainen
Instructor: Docent Ville Pulkki

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Juha Vilkamo
Name of the thesis:	Spatial Sound Reproduction with Frequency Band Processing of B-format Audio Signals
Date:	May 28, 2008 **Number of pages:** 67
Faculty:	Electronics, Communications and Automation
Professorship:	S-89
Supervisor:	Professor Matti Karjalainen
Instructor:	Docent Ville Pulkki

The increase of knowledge in the field of spatial hearing has given birth to various spatial audio reproduction technologies. These include efficient perceptual coding of multi-channel audio, channel conversion technologies and universal audio formats with no restrictions to any specific loudspeaker setup. Directional Audio Coding (DirAC) extends the scope of universal audio reproduction to real sound environments by utilizing existing microphones for analysis and arbitrary loudspeaker setups for synthesis of the perceptually relevant properties of the sound field.

The human spatial hearing functions on the basis of multitude of cues. These cues range from the differences of the sound reaching both ears to the multimodal cues such as the visual cues. The goal of DirAC is to measure and synthesize those sound field properties by the influence of which the auditory cues arise, leaving only the multimodality out of scope.

The particle velocity and the sound pressure in a single measurement point enable the calculation of the sound field intensity and the energy in frequency bands. From these, the direction of arrival and the sound field diffuseness can be formulated. The fundamental assumption of DirAC is that the human auditory cues arise by the influence of these sound field properties along with the monaural spectral and temporal properties. Therefore a successful re-synthesis of these properties is assumed to bring a spatial hearing experience identical to that of the original measurement space.

A real-time linear phase filterbank version of DirAC was implemented. The reproduction quality of DirAC was shown to be excellent in formal listening tests if the number of loudspeakers is adequate and the microphone is ideal. The reproduction quality with standard 5.0 setup and Soundfield ST350 microphone was good. Additional experiments showed that the directional properties of the ST350 microphone collapse at frequencies above 1,5-3 kHz.

Keywords: Spatial sound reproduction, frequency band processing, B-format signal, diffuseness, spatial synthesis

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Juha Vilkamo
Työn nimi:	Tilaäänen toistaminen B-formaattiäänisignaaleista taajuuskaistaprosessoinnin avulla
Päivämäärä:	28.5.2008 **Sivuja:** 67
Tiedekunta:	Elektroniikka, Tietoliikenne ja Automaatio
Professuuri:	S-89
Työn valvoja:	Professori Matti Karjalainen
Työn ohjaaja:	Dosentti Ville Pulkki

Lisääntynyt tietämys tilakuulon toimintaperiaatteista on mahdollistanut lukuisien tilaäänentoistoteknologioiden synnyn. Näihin lukeutuvat muiden muassa monikanavaäänen pakkaus, kanavakokoonpanon muunnokset sekä tilaäänen yleinen kanavariippumaton esitystapa. Directional Audio Coding (DirAC) on teknologia, jolla pyritään analysoimaan ja vastaanottopäässä syntetisoimaan havainnon kannalta oleelliset äänikentän ominaisuudet.

Ihmisen tilakuulo toimii niinsanottujen vihjeiden avulla. Näitä ovat muiden muassa korviin saapuvien äänisignaalien keskinäiset erot sekä moniaistiset vihjeet kuten näköaistista saatava informaatio. DirAC:n tavoitteena on mitata äänitystilassa ja uudelleentuottaa kuuntelutilassa ne äänikentän ominaisuudet, jotka vaikuttavat kuuloaistiin liittyvien vihjeiden syntyyn.

Yhdestä pisteestä mitattavasta hiukkasnopeudesta sekä äänenpaineesta voidaan laskea äänikentän hetkellinen intensiteetti ja energia taajuuskaistoittain. Näistä voidaan puolestaan selvittää äänen tulosuunta sekä diffuusisuus eli hajaantuneisuus. DirAC:n perusoletus on, että ihmisen suuntakuulon vihjeet muodostuvat näiden ominaisuuksien perusteella, äänen taajuus- ja aikarakenteen lisäksi. Toisin sanoen oletus on, että mikäli nämä ominaisuudet onnistutaan uudelleentuottamaan, kuulijan tulisi kokea kuulokokemus, joka vastaisi täysin sitä kuulokokemusta, joka olisi syntynyt alkuperäisessä mittaustilassakin.

Reaaliaikainen lineaarivaiheiseen suodinpankkiin perustuva DirAC-ohjelmisto toteutettiin tutkimuksen yhteydessä. Kuuntelukokeet osoittivat, että riittävällä määrällä kaiuttimia sekä ideaalisella mikrofonilla DirAC:n kyky uudelleentuottaa tilaääntä oli erinomainen. 5.0 -kotiteatterikokoonpanoa sekä Soundfield ST350 -mikrofonia käytettäessä laatu oli hyvä. Lisätutkimukset osoittivat, että ST350 -mikrofonin toimivuus suunta-analyysissä heikkenee voimakkaasti taajuuksilla, jotka ylittävät 1,5-3 kHz.

Avainsanat: Tilaäänentoisto, taajuuskaistaprosessointi, B-formaattisignaali, diffuusisuus, suuntasynteesi

Acknowledgements

The research for this thesis was conducted in the Department of Signal Processing and Acoustics of Helsinki University of Technology. The project was funded by Emil Aaltonen foundation.

I want to thank my instructor Docent Ville Pulkki and my supervisor Professor Matti Karjalainen for guidance and the creation of an excellent research environment. I would also like to thank my co-workers Mr. Jukka Ahonen, Mr. Mikko-Ville Laitinen and Mr. Timo Hiekkanen for collaboration and the ever-present enjoyable mood. I want to express my gratitude also to all other personnel for the positively encouraging motivation throughout the laboratory.

Although last mentioned, but still of highest importance, I would like to thank my family and friends for everything.

Otaniemi, May 28, 2008

Juha Vilkamo

Contents

Abbreviations

ANOVA	Analysis of variance
DirAC	Directional Audio Coding
ERB	Equivalent rectangular bandwidth
IC	Inter-aural coherence
IIR	Infinite impulse response
ILD	Inter-aural level difference
ITD	Inter-aural time difference
ITU	International Telecommunication Union
MDCT	Modified discrete cosine transform
MUSHRA	Multiple stimulus and hidden reference and anchor
OLS	Overlap-save
STFT	Short time Fourier transform
VBAP	Vector base amplitude panning

Chapter 1

Introduction

Spatial hearing is a property virtually always active in our daily lives, and it functions without any conscious effort. Our neural systems have adapted through evolution to efficiently extract a wide variety of information from the vibration of air, in an astonishing precision.

If there would be no spatial processing in the hearing system, our auditory perception would merely be a confusing experience of two slightly different sounds, the ear inputs, without the ability to decompose the sounds to different sources in different locations. In reality, we usually do not even notice from the auditory perception that we have two ears, but the sounds are actually perceived to be located at specific locations in a three-dimensional environment. We do not only localize the sources, but we can sometimes even localize the surrounding surfaces or at least have a general feeling of the characteristics of the space that encompasses us.

Problems arise when the spatial perception is disturbed. For example, there are personal preferences such as not to listen to music with headphones or not to utilize teleconferencing systems due to the distant or otherwise unnatural experience. Real concerts are typically preferred over the ones reproduced with home sound setups. It is desirable in many spatial sound reproduction technologies to overcome the perceptual gap between the reproduced sound and the original "real" sound.

The field of study in this thesis is the recording and reproduction of the spatial sound of any real space with a B-format microphone (Section 5.1) and an arbitrary sound system at the receiver end. There have been numerous microphone techniques for spatial sound reproduction, with certain limitations each. Coincident microphone techniques [1] can be flexible in terms of reproduction in different sound setups but the high coherence between the channels is problematic in terms of the sound spaciousness and coloration due to coherent summation. Furthermore, since there is high coherence between the channels, the source localization tends to collapse to the nearest loudspeaker due to the precedence effect (Section 3.2.2). Spaced microphone techniques [2] avoid the coherence problem but they lack the scalability to different sound reproduction systems. Spaced techniques can achieve a pleasant listening experience, but they are not targeted for the exact replication of the spatial hearing experience and are also unpractical

for many recording situations.

Directional Audio Coding (DirAC) [3, 4] is a psychoacoustically motivated DSP-driven microphone technique which aims for a perfect re-synthesis of the perceptually relevant properties of any sound field. The goal is not only to have an ideal sweet spot performance, but also good quality within the whole listening area between the loudspeakers. DirAC utilizes existing microphones for analysis and is applicable to virtually all sound reproduction systems. Another application of DirAC is as a spatial audio coder, since the output of DirAC analysis can be reduced to a mono channel and a low bitrate stream of spatial parameters. In this thesis, DirAC is studied only as a microphone technique and the bitrate reduction issues are left out of scope.

Spatial audio coding (SAC) [5, 6, 7] shares many principles and problems with DirAC. The fundamental difference is that SAC is designed to analyze, downmix and rebuild existing multichannel recordings while DirAC aims for real spaces. The main field of SAC is bit rate reduction, but it can also be used for combining different channel configurations into one audio stream. SAC improves the overall perceptual quality by extracting the interchannel differences and downmixing the channels, thus increasing available bits per channel for the audio encoder. Parametric stereo is a special case of SAC and is implemented in modern audio codecs such as HE-AAC [8], which is currently widely used in low bitrate audio broadcasting over the Internet.

Spatial audio coding using universal cues [9] is a technology that falls in between SAC and DirAC. The multichannel recording is no longer represented in terms of loudspeaker signals, but in terms of "universal cues". The source audio is recorded multichannel audio such as a 5.1 channel stream, but there is no longer connection to any fixed loudspeaker setup after the encoding is performed. The concept of universal cues is similar to directional cues in DirAC, with the difference that DirAC cues are based on the physical properties of the sound field.

Yet another possible approach to achieve a universal playback of multichannel audio signals is to up-, down- or crossmix the given multichannel audio to fit to the desired sound system. A simple method is to use the available loudspeaker setup to position a set of virtual loudspeakers with panning techniques, although it is possible in this approach that the full potential of the loudspeaker setup is not utilized. A more extensive utilization is achieved with perceptually motivated spatial decomposition and resynthesis of the sound [10]. Channel conversion technologies can in some extent be applied as an extension to any coincident or spaced microphone technique to overcome the problem with the fixed speaker layout.

The following chapters discuss the relevant psychoacoustic principles, the sound reproduction methods, the DirAC method and its implementation, the conducted experiments and finally present the conclusions.

Chapter 2

Physics of sound

The famous riddle "If a tree falls in a forest and no one is around to hear it, does it make a sound?" along with many other interpretations, illustrates the many usages of the word "sound". In this chapter, sound is considered exclusively as a physical phenomenon and thus the falling tree indeed makes a sound. Chapter 3 discusses the subjective aspect of sound, from the viewpoint of perceptual psychology.

In terms of physics, the sound is fundamentally the movement and concentration of the air particles in the molecular level. In the macroscopic scale, these properties are manifested as the statistical quantities of pressure and particle velocity. Sound waves are changes in these quantities that propagate through space.

2.1 Point sources and plane waves

Physical sound is ignited by mechanical movement such as strings in an instrument, vocal chords, impacting objects or movement of touching surfaces. Different types of the ignition produce sounds that can be categorized by the type of vibration. These types are repetitive (tones), non-repetitive (noise) or burst-like (transients).

If a point source is active in a free field, the sound propagates to all directions and the energy density attenuates by $1/r^2$ where r is the distance. The attenuation happens as the same amount of energy is distributed over a larger area as the sphere grows as a function of the radius. The sound also attenuates due to air absorption as the energy especially in high frequencies is transformed to thermal energy, another form of molecular movement. Only few types of sources, such as explosions and specifically designed loudspeakers, can be considered to be omnidirectional point sources, while typical sources radiate sound energy direction-dependently.

A plane wave is a wave front where the propagating wave is not spherical but a plane. Plane waves can be emitted by flat vibrating surfaces. Spherical waves are typically approximated with a plane wave at a sufficient distance from the source. [11]

2.2 Reflections

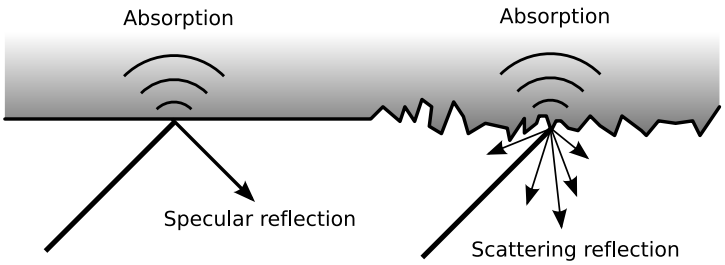

Figure 2.1: Specular and scattering reflections and absorption.

Frequency dependent phenomena occur as a sound wave meets a surface. If the surface size is large and the fine shape is small compared to the wavelength, the surface specularly reflects the sound similarly as a mirror reflects the light (Fig. 2.1). If the fine shape of the surface larger in comparison to the wavelength, a scattering reflection occurs. This means that the sound is reflected as light reflects from a sheet of paper, to all visible directions. If the surface itself is small in comparison to the wavelength, the sound wave passes the surface as if it would not exist. The surfaces also absorb the sound energy frequency dependently. [12]

2.3 Reverberation

The repeatedly reflecting and scattering sound builds up as reverberation (Fig. 2.2). The reverberating sound decays due to the absorption of the reflecting materials and the air. The amount of reverberation is typically discussed in terms of reverberation time, which is defined as the time in which the reverberating energy decays 60 dB. Long reverberation times are typically found in spaces with hard surfaces, since the reflections are less absorptive, and in large spaces, since the reflections occur less often. [12]

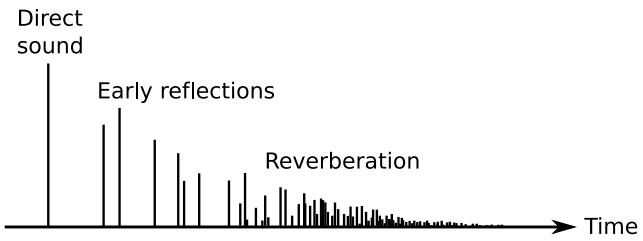

Figure 2.2: An illustration of an impulse response: direct sound, early reflections and reverberation.

2.4 Sound field

The sound field at each position of the space has a certain time-dependent vector quantity of particle velocity $\mathbf{u}$ and a scalar quantity of sound pressure p. From these, the sound field energy and intensity can be derived. The instantaneous energy density of a sound field is

$$E = \frac{1}{2} p_0 \left(\frac{p^2}{Z_0^2} + \|\mathbf{u}\|^2 \right) \tag{2.1}$$

and the instantaneous intensity vector is defined as

$$\mathbf{I} = p\mathbf{u} \tag{2.2}$$

where p_0 is the mean density of the air, $Z_0 = p_0 c$ is the acoustic impedance and c is the speed of sound. The intensity vector points towards the flow of the energy. The diffuseness of the sound field at the location is defined as

$$\psi = 1 - \frac{\| \langle \mathbf{I} \rangle \|}{c \langle E \rangle} \tag{2.3}$$

where $\langle \, \rangle$ denotes the time average. Diffuseness ranges within $0 \leq \psi \leq 1$ and stands for the diffuse fraction of total sound field energy. A completely diffuse sound field means that there is no net transport of acoustic energy, but instead the intensity averages to zero. Completely non-diffuse sound means that there is only one sound source at one location in a free field. In practice, both of these extremes are rare. [13]

A typical example of a highly diffuse sound is reverberation. A vast number of reflections reaches the measurement point from many directions in a limited time. Therefore $\| \langle \mathbf{I} \rangle \|$ becomes small and ψ becomes high. High diffuseness can also occur when there are two or more sources active at opposing directions.

In the scope of this thesis, the term "diffuse sound field" denotes a situation that corresponds to a large number of uncorrelated sound sources distributed in all directions with equal probability.

2.5 Modeling of acoustic spaces

Acoustic spaces can be modeled with numerous methods. Modeling is necessary to predict or simulate the behavior of the sound in a space when the space itself is not available, for example in the design of concert halls or in computer games.

There are numerous approaches in computerized acoustic modeling, ranging from high complexity but more detailed approaches (e.g. finite-difference time-domain method (FDTD) [14]) to computationally efficient but rougher approximations such as ray-based methods [15]. In concert hall design, a scale model is also often used for testing of the acoustic properties of the hall.

The image source method [16] (Fig. 2.3) is a ray-based method, which duplicates the sources in respect to the reflecting surfaces and applies the surface absorption caused by the reflection to the duplicated source. This operation can be repeated to the first order image sources to formulate higher order image sources. The result is a set of image sources in a free field, which is a practical situation for auralization purposes. The image source method gives the direct sound and the early part of the reflections (Fig 2.2). The late reverberation is typically approximated with a separate reverberation algorithm.

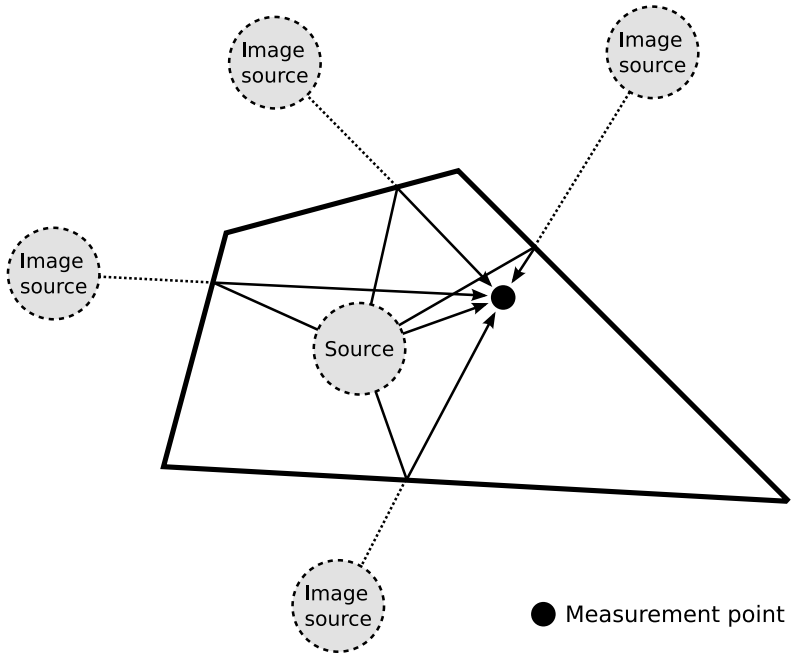

Figure 2.3: Illustration of first order image sources.

The process of creating audible content from the acoustic models is referred to as auralization. Auralization process utilizes the available audio hardware so that the reproduced sound is ideally perceived as if the listener would be in a defined location of the modeled space. To achieve best results, the reproduction should be arranged so that all direct sounds and reflections arrive to the listener as they arrive to the listening position in the model. The sound reproduction technologies that can be utilized also in the auralization of virtual spaces are described in Section 4.

Chapter 3

Psychoacoustics

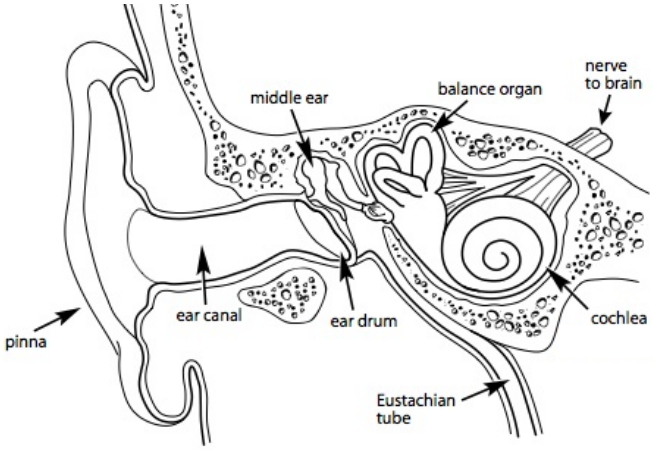

Figure 3.1: Human ear.

The function of ear (Fig. 3.1) is to transform air pressure changes into neural activity in a frequency-selective manner. The air vibration is mechanically transformed into movement of the eardrum, and is further transferred through the ossicles (three connected bones in middle ear) to fluid movement in the inner ear. The fluid movement in turn moves the basilar membrane on which the hair cells are located. The basilar membrane is frequency-selective so that each point in the membrane responds maximally to a certain frequency and less to other frequencies. The hair cells react to the movement by firing neural impulses to the auditory nerves. This is the last part of the ear, but only the starting point of our hearing system. [17]

The neural part of hearing contains a multitude of interconnected neural processes which in the higher levels include highly complex properties such as multimodality [17] and memory (e.g. in lingual perception). Fortunately for audio processing purposes, it is often not necessary to have full understanding of human cognition but to have experimental information about the overall functionality of all processes together. This knowledge is achieved through carefully

planned subjective listening tests and statistical analysis. Psychoacoustics is a term that means the study of human auditory perception. In this chapter, the psychoacoustic principles related to DirAC are explained.

3.1 Critical bands

The phenomenon of critical bands is that within certain frequency bands the perceived loudness depends solely on the signal energy. This property is very convenient in the point of view of computerized audio signal processing since the signal energy is easily calculable. The critical bands are considered to be a valid scale also for spatial hearing [18].

There are two common ways to define the critical bands, the equivalent rectangular bandwidth (ERB) scale [19] which is utilized also in DirAC

$$\triangle f_{\text{ERB}} = 24.7 + 0.108 f_c \tag{3.1}$$

and the Bark scale [20]

$$\triangle f_{\text{Bark}} = 25 + 75 \left[1 + 1.4 \left(\frac{f_c}{1000} \right)^2 \right]^{0.69} \tag{3.2}$$

where $\triangle f$ is the bandwidth and f_c is the center frequency of the frequency band. Both scales are derived from experimental results. The Bark scale is derived from narrowband noise listening tests by having a constant noise energy and adjusting the bandwidth, and finding the bandwidth in which the perceived loudness starts to increase. The ERB scale is derived by notched noise masking listening tests by altering the passband bandwidth and studying the detectability of a sinusoid. The masking noise approach guarantees that the listening is restricted only to the frequency band in question. The ERBs are narrower, and the whole hearing range (20Hz - 20kHz) is divided into 42 bands, while the Bark scale is 24 bands. Also a third-octave band is sometimes used in approximation of the critical bands. These scales are plotted in Fig. 3.2.

3.2 Directional hearing

When a sound source is not in the median plane of the listener (Fig. 3.3), there is a distance difference from the source to the listener's ears, and therefore the input signal of the farther ear is delayed in comparison to the nearer ear. This delay is called interaural time difference (ITD), modeled first in [21]. When a wavefront reaches the listener, the reflections from upper torso and pinnae will direction dependently affect the spectrum of the sound entering the ear canals. This and the head shadowing causes interaural level difference (ILD). ILD is defined as the level difference between right and left ear input in decibels. These binaural cues play an important role in directional hearing along with monaural cues and information from visual and tactile senses. Our hearing system adapts through experience to utilize these spatial cues in

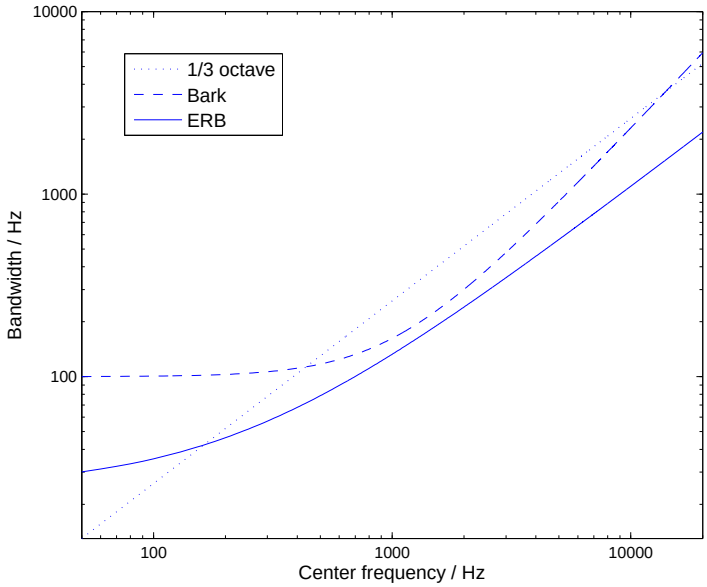

Figure 3.2: Comparison of frequency resolution scales

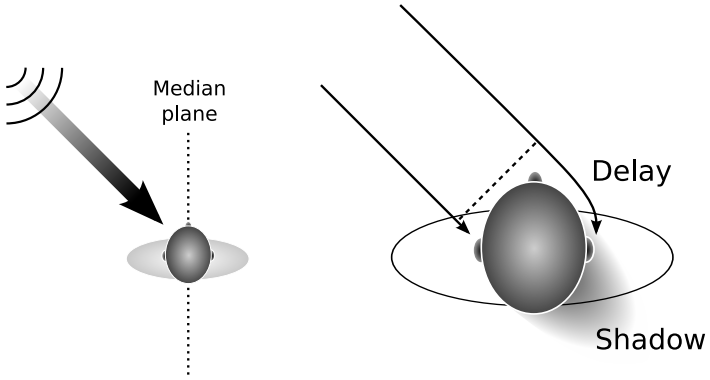

Figure 3.3: Sound arriving to the listener's ears. The delay and head shadowing contribute to the interaural time difference (ITD) and interaural level difference (ILD).

determining the sound source location. The change in the binaural cues in respect to the head movement provides further spatial information.

ILD and ITD are processed in frequency bands. For perception of azimuth, ITD cues are dominant in frequencies lower than approximately 1.5 kHz while ILD cues are dominant above this limit [22]. Additionally, the interaural coherence (IC) is an often used measure to represent the "similarity" of the sound entering the ears. IC is defined as the normalized correlation between ears when the ITD is compensated. The parametric stereo coding algorithms [5, 6] analyze the

inter-channel equivalents of ITD, ILD and IC. In headphone listening, the inter-channel cues are equivalent to inter-aural cues.

3.2.1 Distance cues

There are various cues that affect the distance perception [23]. If the listener can estimate the spectral properties of a source, for example with a human voice, the sound level and the coloration by air absorption are cues for the distance perception. Other cues are the ratio and delay between the direct and reflected sound and change in the spectral properties as the source moves towards or away from the listener.

3.2.2 Precedence effect

In normal listening situations, a considerable part of the sound energy reaching a listener is from non-direct paths from the source. Precedence effect [24, 23] is the phenomenon which emphasizes the first arriving wavefront in source localization. When a same sound is briefly (1-40ms) repeated from arbitrary directions, only one sound source is perceived, in the direction of the first arriving wavefront. The phenomenon is present still if the level of the first arriving sound is as much as 10 dB lower than the latter one. If the arrival interval of the sounds is reduced to less than approximately one millisecond, the directional perception starts to shift towards the second arrival. If there are quick changes in the acoustic environment, the dominance of the first arriving sound can break down temporarily and then again adapt to the new environment. In the light of this knowledge, the precedence effect can be explained as our adaptive way of coping with the everyday sound environments where reflective surfaces are present.

3.2.3 Multimodality

In addition to the auditory information, synchronous visual and tactile sensory input can affect the spatial hearing experience [17]. For example the voice of a person speaking on a television screen is heard from the direction of the image, even if the actual source might be a single loudspeaker at the side of the screen. A simplified generalization of the multimodal perception is that the most likely scenario that matches the sensory inputs becomes the perception.

Multimodality poses a problem for spatial hearing tests. For example, if a perfect auralization of a certain space could be created with headphones, but the listener is in an acoustically very different room, there is an inherent inconsistency in multimodal information. Therefore, even with a perfectly working system the listener may intuitively feel artificialness about the situation.

3.2.4 Timbre

The timbre roughly corresponds to the sound color. American National Standard Association [25] defines timbre as “that attribute of auditory sensation in terms of which a listener can judge two sounds similarly presented and having the same loudness and pitch as dissimilar”. The

definition continues with a further explanation: “Timbre depends primarily on the spectrum of the stimulus, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus”.

Chapter 4

Sound reproduction

In the scope of this thesis a virtual source is defined as a localizable auditory object that may or may not coincide with real sources such as loudspeakers. In many spatial audio technologies including DirAC it is necessary to be able to position virtual sound sources to any direction. In this chapter, the most relevant technologies that fulfill this requirement are discussed.

4.1 Multichannel loudspeaker systems

A virtual source can be positioned in the space between the loudspeakers by panning methods such as Vector Base Amplitude Panning (VBAP) [26]. The precision of virtual sources decreases as the angle between the loudspeakers increases (see the discussion in Section 5.6).

4.1.1 Vector Base Amplitude Panning (VBAP)

VBAP (Fig. 4.1) is a robust technology for positioning virtual sources in the line between two loudspeakers, or in the area between three loudspeakers. VBAP assumes that the loudspeakers are located on a sphere, and the listening point is the center point of the sphere. If there are differences in the distance, they should be compensated with appropriate delays and gains. In VBAP, the imaginary sphere around the listening point is divided into non-overlapping triangular sections with the corners at the loudspeaker locations. When a virtual source is positioned to an arbitrary point, the three loudspeakers that form the corresponding triangle will be active. The gains of these three loudspeakers fulfill the following equation.

$$\mathbf{p} = \mathbf{L}\mathbf{g} = \begin{bmatrix} \mathbf{l}_1 & \mathbf{l}_2 & \mathbf{l}_3 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix} \tag{4.1}$$

where g_n is the gain value of the nth loudspeaker, $\mathbf{p} = \begin{bmatrix} p_x & p_y & p_z \end{bmatrix}^T$ is a unit vector pointing towards the virtual source and $\mathbf{l}_n = \begin{bmatrix} l_{n,x} & l_{n,y} & l_{n,z} \end{bmatrix}^T$ is a unit vector pointing towards the

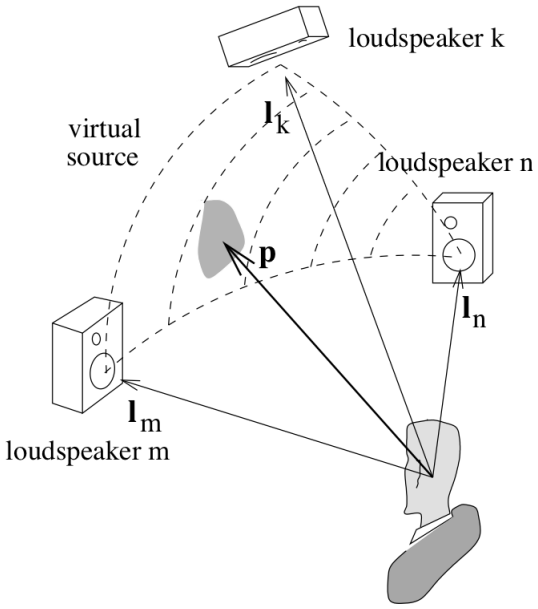

Figure 4.1: Vector Base Amplitude Panning (VBAP)

nth loudspeaker taking part in the amplitude panning. The gain factors for direction indicated by $\mathbf{p}$ can then be calculated by

$$\mathbf{g} = \mathbf{L}^{-1}\mathbf{p} \tag{4.2}$$

These gains are finally normalized to achieve a constant total energy independent of direction

$$\sum_i g_i^2 = 1 \tag{4.3}$$

4.2 Headphones

The three-dimensional positioning of virtual sources can also be performed with headphones by utilizing head-related transfer functions (HRTFs) [27], which are the transfer functions from a point in space to the listener's ears. If a source audio signal is filtered with a pair of HRTFs for a specific location and then listened with headphones, the source is ideally perceived as if it would be in that location. Utilization of HRTFs often suffers from problems that arise from the individuality since the HRTFs depend on the physiological properties of the listener. Artificial heads have been designed to represent an average listener, but experiments have shown that a selected HRTF recording from real humans' ears produces better results [28]. Head tracking enables to keep the virtual sources stable so that they do not rotate along with head rotation. Utilization of

head tracking is also found to reduce possible front/back confusions [29]. Some studies claim that head tracking does not significantly increase the localization precision in comparison to static listening situations [29, 30], although other studies [31] claim otherwise.

4.3 Crosstalk cancelled stereo

When two loudspeakers are used, but headphone-like virtual surround sound is needed, a system based on crosstalk-cancelling can be designed [32]. In this scenario, the signals from both loudspeakers reach both of the listener's ears. The two channels are mixed so that ideally this crosstalk is cancelled as in Fig. 4.2. The signal of left channel (and vice versa) is fed in opposite phase and delayed to the right channel so that when the wavefronts arrive to the right ear, they sum to zero. This situation corresponds to using headphones and enables the techniques described in Section 4.2. In practice, the cancellation of high frequencies is impossible with this method, but the head shadowing compensates this deficiency in some extent. The crosstalk-cancelling signal in turn recursively causes crosstalk which also has to be cancelled.

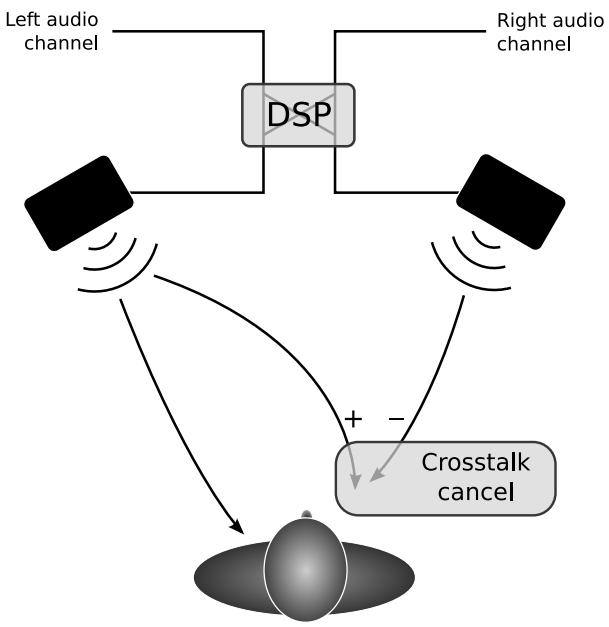

Figure 4.2: Crosstalk cancelled stereo. The crosstalk-cancelling signal in turn recursively causes crosstalk which also has to be cancelled.

The crosstalk cancellation can be problematic since the sound does not propagate only along the direct path, but also reflects from walls and objects. These reflections can be controlled only by using enough acoustic damping in the listening room. Crosstalk canceling is also very sensitive to the listener position and alignment, which makes it impractical for normal listening

situations.

4.4 Wave field synthesis

For a very large array of loudspeakers, a virtual sound source can be positioned more freely by using wave field synthesis (WFS) [33] (Fig. 4.3). In WFS, the physical properties of a wave field are reconstructed, and the virtual source position is no longer limited to the surface between loudspeakers as in VBAP. Instead, the virtual source can also be positioned freely both in front and behind the loudspeakers. The limitation is that the virtual sources can be positioned only so that a ray from the listening point through the virtual source also crosses the loudspeaker array at some point. In DirAC analysis, the fundamental assumption is that the arriving wave fronts are plane waves from infinite distance, and therefore it could be beneficial to be able to synthesize true plane waves with WFS. This approach would extend the sweet spot to the whole effective area of the WFS. The technology however is not practical due to the extensive hardware requirements and so far there has been no research concerning the applicability of wave field synthesis in DirAC.

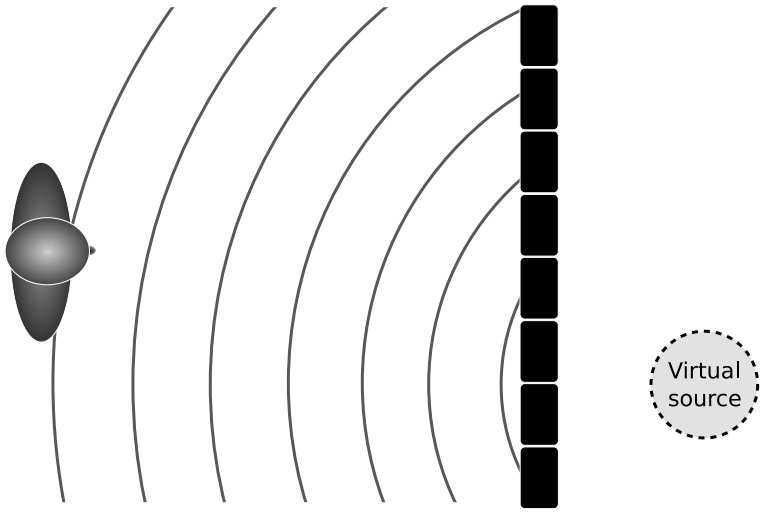

Figure 4.3: Wave field synthesis can produce virtual sources both in front and behind the loudspeaker array.

4.5 Ambisonics

Ambisonics [34] is a microphone technique that shares the starting point and the goal with DirAC. Both techniques aim for the reproduction of two- or three-dimensional sound environments from coincident microphone signals, although the approaches are very different. Am-

bisonics essentially creates virtual directional microphone signals, depending on the microphone type and the loudspeaker setup. In the basic form, the microphone directional patterns of the first and second order Ambisonics utilize the following microphone patterns

$$s_n(t) = a + b\cos(\theta_n) + c\cos(2\theta_n) \tag{4.4}$$

where a, b and c are constants which define the microphone pattern and (θ_n) is the spatial angle from the angle of the loudspeaker. c is zero in first order Ambisonics. A set of directional patterns are illustrated in Fig 4.4, with $a = \frac{1}{3}$, $b = \frac{2}{3}$ and $c = 0$ for first order and $a = \frac{1}{5}$, $b = \frac{2}{5}$ and $c = \frac{2}{5}$ for second order.

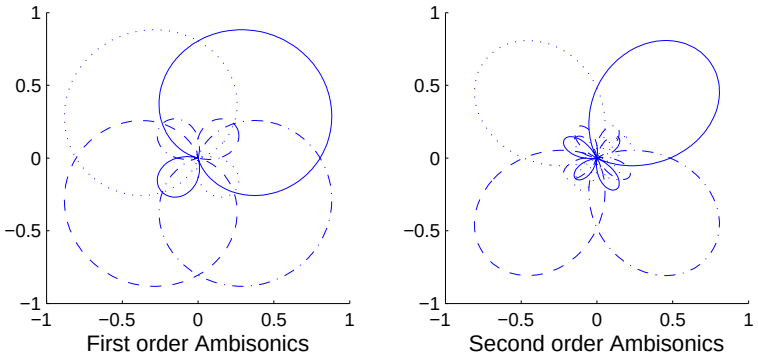

Figure 4.4: Directional microphone patterns of first- and second-order Ambisonics for a loudspeaker setup of four equally distributed loudspeakers.

The problem of Ambisonics is that the high coherence between the channels causes undesirable effects, especially if the density of the loudspeakers is high. The possible problems include comb filtering effects and emphasis of low frequencies. Furthermore, the spatial perception tends to collapse to the nearest loudspeaker in off sweet spot listening due to the precedence effect. The timbral problems are lesser if the usage of the first order Ambisonics is restricted to sparse loudspeaker systems such as the quadraphonic system in Fig. 4.4, or by using higher order microphones. The latter approach is however problematic in terms of quality and availability of existing higher order microphones.

Chapter 5

Directional Audio Coding (DirAC)

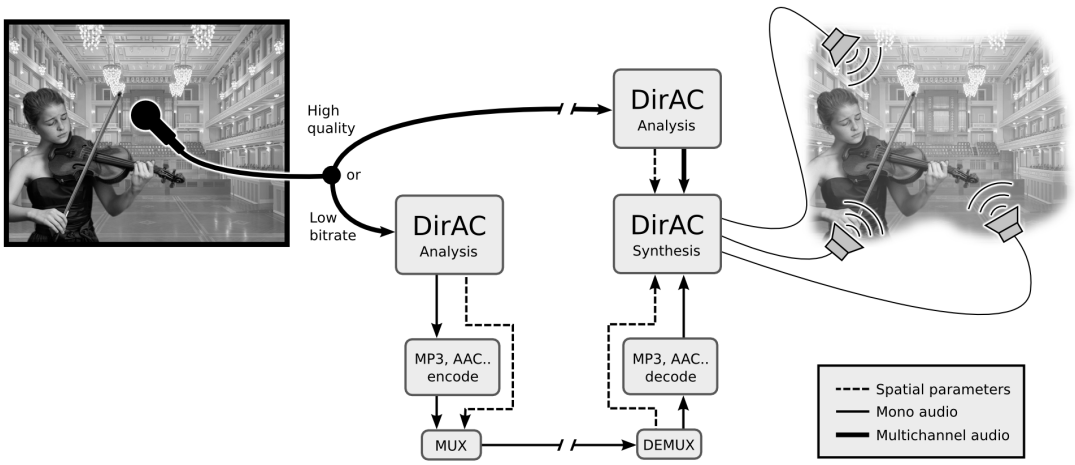

Figure 5.1: High quality (top) and low bit rate (bottom) spatial audio reproduction with DirAC.

DirAC is an active microphone technique which is designed to reproduce an arbitrary sound environment so that all psychoacoustic cues presented in Chapter 3 are preserved in the extent that the reproduced sound causes the same auditory perception as the original sound (Fig. 5.1). The following assumptions are the fundamental principles of DirAC.

- Direction of arrival transforms to the ILD and ITD cues.
- Diffuseness transforms into IC cues.
- One direction of arrival and one diffuseness value in each critical band and in each time instant is a sufficient approximation in all situations.
- Timbre depends on the monaural spectral properties and the spatial distribution of the sound.

- All previously mentioned cues are related to the lower levels of the auditory system, and it is assumed that the higher levels create the spatial impression according to these cues. Therefore the total spatial listening experience is assumed to be preserved when the cues are preserved. The cues are then assumed to be preserved by the proper reproduction of those sound field properties from which the cues arise.

The following chapters often discuss the observations from an informal listening situation. Unless otherwise mentioned, this means a situation where the author with or without other members of the DirAC research team compares a 21-channel three-dimensional virtual reality and a 16-channel three-dimensional DirAC reproduction in an anechoic chamber in sweet spot listening. The setups and the used reference stimuli were the same as in the formal listening tests and are described in Section 7.3.

5.1 Approximation of sound field intensity and energy with B-format microphone

DirAC analysis is performed at a single measurement point based on the sound field intensity in Eq. (2.2) and energy in Eq. (2.1), which are formulated from the particle velocity and the sound pressure. Both of these values can be derived from a B-format microphone signal. B-format microphone has four channels: omnidirectional and three figure-of-eight microphones organized orthogonally as in Fig. 5.2.

The omnidirectional microphone signal is denoted as $w(t)$. The three figure-of-eight microphones $x(t)$, $y(t)$ and $z(t)$ are scaled with $\sqrt{2}$. B-format microphones can be constructed by placing three figure-of-eight microphones and an omnidirectional microphone in the same location, but also by using four subcardioid microphones [35]. A low-cost B-format microphone implementation is possible with an array of omnidirectional microphone capsules [36].

Assuming a planar sound wave [13], the particle velocity can be estimated from the B-format signal by

$$\hat{\mathbf{u}}(t) = -\frac{1}{Z_0\sqrt{2}}\left(x(t)\mathbf{e}_x + y(t)\mathbf{e}_y + z(t)\mathbf{e}_z\right) \tag{5.1}$$

where the term $-\frac{1}{\sqrt{2}}$ term is due to the alignment and scaling of the X-, Y- and Z-channels of the standard B-format microphone. The sound pressure is estimated simply

$$\hat{p}(t) = w(t) \tag{5.2}$$

The estimates for energy in Eq. (2.1) and intensity in Eq. (2.2) are then

$$\hat{E}(t) = \frac{p_0}{Z_0^2}\left(\frac{w^2(t)}{2} + \frac{x^2(t) + y^2(t) + z^2(t)}{4}\right) \tag{5.3}$$

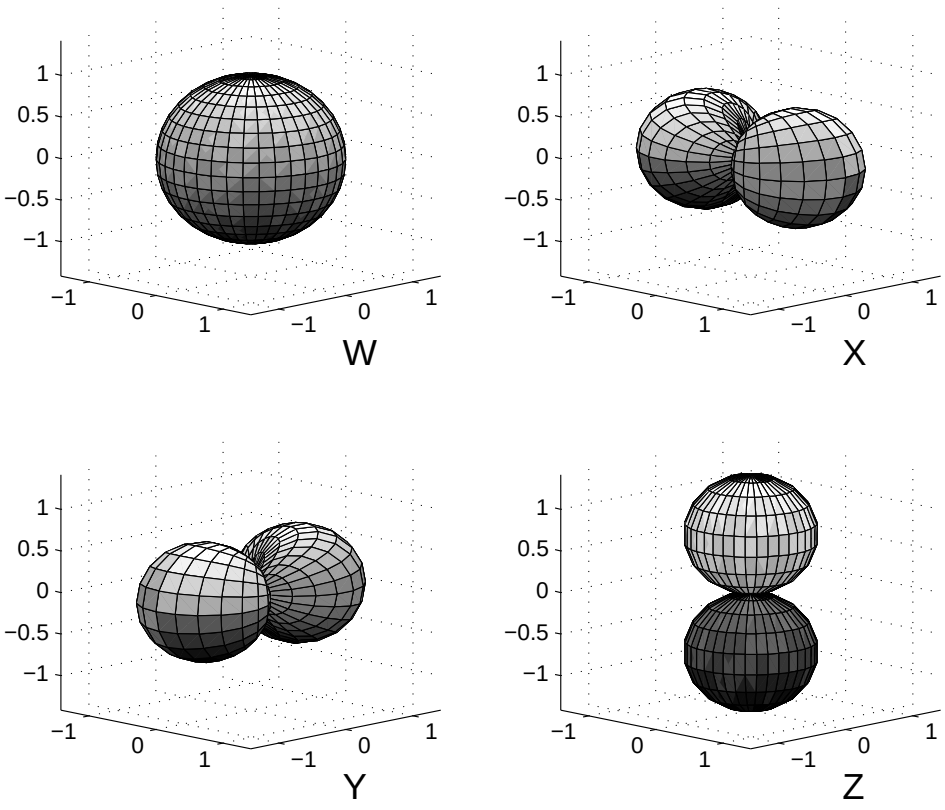

Figure 5.2: The directional patterns of the four microphones in a B-format microphone.

$$\hat{\mathbf{I}}(t) = -\frac{1}{Z_0\sqrt{2}} w(t)\,(x(t)\mathbf{e}_x + y(t)\mathbf{e}_y + z(t)\mathbf{e}_z) \tag{5.4}$$

5.2 Time-frequency analysis

DirAC analysis and synthesis is performed in frequency bands. Figure 5.3 illustrates a block diagram of DirAC which operates on a linear phase filterbank, although DirAC is not restricted only to this type of frequency analysis.

For any time-frequency transform there is a minimum possible area that can be covered in the time-frequency plane [37]. In other words, there is an inherent compromise between time resolution and frequency resolution. This compromise is actualized also in human hearing. In low frequencies, the frequency resolution is high but the time resolution is low, and in high frequencies, the time resolution is high but the frequency resolution is low.

In many audio processing technologies, the frequency analysis is done by utilizing computationally efficient block-wise frequency transforms such as short-time Fourier transform (STFT), utilized in [3, 5, 6, 9] and modified discrete cosine transform (MDCT), utilized especially in the area of audio coding [7, 8]. The computational efficiency of these methods is not limited to the

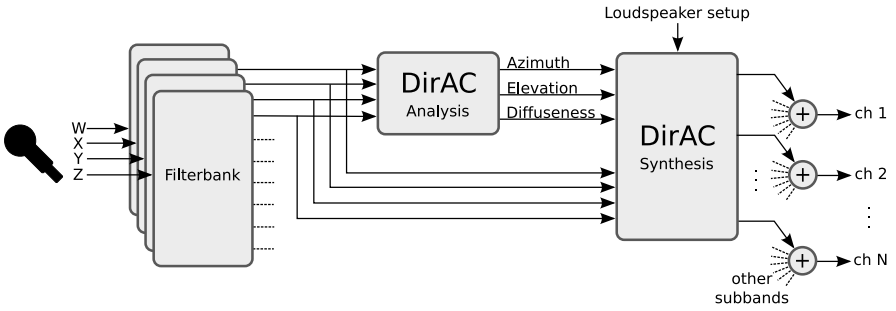

Figure 5.3: DirAC analysis and synthesis is performed in frequency bands.

transform itself, but through them the calculations of convolution and correlation also become very efficient. The properties of these transforms, such as the circularity of the STFT must be fully taken into account to achieve desired functionality.

A computationally expensive but less restrictive approach is to use a filterbank with a property of higher time resolution in the higher frequency bands and higher frequency resolution in the lower bands. The filterbank approach is time invariant and avoids the possible temporal artefacts of block processing methods. A detailed study of the time-frequency properties of the described methods is in Section 7.1.

5.2.1 Analysis of direction of arrival and diffuseness

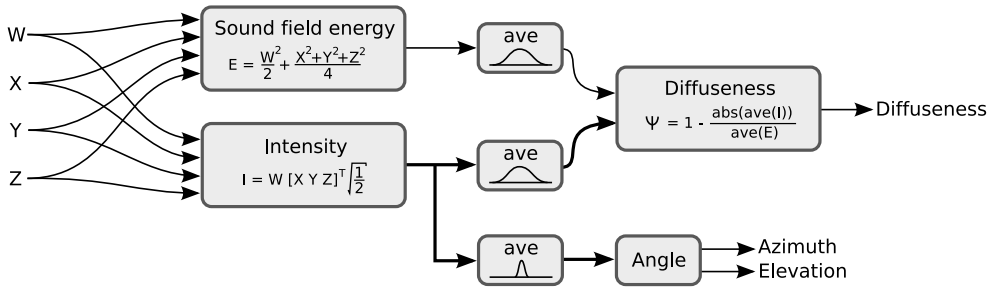

Figure 5.4: DirAC analysis with a B-format signal.

The direction of arrival is defined as the opposite direction of the intensity vector in Eq. (2.2). The multiplier $\frac{1}{Z_0\sqrt{2}}$ in the equation can be discarded in this analysis as the constant gain does not affect the analyzed direction. The estimates of azimuth and elevation are

$$\hat{\theta}(t) = \begin{cases} \arctan\left(\frac{\langle y(t)w(t)\rangle}{\langle x(t)w(t)\rangle}\right) & , \langle x(t)w(t)\rangle \geq 0 \\ \arctan\left(\frac{\langle y(t)w(t)\rangle}{\langle x(t)w(t)\rangle}\right) - 180^o & , \langle x(t)w(t)\rangle < 0. \end{cases} \tag{5.5}$$

$$\hat{\varphi}(t) = \arctan\left(\frac{\langle z(t)w(t)\rangle}{\sqrt{\langle x(t)w(t)\rangle^2 + \langle y(t)w(t)\rangle^2}}\right) \tag{5.6}$$

When Eq. (5.3) and Eq. (5.4) are substituted to Eq. (2.3), we get the estimate of the diffuseness from B-format signal

$$\hat{\psi}(t) = 1 - \frac{\| \left\langle \hat{\mathbf{I}}(\mathbf{t}) \right\rangle \|}{c\left\langle \hat{E}(t) \right\rangle} = 1 - \frac{\sqrt{\langle x(t)w(t)\rangle^2 + \langle y(t)w(t)\rangle^2 + \langle z(t)w(t)\rangle^2}}{\sqrt{2}\left\langle \frac{w^2(t)}{2} + \frac{x^2(t)+y^2(t)+z^2(t)}{4} \right\rangle} \tag{5.7}$$

Note that the analysis for azimuth, elevation and diffuseness depends solely on microphone signals $w(t)$, $x(t)$, $y(t)$ and $z(t)$. The block diagram of this analysis is presented in Fig. 5.4. For simplicity, the time-dependency is omitted in the equations that follow.

5.3 DirAC synthesis

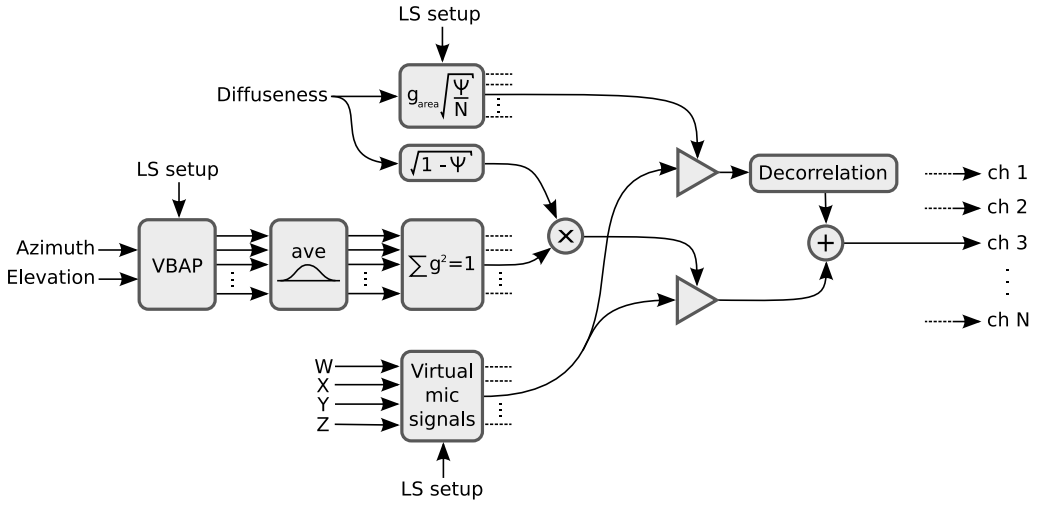

Figure 5.5: DirAC synthesis in one frequency band (excluding gain compensations, Sections 5.4.1 and 5.4.2).

DirAC synthesis (Fig. 5.5) can be considered to be an independent process from the analysis. As the goal of the analysis part is to extract the perceptually relevant sound field properties (direction and diffuseness), the goal of the synthesis part is then to utilize the available loudspeaker setup to create a sound field that has these properties. The focus in this thesis is in the multichannel loudspeaker playback even though DirAC is also scalable to the other sound setups mentioned in Chapter 4.

5.3.1 Loudspeaker gains for non-diffuse sound

The non-diffuse sound is synthesized by positioning the non-diffuse fraction of the total sound energy with VBAP to the direction determined in the analysis. The total gain of the non-diffuse sound for the nth loudspeaker is

$$g_{\text{nondiff}}(n) = g_{\text{vbap}}(n, \theta, \varphi)\sqrt{1-\psi} \tag{5.8}$$

where $g_{\text{vbap}}(n, \theta, \varphi)$ is the gain determined by VBAP for azimuth θ and elevation φ for the n:th loudspeaker.

5.3.2 Loudspeaker gains for diffuse sound

The diffuse sound is synthesized by distributing and decorrelating (Chapter 5.3.3) the diffuse fraction of the total sound energy to all loudspeakers. The gain of the diffuse sound for nth loudspeaker is

$$g_{\text{diffuse}}(n) = g_{\text{area}}(n)\sqrt{\frac{\psi}{N}} \tag{5.9}$$

where N is the number of loudspeakers and $g_{\text{area}}(n)$ is the gain multiplier for n:th loudspeaker defined as

$$g_{\text{area}}(n) = \sqrt{\frac{A(n)}{E\left[A(n)\right]}} \tag{5.10}$$

The area $A(n)$ is defined as the area of the surface composed of the points on a listening point centered sphere for which the Euclidean distance from n:th loudspeaker is smaller than for any other loudspeaker. $E\left[A(n)\right]$ denotes the expectation of $A(n)$, i.e. the average area. This gain compensation balances the spatial distribution of the produced diffuse field. If for example 5.0 sound setup would be used with constant $g_{\text{area}}(n) = 1$, the diffuse sound might be perceived to be mainly coming from the direction of the frontal loudspeakers. The $g_{\text{area}}(n)$ is higher at the directions where the loudspeakers are sparse.

5.3.3 Decorrelation

Decorrelation of the diffuse part of the signal is a complex issue which has been discussed within many studies concerning spatial audio synthesis [6, 38, 39, 40]. An ideal decorrelator can be outlined as follows:

1. A decorrelated signal listened separately should be indistinguishable from the original audio. In other words, the timbre should not change.

2. The decorrelated signal should be incoherent with the original sound. Decorrelated sound added to original should be perceived as the original sound, but 3 dB louder.

3. Multiple decorrelated versions of a signal should be incoherent with each other.

In practice, these idealities are not met in existing decorrelation methods. The decorrelators are instead heuristically tuned to maximize the decorrelation property and minimize the audible artefacts. Typically there is a compromise between these two goals.

For filterbank-based applications, a very convenient approach is to use channel- and frequency band-dependent time-invariant delays [40]. The computational load is very low since only a ring memory buffer is required without any arithmetic operations. When the decorrelation is performed in this way, the spectral and in some extent temporal resemblance in respect to the original sound is preserved. The following design boundaries for the delays were selected, following the principles of the precedence effect:

1. The minimum delay should be at least 1 millisecond. Then the precedence of non-diffuse sound is not affected by the delayed diffuse sounds. An additional boundary of 1-3 milliseconds can be added to guarantee the precedence in a larger listening area.

2. The maximum delay should be within the range in which the delayed sound is fused to the possible non-diffuse sound so that it is not heard as a separate echo. This limit is 5 milliseconds for clicks and 40 milliseconds for complex signals [23]. It can be reasoned from this knowledge that the maximum delay in decorrelation process should be less than 5 milliseconds for high frequencies and less than 40 milliseconds for low frequencies.

3. The delays are adjusted so that the consecutive frequency bands are in phase at the boundary frequency. This is to avoid the loss of energy in the transition bandwidths.

5.4 Directional microphones in DirAC synthesis

As explained in the previous chapters, DirAC synthesis is performed by creating the direct and diffuse parts of the sound by positioning the non-diffuse fraction of the total energy to the desired direction and distributing and decorrelating the diffuse fraction to all loudspeakers. The most straightforward choice for the source audio is the omnidirectional microphone signal.

In practice, there are several disadvantages when only omnidirectional signal is used. One problem is that since the source data for all loudspeakers is identical, i.e. fully correlated, the decorrelation process has to be stronger. This leads to the requirement of longer frequency-dependent delays or more prominent artificial reverberators, which both lead to audible artefacts. If the source data would be less correlated to begin with, more subtle decorrelators would be sufficient to produce a sound that would be perceived equivalently to a diffuse sound field.

Other problems with usage of only omnidirectional signal are also present. Let us consider a situation with a continuous sound source and a mirror reflection from the wall as in Fig. 5.6. In this case, a listener will without fail hear the sound coming from the direction of the source due to the precedence effect. In DirAC analysis however, the analyzed direction of arrival vector will

fluctuate between the source direction and the reflection direction. The problem arises since it is necessary, as it will be explained in Section 5.5, to average many parameters such as loudspeaker gains over time to avoid audible artefacts. This sluggishness can cause spatial smearing of the reproduced virtual source towards the reflection or towards another virtual source.

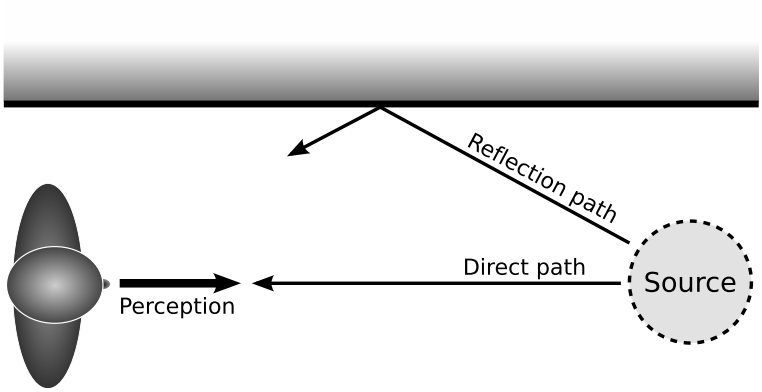

Figure 5.6: An example situation with a direct sound and one reflection.

The situation of Fig. 5.6 also illustrates another problem which arises due to the usage of omnidirectional microphone in synthesis. The sound from the reflection path is superimposed to the sound from the direct path, and therefore the microphone signal is comb filtered. Comb filtering occurs also in both of the ears, but since they are located separately the comb filter is different for each ear, and therefore there is more information available for the detection of timbre. Informal listening tests showed that if DirAC synthesis is performed with omnidirectional microphone signal, the comb filtering is in some cases prominent.

The above limitations of the omnidirectional synthesis limits its usage to medium quality DirAC applications. A way to address all of the problems described above is by using directional microphones pointing towards the angles of the loudspeakers. This leads to a situation where each loudspeaker signal is relatively highly correlated with the adjacent loudspeakers, but less correlated with loudspeakers far away. The relatively high correlation between adjacent loudspeakers enables the virtual source positioning approximately as in amplitude panning techniques. The lower overall inter-channel coherence enables lighter decorrelation and therefore reduces the possibility of artefacts. Also, the precedence effect and therefore the source stability improves as the reflections are separated from the source and additionally the unwanted comb filtering effect is significantly reduced.

5.4.1 Arbitrarily shaped microphone signals: Accurate gain compensation

In addition to the benefits of the utilization of directional microphones, the off-center attenuation (Fig. 5.7) also causes undesirable effects:

1. The energy of the synthesized non-diffuse sound is affected, except when the diffuseness

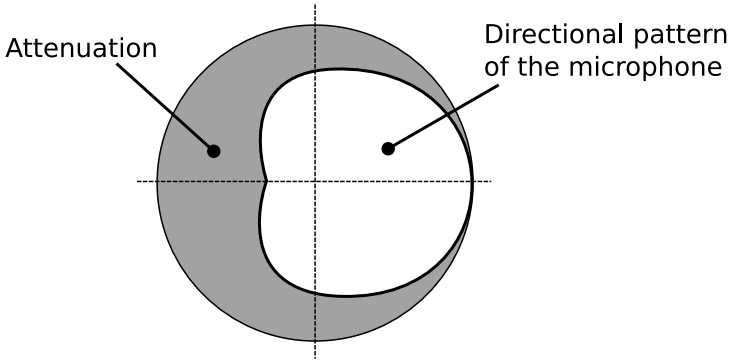

Figure 5.7: The directional pattern of a microphone. The gray area illustrates the attenuation in off-center directions. This energy loss must be taken into account in DirAC synthesis.

is zero and the source is exactly at the angle corresponding to a loudspeaker in the receiver end.

2. The energy of the synthesized diffuse sound is affected.

3. The direct sound in the measurement space affects the spatial distribution of the synthesized diffuse sound.

Simply stated, once the directional microphones are introduced, the DirAC synthesis no longer produces the sound field properties as intended. To counter this, a theoretically accurate, although unstable energy compensation scheme is derived. The next section explains a simplified and practical scheme.

In the following formulation, it is assumed for simplicity that the directional patterns of the microphones are rotationally symmetric and point towards the corresponding loudspeaker. This allows a notation that the microphone directional pattern is dependent only on the spatial angle γ_n from the direction of the nth loudspeaker. Let us define $g_{\text{mic}}(\gamma_n)$ as the gain of the microphone pattern at spatial angle γ_n from nth loudspeaker and $g_{\oint}$ as the microphone's gain for the enveloping diffuse field. As in DirAC generally, an approximation is made that in each time instant and in each frequency band, the sound field is a perfectly diffuse sound field superimposed with one source in a free field. E_{tot} is defined as the total energy of the sound field, composed of the non-diffuse part $(1-\psi)E_{\text{tot}}$ and the diffuse part ψE_{tot}. For simplicity, in the following equations the loudspeaker gains resulting from all processes of DirAC for non-diffuse sound and diffuse sound are denoted as $g_{\text{ND,n}}$ and $g_{\text{D,n}}$, respectively. From the above assumptions, the

sound energy that is produced with nth loudspeaker can be formulated with

$$E_n = (g_{\mathrm{ND},n}^2 + g_{\mathrm{D},n}^2) \underbrace{\left[g_{\mathrm{mic}}^2(\gamma_n)(1-\psi) + g_{\oint}^2 \psi \right]}_{\text{attenuation}} E_{\mathrm{tot}} \tag{5.11}$$

To compensate the attenuation of the directional microphone, a compensating gain factor g_n is added so that

$$g_n^2(\psi, \gamma_n) E_{\mathrm{n}} = (g_{\mathrm{ND},n}^2 + g_{\mathrm{D},n}^2) E_{\mathrm{tot}} \tag{5.12}$$

From Eq. (5.11) and Eq. (5.12)

$$g_n(\psi, \gamma_n) = \frac{1}{\sqrt{g_{\mathrm{mic}}^2(\gamma_n)(1-\psi) + g_{\oint}^2 \psi}} \tag{5.13}$$

This gain compensation depends on the diffuseness and the spatial angle between the direction of arrival and the loudspeaker. The latter dependence is problematic since the direction vector usually fluctuates fast. In addition, g_n can have infinite values. To avoid artefacts, the maximum of g_n must be limited and direction of arrival vector must be slowed down. Even with these measures, it is difficult to avoid artefacts, especially when real microphones are used instead of simulated ideal microphones. In the next section, another gain compensation approach is presented, which is less accurate but robust for artefacts.

5.4.2 Arbitrarily shaped microphone signals: Robust gain compensation

Another approach to compensate the negative effects of directional microphones is to not to consider the energy losses of a specific loudspeaker but the total energy losses for the non-diffuse and the diffuse sound. As an addition to the approximations in the previous section, a further approximation is made that the source direction corresponds exactly to the direction of a loudspeaker in the receiver end. By these means it is possible to formulate a smoothly behaving overall gain compensating scheme that does not depend on the spatial angle γ_n. For each frequency band, the energy of the synthesized as the non-diffuse sound is

$$E_{\mathrm{nondiff}} = (1-\psi) \underbrace{\left[(1-\psi) g_{\mathrm{mic}}^2(0) + \psi g_{\oint}^2 \right]}_{\text{attenuation}} E_{\mathrm{tot}} \tag{5.14}$$

Since E_{nondiff} should be equal to $(1-\psi)E_{\mathrm{tot}}$, the gain compensation is

$$g_{\mathrm{nondiff\_c}}(\psi) = \frac{1}{\sqrt{(1-\psi) g_{\mathrm{mic}}^2(0) + \psi g_{\oint}^2}} \tag{5.15}$$

Similarly, the expectation of the energy of the diffuse sound for one channel is

$$E\left[E_{\text{diff}}\right] = \frac{\psi}{N}\underbrace{\left[(1-\psi)g_{\oint}^2 + \psi g_{\oint}^2\right]}_{\text{attenuation}} E_{\text{tot}} = \frac{\psi}{N} g_{\oint}^2 E_{\text{tot}} \tag{5.16}$$

Since $E\left[E_{\text{diff}}\right]$ should be equal to $\frac{\psi}{N}E_{\text{tot}}$, the compensation gain for diffuse sound is

$$g_{\text{diff\_c}} = \frac{1}{g_{\oint}} \tag{5.17}$$

The above gain for the diffuse sound is constant, and the gain for the non-diffuse sound depends only on the smoothly behaving diffuseness value. These gains restore the overall level and the relative levels of the non-diffuse and the diffuse sounds. The imbalances in the spatial distribution of the diffuse sound remain but informal listening tests suggested that this had little or no perceptual significance. This type of compensation is not vulnerable to the artefacts like the accurate compensation in the previous section.

5.4.3 Virtual directional microphones from B-format

The B-format signal (Section 5.1), from which DirAC analysis is typically performed, gives the omnidirectional signal and three figure-of-eight signals measured at a single position. By linear combination of these signals it is possible to create several types of virtual directional microphones:

$$s_n(t) = \frac{2-\kappa}{2}w(t) + \frac{\kappa}{2\sqrt{2}}\left[\cos(\theta_n)\cos(\varphi_n)x(t) + \sin(\theta_n)\cos(\varphi_n)y(t) + \sin(\varphi_n)z(t)\right] \tag{5.18}$$

where s_n is the virtual microphone signal, θ_n the azimuth, φ_n the elevation of the nth loudspeaker and $0 \leq \kappa \leq 2$ is the value defining the directional properties of the virtual microphones. The effect of κ is illustrated in Fig. 5.8. A computationally efficient way of calculating these signals is presented in Section 6.2.

The directional pattern as a function of a spatial angle from the maximum direction is

$$g_{\text{mic}}(\gamma_n, \kappa) = \frac{2-\kappa}{2} + \frac{\kappa}{2}\cos(\gamma_n) \tag{5.19}$$

The microphone gain $g_{\oint}(\kappa)$ for a diffuse field is

$$g_{\oint}(\kappa) = \sqrt{1 - \kappa + \frac{\kappa^2}{3}} \tag{5.20}$$

5.4.4 Choice of the directional pattern of virtual directional microphones

Several types of virtual microphone signals can be created from a B-format signal by altering κ in Eq. (5.18). By informal listening, the perceptually best results were achieved with $\kappa = 2$

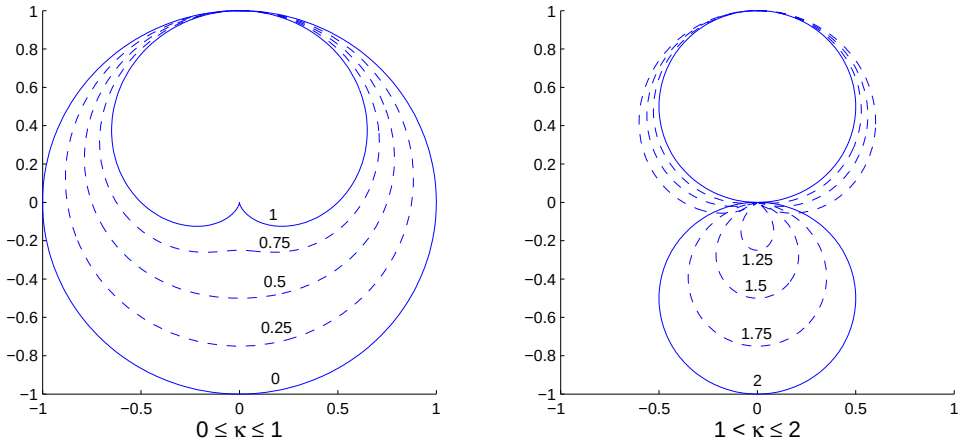

Figure 5.8: The virtual microphone directional pattern can be altered by adjusting κ.

which means that the directional pattern is figure-of-eight. This is a reasonable selection also if one considers that the narrower directional patterns enable better separation of a source and early reflections. The figure-of-eight shape is the narrowest possible pattern that has unity gain in the zero angle and equal or less gain in all other angles.

In DirAC, the back lobes of the figure-of-eight directional patterns have less significance than in passive microphone techniques. In the situation of a highly non-diffuse sound, the back lobes are not problematic since the loudspeaker gains in those directions are zero or close to zero. In the situation of a fully diffuse sound field on the other hand, the back lobes do not pose problems either since the sound will be in any case decorrelated and distributed to all loudspeakers. In fact, if the original diffuse sound field is polarized to an axis, a plane or an ellipsoid, the figure-of-eight of shape can be beneficial since the reproduced diffuse sound is emphasized in the same axes as the original. With moderately diffuse sound fields, it is possible that the back lobes bring undesired effects. The figure-of-eights were used nevertheless due to their superior performance in informal listening tests.

5.5 Temporal averaging

Many parameters such as the loudspeaker gains require temporal averaging to avoid audible distortion due to quick changes. The drawback of applying averaging is that the system can become sluggish. An implementational goal is to tune the averaging windows so that the responses are fast enough for accurate DirAC analysis and synthesis but slow enough to avoid perceivable artefacts.

The averaging window shape is also important. The literary review in [18] implies that there are multiple candidates for the perceptual time integration window. These include a double-sided exponentially decaying window, a Gaussian window and a rectangular window. It was also

suggested that the integration could be different for different auditory processes. Considering the double-sidedness of the suggested window shapes and also that in practical applications it is necessary to use finite length windows unless recursive averaging is utilized, a reasonable approach is to average the data with a Hanning window (Fig. 5.9 right). A perceptually not valid, but computationally efficient approach is to use a first order IIR lowpass filter (Fig. 5.9 left)

$$\hat{a}(t) = \beta_a \hat{a}(t-1) + (1-\beta_a)a(t) \tag{5.21}$$

where $0 \leq \beta_a \leq 1$ is the factor which defines the decay rate of the IIR window and a stands for the parameter to be slowed down.

A causal implementation of these averaging filters causes lag in the estimated parameters. A heuristic solution to this problem is to shift the averaging window so that the gravity point is at the origin. In practical implementations, this means that the input audio must be buffered so that enough data is available for the averaging. For symmetric windows, the gravity point is the center point. For IIR window, the distance of the gravity point from the beginning of the window is formulated with

$$\Delta n_{\beta_a} = -\frac{1}{\log_2(\beta_a)} \tag{5.22}$$

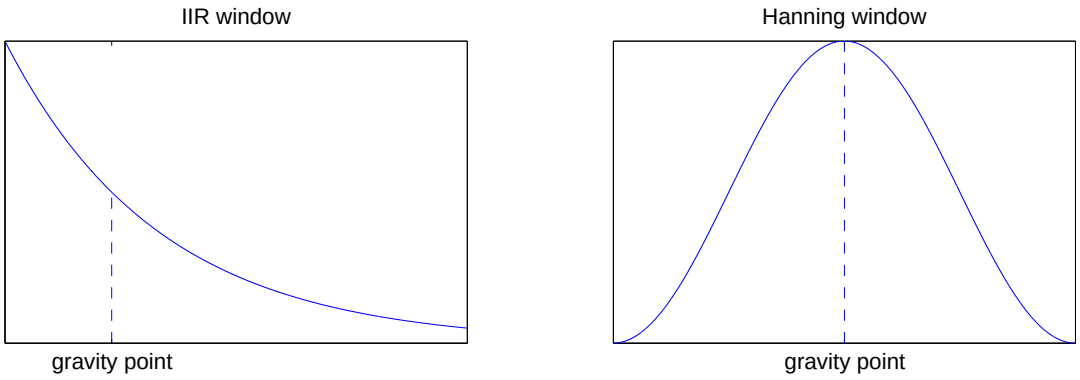

Figure 5.9: The IIR window (left) and the Hanning window (right). These windows can be used in temporal averaging of DirAC parameters.

5.5.1 Temporal averaging of intensity vector for analysis of direction of arrival

The informal listening tests suggest that the averaging of the intensity vector for the analysis of the direction of arrival does not serve any purpose, except in the special case when the direction of arrival dependent gain compensation for virtual microphones (Section 5.4.1) is used. In this case the intensity vector must be averaged and this may result in sluggishness in the reproduction.

5.5.2 Temporal averaging of intensity vector and energy for analysis of diffuseness

Diffuseness (Eq. 2.3) is calculated from the time averages of the sound field intensity and energy. The equation however does not define the length nor the type of the averaging windows. In the point of view of DirAC, the averaging window should be designed to best match the integration in human hearing. Long window length in the intensity and energy averaging guarantees more accurate estimate for the diffuseness with stationary signals, but the diffuseness can be overestimated when the sources move since the intensity vector will have smaller absolute values. Too short averaging windows on the other hand can underestimate the diffuseness.

5.5.3 Temporal averaging of loudspeaker gains

Changes in the analyzed direction of arrival cause fast changes in the loudspeaker gain factors, which in turn cause audible distortion. Averaging of these gains is a straightforward method to remove the distortion. The design principle for the gain averaging window is that it should be just long enough to prevent the artefacts, but as short as possible to minimize the possible sluggishness in the synthesis of the non-diffuse sound. The gain averaging also causes that there are typically more than two or three loudspeakers active simultaneously in the synthesis of the non-diffuse sound.

5.6 Loudspeaker setup

The number and the layout of loudspeakers required for the reproduction of the spatial impression of a diffuse sound field was studied in [41]. The study was performed in an anechoic chamber by comparing the spatial impressions of different loudspeaker layouts with a reference setup of 24 evenly distributed loudspeakers in the horizontal plane. The comparison was performed by subjective listening tests and by inter-aural cross-correlation analysis with a dummy head. The results indicated that with evenly distributed loudspeakers, at least six loudspeakers were needed for the reproduction of a spatial impression indistinguishable from the case of 24 loudspeakers. In the light of these results, it is reasonable to assume that increasing the number of the loudspeakers further from 24 up to infinity does not bring any perceptual difference either. Assuming this to be true, six loudspeakers should be able to produce a sound perceptually equivalent to a horizontal diffuse sound field. An interesting and very convenient result of this study was that the standard 5.0 loudspeaker setup also performed well in this respect.

The three-dimensional loudspeaker setups were not included in the study. DirAC operates also in three dimensions, and therefore this is also a field of interest. Since the loudspeaker density of the three-dimensional 16 channel loudspeaker setup used in our experiments was in all directions of higher density than the six equally distributed horizontal loudspeakers, it is assumed with the same rationale that this 16 channel layout is adequate for the reproduction of

the spatial impression of a three-dimensional diffuse sound field.

The sparseness of the loudspeakers has also an impact on the synthesis of the non-diffuse sound, the accuracy of which largely depends on the accuracy of the VBAP. The performance of VBAP decreases as the angle between the loudspeakers increases. For example, let us consider the standard 5.0 loudspeaker setup in Fig. 5.10. The angle between the frontal loudspeakers, including the center loudspeaker, is $30°$, which gives a relatively precise positioning of the virtual sources in the line in between. The virtual sources between the side front loudspeaker to the closest rear loudspeaker $80°$ apart are less defined. The $140°$ angle between the rear loudspeakers is too large and causes severe ambiguity in localization. Additionally, the listener positioning becomes more relevant as the angle between the loudspeakers increases, since if the listener is close to a loudspeaker that is active in amplitude panning, the virtual source is drawn towards the closer loudspeaker due to the precedence effect.

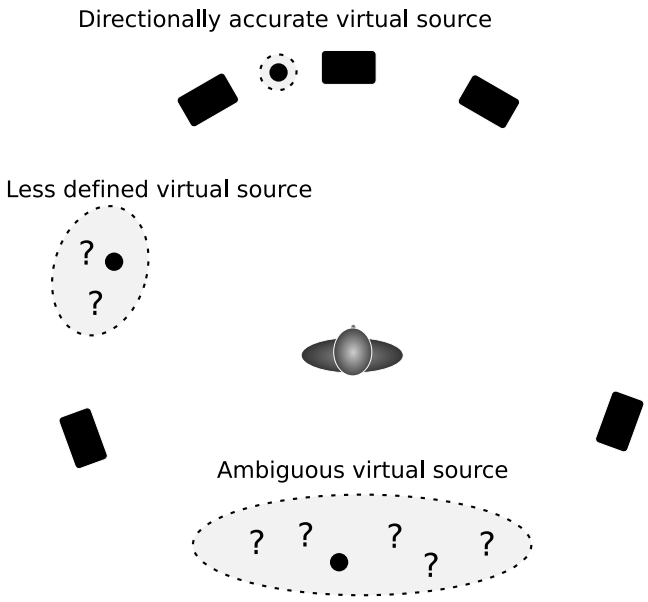

Figure 5.10: An illustration of the precision of Vector Base Amplitude Panning (VBAP) with 5.0 loudspeaker setup.

5.7 Dimensionality and non-surrounding loudspeaker setups

DirAC analyzes the sound field in three dimensions. It is however usual that the loudspeakers are located only in the horizontal plane, and therefore there is a need for a method for dimensionality reduction. If there is access to the original B-format microphone signal, the most straightforward method is to suppress the z-signal (the vertical figure-of-eight microphone) from the input data

before the DirAC analysis. This procedure restricts the intensity vector to the horizontal plane and all vertical energy is analyzed as being part of the diffuse sound. This is the preferred behavior in typical environments where the most of the vertical energy is due to the reverberation and does not contain vital localization information. A localization error will occur in the situations where the original sources are elevated, because of the flattening to the horizontal plane. Also, the diffuseness of the sound field will be overestimated.

Another problematic task is to perform the DirAC synthesis with non-surrounding loudspeaker setups. These are for example a setup with loudspeakers only at non-negative elevations and a setup with all loudspeakers only at frontal directions. In these cases, it is desirable to synthesize the non-diffuse sound normally in the covered directions, but to deal with the non-covered directions in a reasonable manner. A possible solution is to extend the loudspeaker setup virtually as if the loudspeakers would form a fully surrounding sphere or circle, but then map the energy of the non-existing loudspeakers as diffuse sound to the real loudspeakers.

In the implementation of this thesis, only surrounding loudspeaker setups were used, and the z-signal suppression method was used in dimensionality reduction.

Chapter 6

Implementation

A real-time filterbank-based DirAC processing software was implemented for Mac OS X.

6.1 Filterbank design

The design criteria of the filterbank were selected as follows:

1. Bandwidth of each subband is one ERB or otherwise a constant number of ERBs.
2. The filterbank is linear phase. This guarantees predictable behavior in the overlapping sections of adjacent subbands and avoids temporal artefacts.
3. The filterbank is perfect reconstruction. Combining the frequency bands of any signal will produce the same signal, with only a constant delay.

These properties can be fulfilled with filterbank design by the windowing method. The filter is designed in the Fourier transform domain by setting the response of the desired passband frequencies to unity and others to zero (Fig. 6.1 left). The inverse Fourier transform gives a sinc-type time domain response with the center peak at zero. This response is then windowed with a Hanning window (Fig. 6.1 right) to avoid the discontinuities in the edges and delayed to achieve causality. The windowing in the time domain equals to convolution in the frequency domain, and therefore the frequency response will spread in respect to the frequency response of the window. This is disadvantageous especially in the lowest frequencies where the bandwidths are narrow. The perfect reconstruction properties of the filterbank are preserved in the windowing process.

The filter length was selected to be 4096 samples. Increasing window length increases computational complexity but reduces the spectral spreading caused by windowing.

6.2 Computational optimization

Real-time directional analysis and synthesis with a large number of loudspeakers requires computational optimization. The utilization of a direct convolution is problematic for a real-time

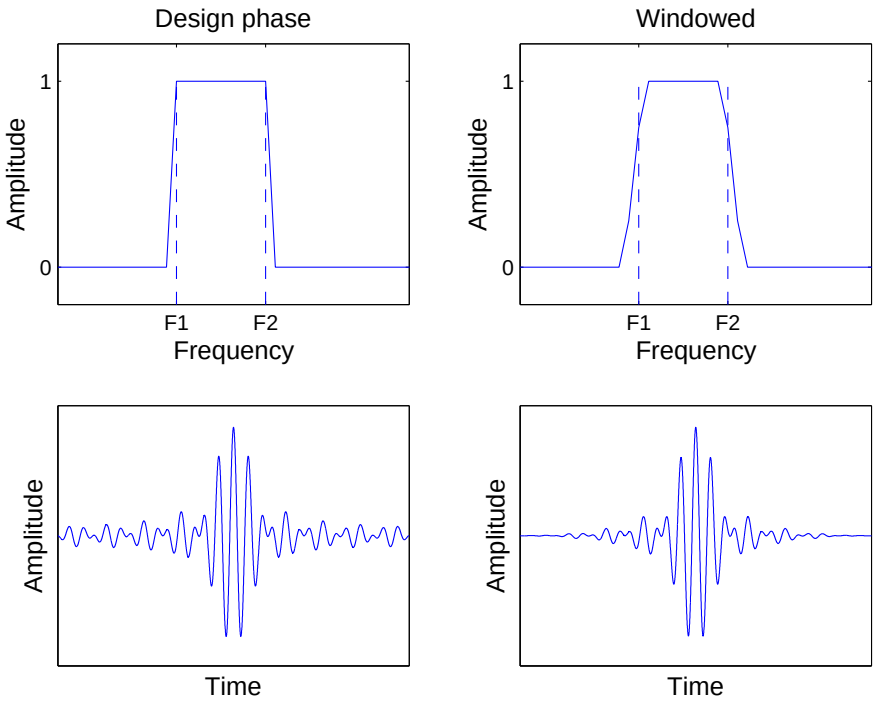

Figure 6.1: The design of a bandpass filter with windowing method. $F1 = 0.005F_s$ and $F2 = 0.007F_s$. The windowing causes spectral spreading.

application since the length of the filterbank filters is several thousand samples. The solution is to use block-wise processing and perform the convolution as a multiplication in the frequency domain by utilizing short time Fourier transform (STFT). An issue in this approach is that the STFT fundamentally considers the signal as being an infinitely long signal repeating itself at the intervals of the block length, and therefore the simple multiplication in STFT domain results as a circular convolution in time domain as in Fig. 6.2.

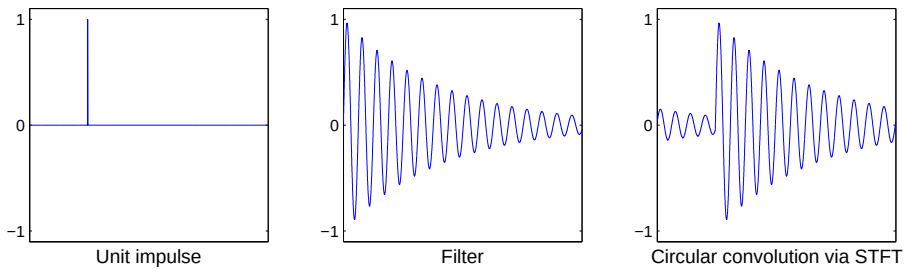

Figure 6.2: A multiplication in STFT domain corresponds to a circular convolution in time domain.

The overlap-save (OLS) method [42] (Fig. 6.3) avoids the circularity problem by doubling the filter length by zero padding. The computational complexity comparison of convolution with OLS method and regular convolution is shown in Fig. 6.4. With a filter length of 4096 samples, the OLS method requires 99.4% less multiplications than a direct convolution. The drawback of the block methods such as the OLS is that they introduce a delay equal to the block length. In these comparisons, the number of multiplications required by a real-valued STFT was $\frac{1}{2}N\log_2(N) - \frac{3}{2}N + 2$, where N is the length of the transform [43].

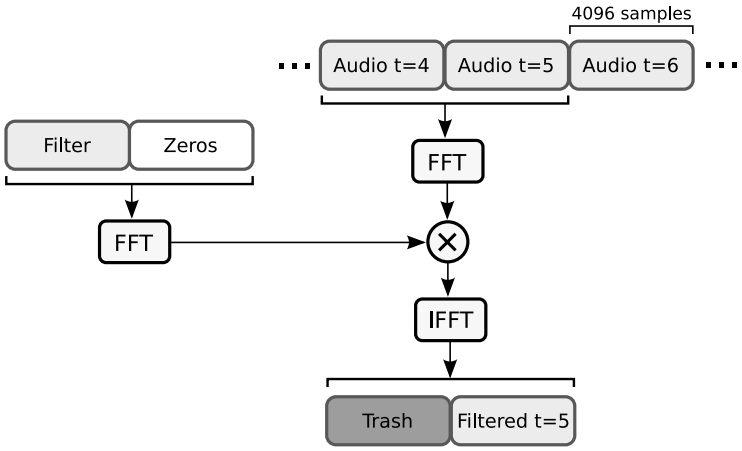

Figure 6.3: Convolution with STFT using overlap-save (OLS) method. Multiplication in frequency domain equals circular convolution in time domain. Circular aliasing is avoided with zero padding of the filter.

Downsampling a frequency band after the filterbank allows a reduction of the number of the calculations in the subsequent processing proportionally to the downsampling factor. In the synthesis part, the downsampled and processed signal is then upsampled with the same factor. A synthesis filter must be applied to attenuate the frequency domain aliasing components. The utilization of the downsampling reduces the total computational complexity to a fraction, allowing the filterbank implementation of DirAC to run in real-time on a modern desktop computer, even with a large number of loudspeakers.

Virtual directional microphone signals from B-format signal (Eq. (5.18)) can be efficiently calculated by taking into account the symmetry in the loudspeaker positioning. For instance, two loudspeakers of elevations φ and $-\varphi$ share the same absolute value of the multiplier for z-channel. More broadly, the z-channel part for all loudspeakers sharing the same $|\varphi|$ can be calculated with only one multiplication per sample. The symmetry in respect to azimuth can also be taken into account correspondingly. The multiplications necessary to calculate the virtual microphone signal at θ azimuth and φ elevation enables the calculation of the virtual microphone signals for total of eight loudspeakers without any additional multiplications. Table 6.1 illustrates

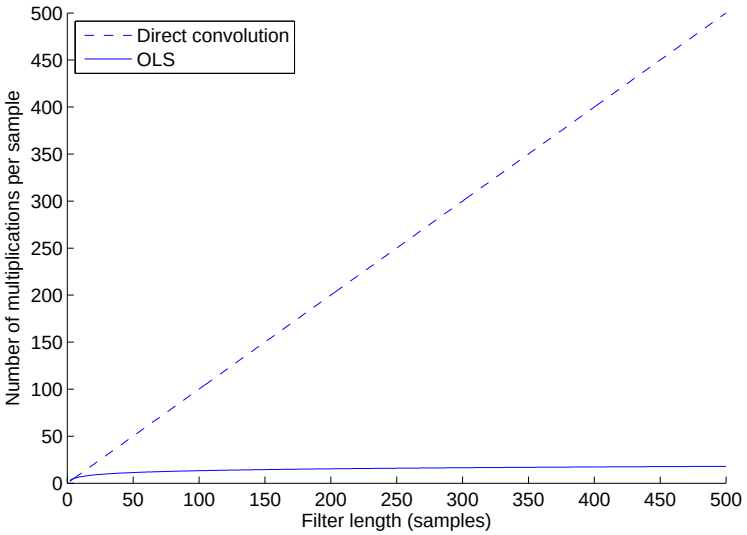

Figure 6.4: Number of arithmetic operations (multiplications and additions) of direct convolution and overlap-save method

the symmetry in an example case of a loudspeaker setup with four loudspeakers.

Table 6.1: B-format signal multipliers for an example loudspeaker setup.

LS azimuth	LS elevation	W multiplier	X multiplier	Y multiplier	Z multiplier
θ_1	φ_1	$\frac{2-\kappa}{2}$	$\frac{\kappa}{2}\cos\theta_1\sin\varphi_1$	$\frac{\kappa}{2}\sin\theta_1\sin\varphi_1$	$\frac{\kappa}{2}\cos\varphi_1$
$-\theta_1$	φ_1	$\frac{2-\kappa}{2}$	$\frac{\kappa}{2}\cos\theta_1\sin\varphi_1$	$-\frac{\kappa}{2}\sin\theta_1\sin\varphi_1$	$\frac{\kappa}{2}\cos\varphi_1$
$\pi-\theta_1$	$-\varphi_1$	$\frac{2-\kappa}{2}$	$-\frac{\kappa}{2}\cos\theta_1\sin\varphi_1$	$\frac{\kappa}{2}\sin\theta_1\sin\varphi_1$	$-\frac{\kappa}{2}\cos\varphi_1$
θ_2	φ_1	$\frac{2-\kappa}{2}$	$(*)\frac{\kappa}{2}\cos\theta_2\sin\varphi_1$	$(*)\frac{\kappa}{2}\sin\theta_2\sin\varphi_1$	$\frac{\kappa}{2}\cos\varphi_1$

(*) unique calculations

6.3 Determination of the length of averaging windows and decorrelation delays

The approximation of the optimal averaging window lengths and the delay lengths for the decorrelation processes is straightforward by using the implemented real-time application in an anechoic chamber for comparing a reference sound created with a virtual reality and a sound reproduced with DirAC. Table 6.2 presents the heuristically selected values and description of the problems encountered with too high or too low values. A graphical representation of these values is in Fig. 6.5.

In terms of sound reproduction quality, it was critical to carefully adjust the range of the de-

lays for decorrelation. The need for precision with gain averaging window length was lesser. Variation of tens of percents of the window length was required to have any perceivable effect. The intensity vector averaging for the analysis of direction of arrival was set to zero since no perceivable benefits was found in selecting otherwise. The intensity vector and energy averaging window lengths for the analysis of diffuseness were noticed to have a very large range of perceptually equally performing values. If the window length is set too low however, the resulting artefacts were loss of spaciousness and bubbling. The maximum window lengths were limited to 200 ms (see Fig. 6.5) to reduce the memory requirements of the non-downsampling implementation.

Table 6.2: Selected Hanning window lengths for averaging and decorrelation delays. T_m is the period time of the middle frequency of the frequency band, N_{band} is the frequency band index, N_{max} is the number of frequency bands and WL is the window length

Parameter	Value	Problems with too low value	Problems with too high value
Gain averaging WL	170*T_m; If less than 50 ms then 50 ms; If more than 200 ms then 200 ms	Bubbling, distortion	Spatial sluggishness, high memory requirements if no downsampling is used
Minimum delay for decorrelation	5 ms	Coherent summation with non-diffuse sound	Echo artefact
Maximum delay for decorrelation	$\frac{N_{\text{band}}}{N_{\text{max}}} * 12$ ms $+ \frac{N_{\text{max}}-N_{\text{band}}}{N_{\text{max}}} * 22$ ms	Loss of spaciousness	Echo artefact
Intensity averaging WL for directional analysis	0 ms	-	Spatial sluggishness
Intensity and energy averaging WL for diffuseness analysis	70*T_m; If more than 200 ms then 200 ms	Underestimation of ψ, lack of spaciousness, bubbling	Memory requirements, possible overestimation of ψ

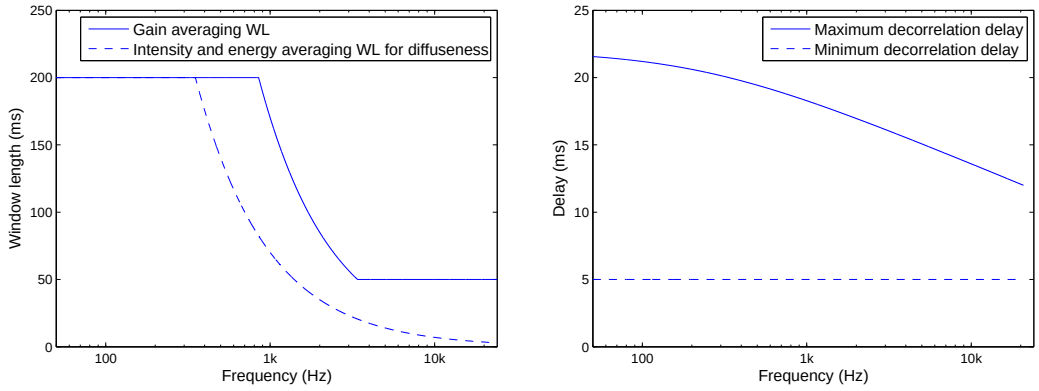

Figure 6.5: The averaging window lengths (Hanning) and decorrelation delay limits that were selected with informal listening and real-time parameter adjustment.

Chapter 7

Experiments

7.1 Properties of the time-frequency transforms

In this section, different time-frequency transforms are compared in terms of frequency resolution and time resolution. The compared transforms were the linear-phase filter bank used in the current DirAC implementation, a gammatone filterbank which approximates the transform of the human cochlea and the STFT, an efficient block-wise transform. The essential purpose of this study was to determine the suitability of the filterbank and the STFT approaches for frequency band audio processing. It should be noted that even though the gammatone filterbank reflects the properties of the ear, it is not an ideal choice in many situations due to following reasons:

1. The off-band attenuation of gammatone filters is low.
2. Gammatone filterbank is not a perfect reconstruction filterbank.
3. The psychoacoustic energy-loudness relation of the equivalent rectangular scale is in many situations more practical than the accurate modeling of the inner ear.

Thus the properties of the gammatone filterbank are not necessarily the design goals for the linear phase filter bank. In this study, the middle frequencies of the gammatone filters were selected to be the center frequencies of one ERB wide frequency bands. The gammatone filters were designed as in [44], with expectation that no adjustment was performed for the phase of the filter. The equation for the filters is

$$h(t, n_{\text{erb}}) = t^3 e^{-1.019\Delta\omega(n_{\text{erb}})t} \cos(\omega_c(n_{\text{erb}})t) \tag{7.1}$$

where n_{erb} is the ERB band index, $\Delta\omega(n_{\text{erb}})$ is the bandwidth and $\omega_c(n_{\text{erb}})$ is the center angular frequency of the band. An $N = 1024$ sample STFT was selected as a comparison to these filterbanks. Sample rate $F_s = 44.1$ kHz was used in the analysis.

To be able to study and illustrate the time-frequency properties of the above transforms, a set of measures are now defined following the procedure in [44]. These are center angular frequency

ω_0, center time t_0, the standard deviation of the frequency response σ_ω and the standard deviation of the impulse response σ_t. These are defined for signal $h(t)$ and its Fourier transform $H(\omega)$:

$$\omega_0 = \frac{1}{\|H\|^2} \int_0^\infty \omega |H(\omega)|^2 d\omega \tag{7.2}$$

$$\sigma_\omega^2 = \frac{1}{\|H\|^2} \int_0^\infty (\omega - \omega_0)^2 |H(\omega)|^2 d\omega \tag{7.3}$$

$$t_0 = \frac{1}{\|h\|^2} \int_{-\infty}^\infty t |h(t)|^2 dt \tag{7.4}$$

$$\sigma_t^2 = \frac{1}{\|h\|^2} \int_{-\infty}^\infty (t - t_0)^2 |h(t)|^2 dt \tag{7.5}$$

The time-frequency properties of an STFT cannot be directly calculated with Eq. (7.2 - 7.5). The time resolution of the STFT was formulated by assuming that the signal is windowed with a N-length Hanning window prior to the transform. The time resolution was acquired by analyzing the Hanning window with Eq. (7.4) and Eq. (7.5). The frequency resolution was simply defined as $\sigma_\omega = n_{\text{bins}} \frac{\pi}{N}$, where n_{bins} is the number of the frequency bins combined to best match the ERB resolution. Figure 7.1 illustrates the time- and frequency resolutions of the discussed transforms and of an overlapping linear phase filterbank which will be explained below.

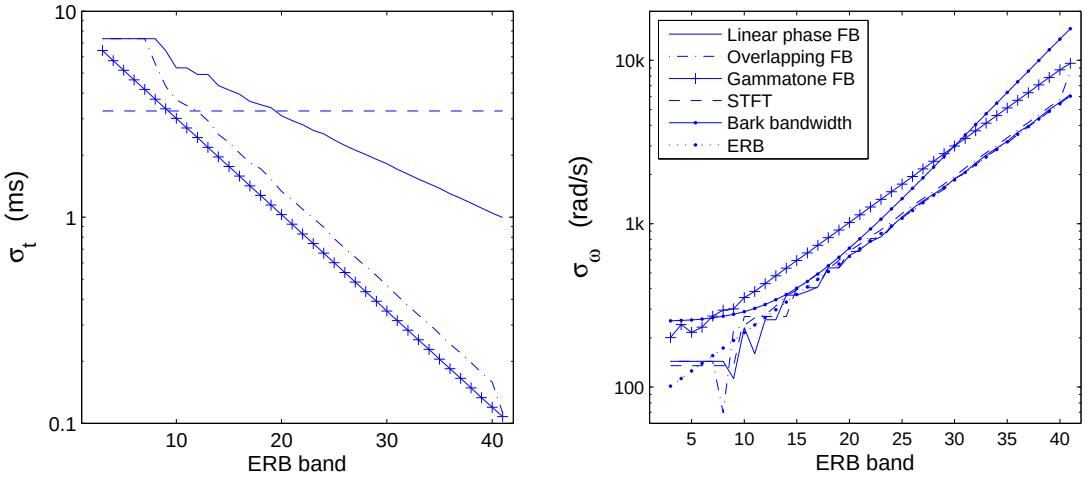

Figure 7.1: The time resolutions (left) and frequency resolutions (right) of different frequency transforms. The zig-zag is due to the precision of the used filter design functions.

From Fig. 7.1 it can be seen that both the STFT and the linear phase filterbank follow closely the ideal ERB scale, with the limitations that the quantization causes zig-zag which is most obvious at the lowest frequency bands. The frequency selectivity of the gammatone filterbank is rougher and somewhat matches the Bark scale.

The time resolution scale (Fig. 7.1 left) shows how the distance between the resolutions of the gammatone filterbank and the linear phase filterbank increases towards higher frequencies. The reason for this behavior is that the transition bandwidths between the filters of the linear phase filterbank remain very narrow throughout the frequency range which leads to longer time domain responses. A test was performed to allow the overlap between the frequency bands to be proportional to the bandwidth of the frequency band. This procedure brought a very desirable time resolution as can be seen from Fig. 7.1. For the current DirAC implementation however, the minimally overlapping filterbank was used since the extensive overlapping causes problems with the used decorrelation method. The problem is that there is an energy loss in the transition bandwidths due to the incoherent summation. In the used decorrelation method the energy loss is minimized by adjusting the delays so that they are in phase in the middle of the transition bandwidth. This approach performs well only if the transition bandwidths are narrow.

7.2 Precision of a Soundfield ST350 B-format microphone

The non-idealities of the directional patterns of real microphones can lead to errors in the directional analysis. In this section, the performance of the Soundfield ST350 is studied using existing impulse response measurements [36]. The measurements were performed in 36 evenly distributed horizontal directions. Since the microphone capsules of ST350 are not located in exactly the same position, the inaccuracy is expectably higher in high frequencies, where the wavelengths are close to the proportions of the microphone.

The directional analysis was performed by dividing the 4-channel B-format impulse responses to ERB bands with a linear-phase filterbank and analyzing the subbands with DirAC. The directional error was defined as the absolute value of the spatial angle between the actual direction and the direction analyzed with DirAC. The average and the maximum values of the directional errors are shown in Fig. 7.2.

The second part of the experiment was to measure the error in diffuseness analysis. A reference horizontal diffuse sound field was created by having 36 uncorrelated virtual white Gaussian noise sources distributed evenly in the horizontal plane. The non-diffuse part was a white Gaussian noise source at zero azimuth. The sound field diffuseness was controlled with the relative energies of the diffuse and the non-diffuse sound. The B-format signal was calculated by convolving the virtual source signals with the corresponding measured impulse responses. The resulting B-format signal was then analyzed using ERB scale linear-phase filterbank, and the diffuseness was analyzed with DirAC. Figure 7.3 shows the analyzed diffuseness and Fig. 7.4 shows the errors, i.e. the distance between actual and analyzed diffuseness.

The results show that the Soundfield ST350 microphone performs relatively well up to 20th to 25th ERB band (1,5 - 3 kHz). In the 20th band, the directional error is in average 6° and maximally 24°. This scale of error is clearly noticeable when a reference signal is available, but is nevertheless within a reasonable range. In higher frequencies, the accuracy both in directional

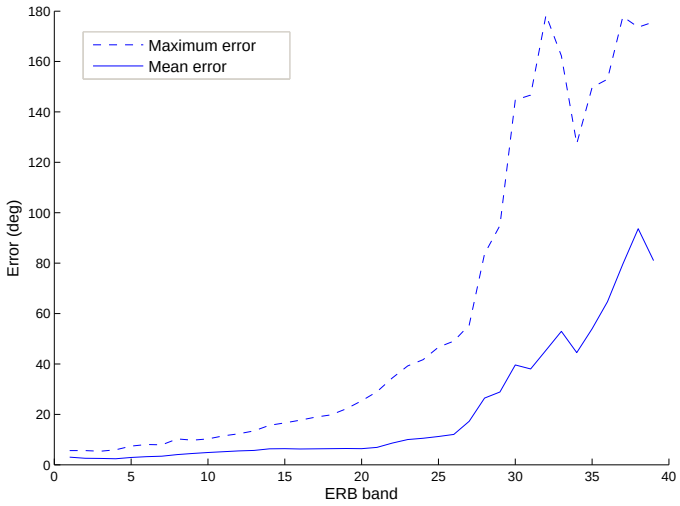

Figure 7.2: Error in directional analysis with Soundfield ST350 microphone.

and diffuseness analysis is poor. The directional error however becomes less significant as the high diffuseness value reduces the energy of the non-diffuse sound. The key result of this study is that regardless of the recorded sound environment, when ST350 is used, the high frequency range will be analyzed as being highly diffuse.

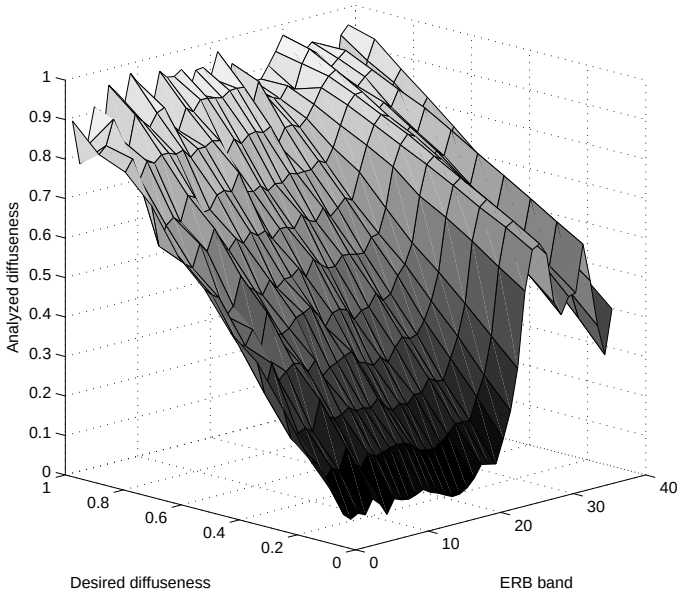

Figure 7.3: Diffuseness analysis with Soundfield ST350 microphone.

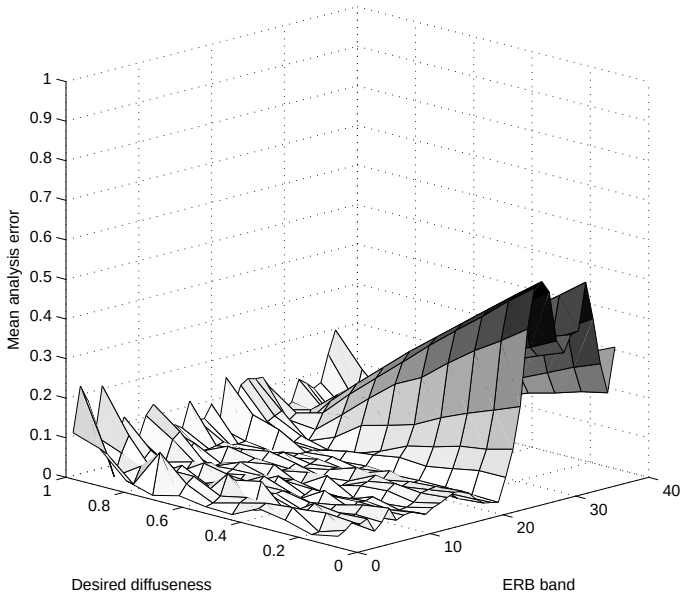

Figure 7.4: Errors in diffuseness analysis with Soundfield ST350 microphone.

7.3 Listening tests

Listening tests were organized to evaluate the perceptual quality of DirAC in sweet spot listening in an anechoic chamber, using reference sounds created with virtual acoustics.

7.3.1 Subjects and test setup

The listening tests included 14 listeners with tested normal hearing (no more than 15 dB hearing loss at any frequency), all of which were participants of a communication acoustics course in Helsinki University of Technology. The participants were awarded with bonus points in course grading. None of the subjects were members of the DirAC development team.

The listening test was designed according to the principles of MUltiple Stimulus and Hidden Reference and Anchor (MUSHRA) [45]. The test subjects were asked to rate the overall reproduction quality of test samples in respect to the reference. Subjects were instructed to consider the sound color, spaciousness, source directions and distances, but were instructed to give an overall rating including all perceivable aspects according to their personal preferences. During the listening tests, subjects were sitting in the anechoic chamber so that their head was located at the center point of the loudspeakers and were encouraged to rotate their heads while evaluating the test samples.

An implemented listening test software (Fig. 7.8) allowed the users to freely switch between different test samples and the reference on the run, browse the playing sound and to start and stop the playback. The test subjects interfaced with the software with a touch screen installed on

a tripod. The test was conducted in dark to minimize the visual cues. The order of test cases and the samples was randomized for each subject.

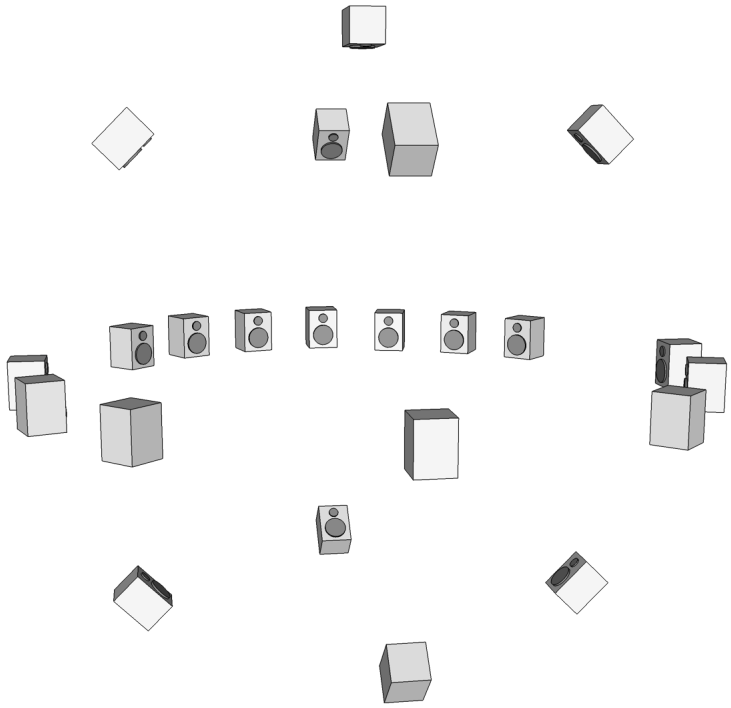

Figure 7.5: The loudspeaker configuration in the anechoic chamber.

The listening tests were performed in an anechoic chamber using 23 Genelec 8030A loudspeakers distributed on a sphere according to Fig. 7.5. This setup was utilized in four different configurations: 21-channel sphere, 16-channel sphere, standard 5.0 surround and quadraphonic 4.0. These configurations are illustrated in Figs. 7.6 and 7.7. The loudspeaker gains were adjusted within 1 dB range with measurements and the distance differences were compensated with channel delays. The angles of the loudspeakers in the 16-channel setup were measured with a theodolite. The measurements are listed in Table 7.1.

7.3.2 Reference stimuli

The 21-channel configuration was used for playing the reference sounds which were virtual sources located in virtual spaces. The loudspeaker density was higher in the frontal directions, since the virtual sources were located in the frontal horizontal plane in all cases except one. The better resolution guarantees that the direct sounds and early reflections of the virtual rooms are less quantized into same loudspeakers, and thus some unnatural coloration due to comb filtering is avoided.

The early reflections up to fourth reflection of three different virtual spaces were created with

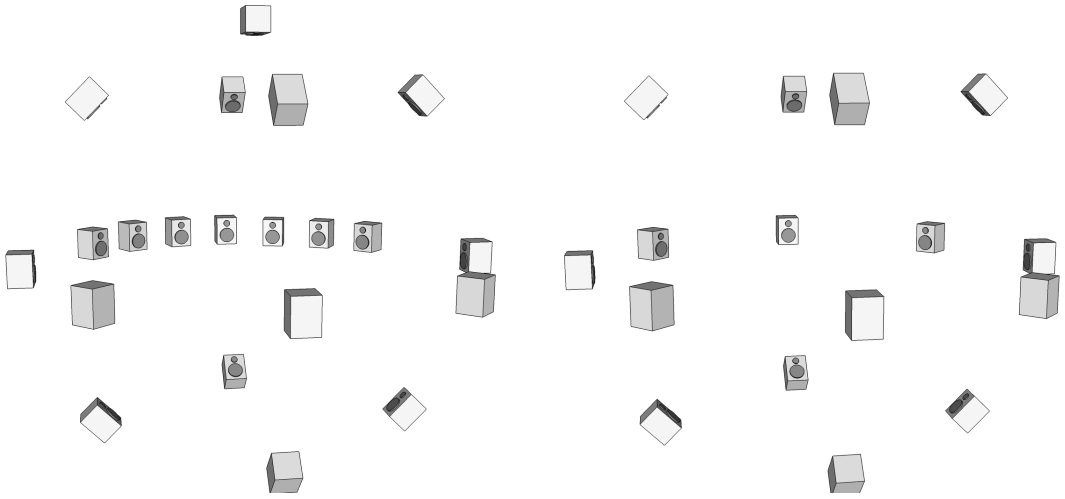

Figure 7.6: The three dimensional setups: 21-channel sphere and 16-channel loudspeaker sphere.

Figure 7.7: The horizontal setups: 5.0 surround and quadraphonic 4.0.

Digital Interactive Virtual Acoustics (DIVA) software [46] using image source method by Tapio Lokki from Helsinki University of Technology. The direct sound and the reflections were quantized to the closest loudspeakers to avoid the usage of amplitude panning techniques. The late reverberation was simulated with exponentially decaying white Gaussian noise with a decay rate for each octave band according to the reverberation times of the modeled rooms. A fourth room type, a reverberation hall, was created by using the early reflections of a concert hall but by multiplying the reverberation times by three. The reverberation hall was included for evaluating the perceptual reproduction capabilities of DirAC in the situation of diffuse reverberation. The onset of the late reverberation was tuned so that it smoothly fades in as the early reflections become sparse, so that the energy decay rate of the resulting impulse response is approximately constant at all time intervals. The reverberation times of the virtual spaces used in the listening tests are plotted in Fig. 7.9, and the impulse responses are plotted in Fig. 7.10.

The listening test consisted of 11 different reference stimuli representing different types of sounds in different acoustic environments. These stimuli are described in Table 7.2. The snare drum fill was included since it was noticed to be critical in revealing the possible defects in DirAC synthesis. The music sample, a 31 second clip from *Beatles - Being for the Benefit of Mr.*

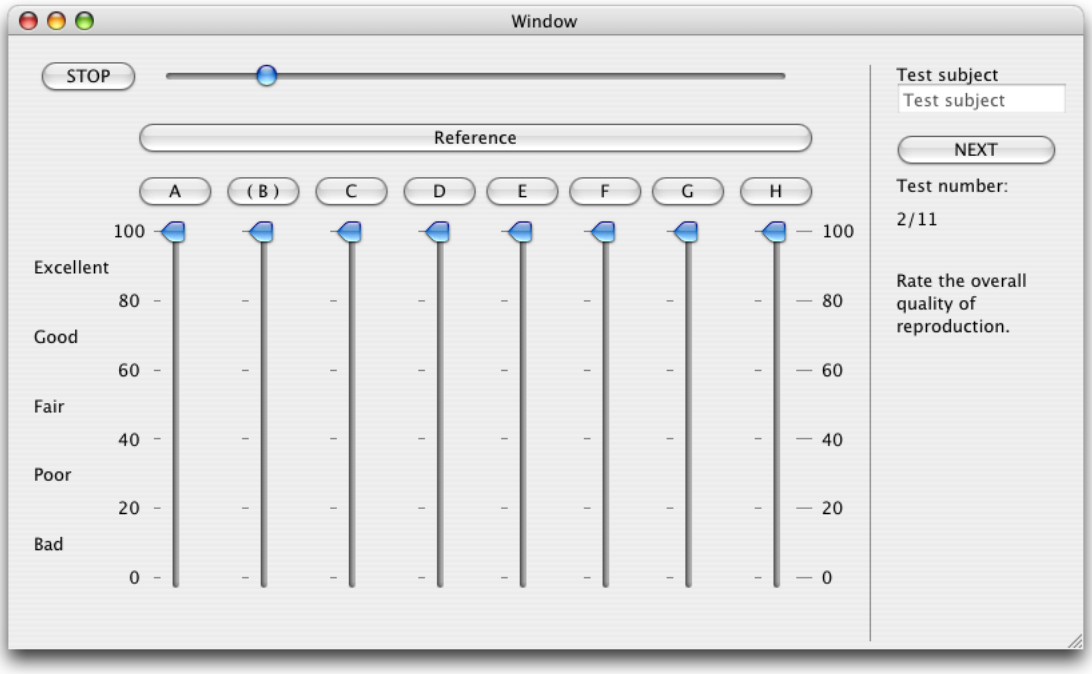

Figure 7.8: User interface of the listening test software.

Table 7.1: Locations and angular distances from the desired locations of the loudspeakers in the 16-channel setup.

Index	Azimuth	Elevation	Dislocation
1	0.0°	0.0°	0.0°
2	45.6°	0.0°	0.6°
3	91.1°	0.0°	1.1°
4	137.6°	0.0°	2.6°
5	180.8°	0.0°	0.8°
6	228.2°	0.0°	3.2°
7	271.3°	0.0°	1.3°
8	315.9°	0.0°	0.9°
9	358.1°	-46.7°	2.1°
10	90.8°	-45.7°	0.9°
11	180.4°	-46.9°	1.9°
12	270.2°	-46.2°	1.2°
13	356.9°	49.6°	5.2°
14	89.7°	49.8°	4.8°
15	180.1°	47.8°	2.8°
16	266.1°	46.1°	2.9°
		Mean	2.0°

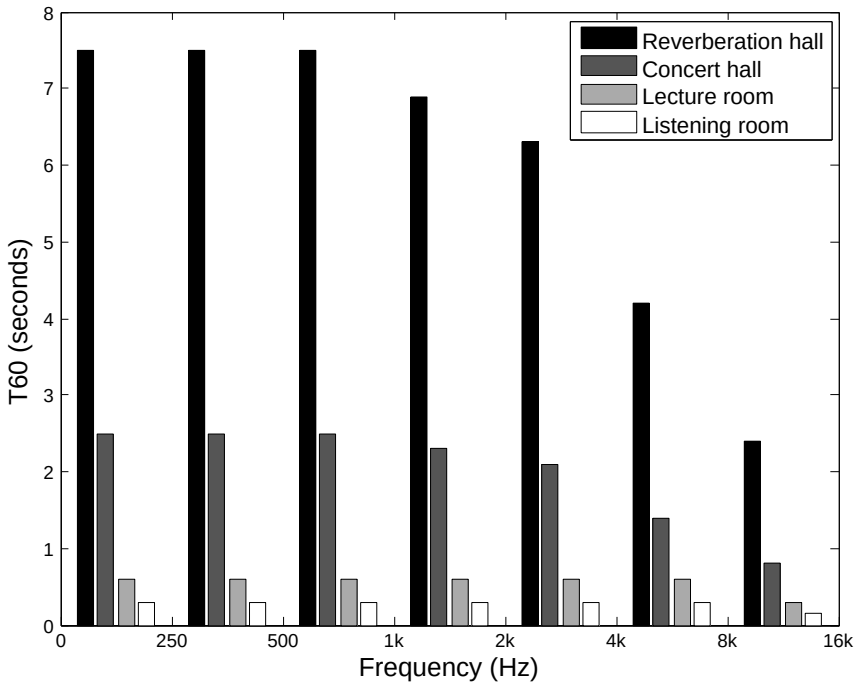

Figure 7.9: Reverberation times (T_{60}) of the virtual rooms. The reverberation hall is geometrically identical to the concert hall, but with exactly three times longer reverberation time.

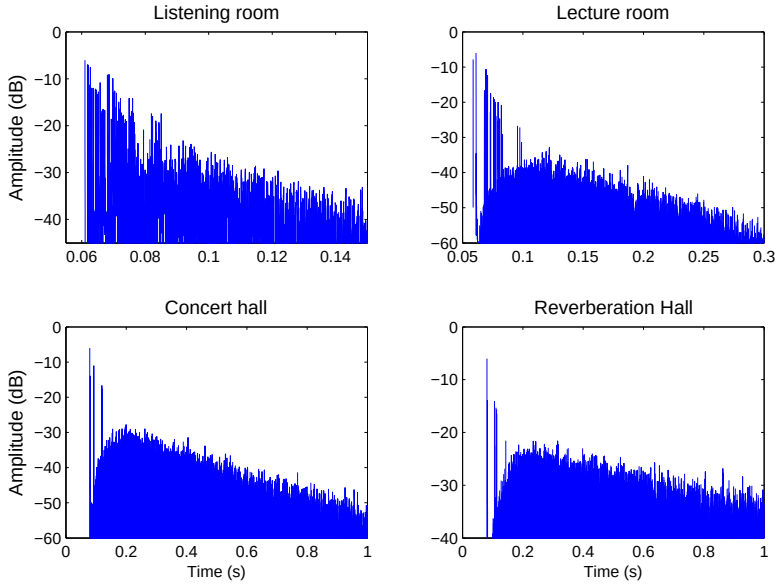

Figure 7.10: Impulse responses of the virtual rooms. Notice the different axes.

Kite!, was selected because the majority of instruments in the piece were panned completely to either of the channels, thus having a strong stereo image. The test cases with one source had the source located in zero azimuth. In all test cases with two active sources, the sources were located at $\pm 30^\circ$. In Dry_speech, two female speakers and one male speaker were located in front, 90° right and 90° left. In Hall_orch and in Rev_hall the separate instruments were positioned in the virtual stage according to the traditional setup with real orchestra. The listening position was selected so that the leftmost instrument was at -30° and the rightmost at 30°.

Table 7.2: Reference stimuli.

Abbreviation	Room type	Source data	Number of sources
List_snare	Listening room	Snare drum fill	1
ListR_music	Listening room	Stereophonic music	2
ListR_sing	Listening room	Singers	2
LectR_snare	Lecture room	Snare drum fill	1
LectR_music	Lecture room	Stereophonic music	2
LectR_sing	Lecture room	Singers	2
Hall_snare	Concert hall	Snare drum fill	1
Hall_orch	Concert hall	Orchestra	14
Hall_sing	Concert hall	Singers	2
Rev_hall	Reverberation hall	Orchestra hits	14
Dry_speech	Anechoic chamber	Speech	3

7.3.3 Test stimuli

The test stimuli were created by reproducing the reference stimuli with five different methods, and including a hidden reference and two anchors (Table 7.3). The DirAC processing was performed with the implemented software. The Ambisonics decoding was performed with a command line decoder utility [47] with default settings. The loudness levels of each sample were adjusted by three expert listeners to best match the reference.

The DirAC was run on default settings in all test cases except Dry_speech, where the maximum decorrelation delays were shortened to 10 milliseconds. Without this procedure, the decorrelation process would have been audible as increased spaciousness. Downsampling was disabled and the samples were processed offline.

The ST350 microphone signals were equalized before DirAC processing. The equalization filter was a constant minimum phase IIR filter which was designed to flatten the power sum spectrum of a selection of impulse responses of the ST350 microphone. The selection was the responses from those loudspeakers that were used in the DirAC synthesis to the virtual figure-of-eight shaped microphones pointing towards the loudspeakers in question. The selection of the responses for equalization was therefore different for D16_st350 and D5_st350. The virtual figure-of-eight pattern was used since it was also used in DirAC synthesis. This procedure also equalizes the effect of double listening the loudspeaker, i.e. that the sound is first reproduced with

Table 7.3: Compared test stimuli.

Abbreviation	Description
Ref	Hidden reference
D16_ideal	DirAC using a simulated B-format microphone and a three-dimensional 16 channel loudspeaker setup
D16_st350	DirAC using a Soundfield ST350 microphone and a three-dimensional 16 channel loudspeaker setup
D5_st350	DirAC using a Soundfield ST350 microphone and a 5.0 channel loudspeaker setup
A4	Ambisonics using a simulated B-format microphone and a quadraphonic loudspeaker setup
A16	Ambisonics using a simulated B-format microphone and a three-dimensional 16 channel loudspeaker setup
LP	Anchor 1: The reference sound lowpass filtered with a second order Butterworth lowpass filter with passband frequency at 2000 kHz
Mono	Anchor 2: All reference channels summed together and played back on the front loudspeaker

the loudspeakers, recorded with a microphone and then again reproduced with the loudspeakers.

7.4 Results

The results of individual subjects were normalized with respect to mean and standard deviation according to the recommendation ITU-R BS.1284-1 [48]. The effect of this normalization can be seen in Fig. 7.11. The overall results with 95% confidence intervals are shown in Fig. 7.12 and the results of individual test cases are shown in Fig. 7.13.

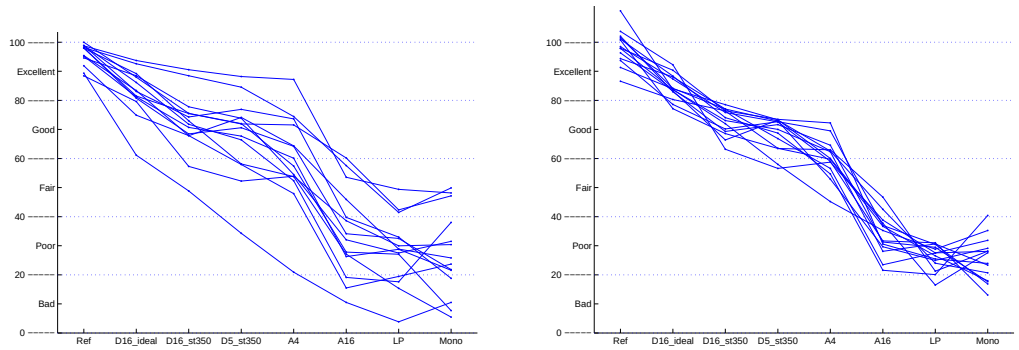

Figure 7.11: The mean perceived quality of reproduction of individual subjects before and after the normalization with respect to the mean and the standard deviation according to the recommendation ITU-R BS.1284-1 [48].

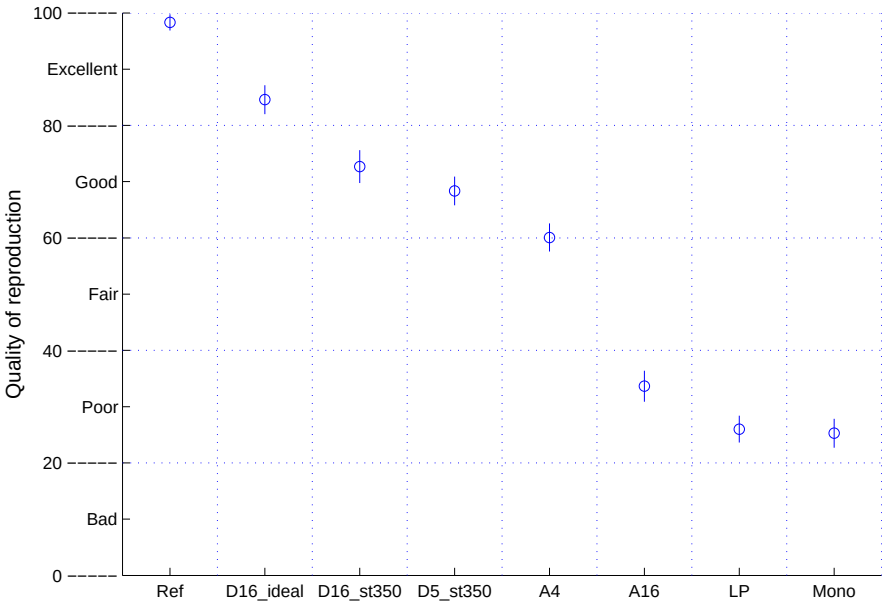

Figure 7.12: The mean perceived quality of reproduction with 95% confidence intervals.

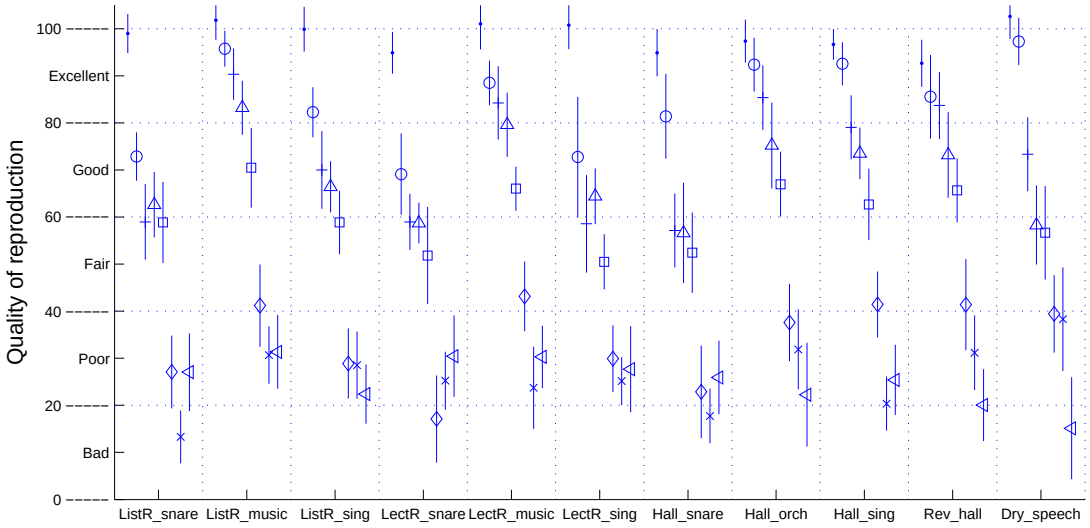

Figure 7.13: The mean perceived quality of reproduction of individual test cases with 95% confidence intervals: Ref (•), D16_ideal (○), D16_st350 (+), D5_st350 (△), A4 (□), A16 (◊), LP (×), Mono (◁).

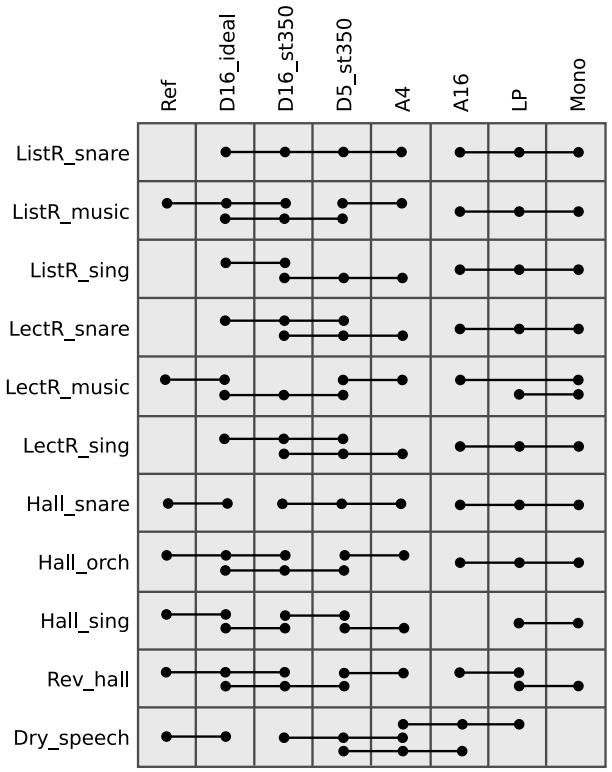

Figure 7.14: The results of the multiple comparison analysis with one-way ANOVA. A dot connection denotes that the means are not significantly different.

The one-way ANOVA multiple comparisons function of Matlab was used for the analysis of statistical differences between the reproduction methods. The overall scores of the reproduction technologies were significantly different from each other except the D16_st350 in respect to D5_st350 and LP in respect to Mono.

Multiple comparison analysis was also performed for each test case separately. The results of these comparisons are illustrated in Fig. 7.14. The results indicated that D16_ideal was significantly different in respect to the reference only in four of the eleven scenarios. In all test cases, all DirAC configurations performed equally or better than the Ambisonics configurations. In these comparisons however, it should be noted that the populations of each item was only 14, thus leading to lower number of differences between populations.

7.4.1 Discussion

The situation with 16 loudspeakers and the usage of simulated microphone is a beneficial setting for DirAC to operate. The setting in anechoic chamber on the other hand allows precise perceptual comparison due to the absence of reverberation which could mask the possible arte-

facts caused by the reproduction technologies. Therefore it can be stated that with excellent recording and reproduction hardware, DirAC can excellently reproduce the typical spatial sound environments.

The setup of 16 loudspeakers in three dimensions in an anechoic chamber however is a rare listening condition. Furthermore, in other scenarios than channel upmix of existing recordings, the real recordings are performed with a real microphone instead of a simulated one. The results indicated that if the ideal microphone is replaced with a Soundfield ST350, the perceived quality drops from excellent to good. This underlines the expected fact that the perceptual quality of DirAC depends directly on the quality of the microphone.

There was no statistical degradation in the perceived quality of reproduction when the ST350 microphone was still used, but when the reproduction was performed with the standard 5.0 setup instead of the 16 loudspeaker sphere. Considering this result and the fact that the room responses were three-dimensional, it can be concluded that the dimensionality reduction is of low relevance in the spatial perception of the room properties as long as the sources themselves are not elevated. The result also shows that DirAC scales very well to different loudspeaker setups. Both of these properties are very practical in a typical end-user point of view.

The DirAC outperformed the first-order Ambisonics in all tested configurations even with the handicap that a real microphone was used instead of the simulated one. The poor result of the 16-channel Ambisonics can be explained with the high coherence between the channels, which leads to severe coloration effects in the listening point.

Chapter 8

Conclusions and Future Work

This thesis started with an introduction to the current trends in the reproduction of spatial audio. The following chapters explained the relevant physical and perceptual aspects of sound, the reproduction methods and the principles, implementation and subjective evaluation of DirAC. Additional experiments were performed on the properties of the time-frequency transforms and on the effect of utilization of a Soundfield ST350 microphone instead of a simulated, ideal microphone.

DirAC is a perceptual DSP-oriented technology which functions as a layer on top of the coincident microphone techniques. This thesis focused on the first order B-format microphones. DirAC was shown to bring more accurate sound reproduction and scalability to arbitrary loudspeaker setups in comparison to Ambisonics, which represented the coincident microphone techniques. Although not proven in this thesis, similar enhancement should be achieved also by applying DirAC to higher order microphones.

The listening tests were conducted in an anechoic chamber. The results indicated that with these stimuli the perceptual quality of DirAC with an ideal B-format microphone and 16 loudspeakers was excellent. With a real Soundfield ST350 microphone, the perceptual quality of DirAC was good, both with the 16 channel setup and the standard 5.0 setup. The ST350 microphone was equalized with a filter derived from impulse response measurements. Quadraphonic and 16 channel first order Ambisonics using ideal microphones were tested as comparison. The quadraphonic version performed fair/good and the 16 channel version performed poorly.

The directional properties of the ST350 microphone were shown to be distorted in the frequencies above 1,5-3 kHz. This affects the DirAC analysis and synthesis by causing the overestimation of diffuseness, which is audible as additional undesired ambience in the corresponding frequency range.

The study of different time-frequency transforms showed that there are tradeoffs between the transform properties including the time resolution, the frequency resolution and the narrowness of the transition bandwidth. Since in this application it was required to have minimum overlap between the frequency bands, and also that the filterbank had to be linear phase, the result was

that the time resolution of the designed filterbank was clearly lower than the resolution of the gammatone filterbank, which represented the human hearing resolution.

This thesis showed that good quality reproduction of spatial audio is possible with a single B-format microphone and a set of five loudspeakers, a setting that is already relatively common as part of home theaters. DirAC is not restricted to any loudspeaker setup, but liberates the users to design their own setup freely. Furthermore, DirAC functions fully automatically and is therefore easily applicable to non-professional purposes as well.

The future work includes listening tests in a listening room, both in sweet spot and off sweet spot. These tests are performed to establish knowledge how well DirAC performs in typical real-life conditions.

Bibliography

[1] S. P. Lipshitz, "Stereo microphone techniques... are the purists wrong?," *Journal of the Audio Engineering Society*, vol. 34, no. 9, pp. 716–744, 1986.

[2] F. Rumsey, *Spatial Audio*. Focal Press, 2001.

[3] V. Pulkki and C. Faller, "Directional audio coding: Filterbank and STFT-based design," *120th AES Convention*, Paris, 2006.

[4] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, 2007.

[5] F. Baumgarte and C. Faller, "Binaural cue coding - part I: psychoacoustic fundamentals and design principles," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 509–519, November 2003.

[6] F. Baumgarte and C. Faller, "Binaural cue coding - part II: schemes and applications," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 520–531, November 2003.

[7] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, and P. Kroon, "Spatial audio coding: Next-generation efficient and compatible coding of multi-channel audio," *117th AES convention*, San Fransisco, 2004.

[8] I. 14496-3:2001/Amd.3:2004, "Information technology - coding of audiovisual objects - part 3: audio."

[9] M. M. Goodwin and J.-M. Jot, "A frequency-domain framework for spatial audio coding based on universal spatial cues," *120th AES Convention*, Paris, 2006.

[10] C. Faller, "Multiple-loudspeaker playback of stereo signals," *Journal of the Audio Engineering Society*, vol. 54, pp. 1051–1064, November 2006.

[11] L. E. Kinsler and A. R. Frey, *Fundamentals of acoustics*. John Wiley & Sons, Inc., 1950.

[12] F. A. Everest, *Master handbook of acoustics*. McGraw-Hill, 2001.

[13] F. Fahy, *Sound Intensity*. Elsevier Science Publishers Ltd., Essex, England, 1989.

[14] D. Botteldooren, "Finite-difference time-domain simulation of low-frequency room acoustic problems," *Journal of the Acoustic Society of America*, vol. 98, pp. 3302–3308, December 1995.

[15] H. Kuttruff, *Room Acoustics*. Elsevier Applied Science, London, Uk, 3rd edition, 1991.

[16] J. B. Allen and D. A. Berkeley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustic Society of America*, vol. 65, pp. 943–950, April 1979.

[17] J. Blauert, *Communication Acoustics*. Springer-Verlag Berlin Heidelberg, 2005.

[18] J. Merimaa, *Analysis, synthesis, and perception of spatial sound - binaural localization modeling and multichannel loudspeaker reproduction*. Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 2006.

[19] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[20] E. Zwicker, G. Flottorp, and S. Stevens, "Critical bandwidth in loudness summation," *Journal of the Acoustic Society of America*, vol. 29, pp. 548–557, 1957.

[21] L.A.Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology*, vol. 41, pp. 35–39, 1948.

[22] Lord Rayleigh, "On or perception of sound direction," *Philosophical Magazine*, vol. 13, pp. 214–232, 1907.

[23] B. C. J. Moore, *An introduction to the psychology of hearing*. Academic press, 1997.

[24] J. Blauert, *Spatial Hearing*. The MIT Press, Cambridge, Massachusetts, USA, revised edition, 1997.

[25] *USA Standard Acoustic Terminology*. American National Standards Institute, 1960.

[26] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, pp. 456–466, June 1997.

[27] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *Journal of the Acoustic Society of America*, vol. 43, pp. 300–321, May 1995.

[28] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Evaluation of artificial heads in listening tests," *Journal of the Audio Engineering Society*, vol. 47, pp. 83–100, March 1999.

[29] 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," *Journal of the Acoustic Society of America*, vol. 49, pp. 904–916, October 2001.

[30] W. Hess, "Influence of head-tracking on spatial perception," *117th AES convention*, San Fransisco, 2004.

[31] W. R. Thurlow and P. S. Runge, "Effect of induced head movements on localization of direct sounds," *Journal of the Acoustic Society of America*, vol. 42, pp. 480–488, April 1967.

[32] D. Begault, *3-D sound for virtual reality and multimedia*. NASA, 2000.

[33] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *Journal of the Acoustic Society of America*, vol. 93, pp. 2764–2778, May 1993.

[34] M. J. Gerzon, "Periphony: with height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, pp. 2–10, January/February 1973.

[35] "St350 portable microphone system, user manual."

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

[37] D. Gabor, "Theory of communication," *J. IEE (London)*, vol. 93, pp. 429–457, November 1946.

[38] J. Merimaa and V. Pulkki, "Spatial impulse response rendering II: Reproduction of diffuse sound and listening tests," *Journal of the Audio Engineering Society*, vol. 54, pp. 3–20, January/February 2006.

[39] J. Engdegård, H. Purnhagen, J. rödén, and L. Liljeryd, "Synthetic ambience in parametric stereo coding," *116th AES convention*, Berlin, 2004.

[40] M. Bouéri and C. Kyirakakis, "Audio signal decorrelation based on a critical band approach," *117th AES convention*, San Fransisco, 2004.

[41] K. Hiyama and S. Komiyama, "The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field," *113th AES convention*, Los Angleles, 2002.

[42] A. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[43] H. Sorensen, D. Jones, M. Heideman, and C.Burrus, "Real-valued fast fourier transform algorithms," *IEEE Transactions on Acoustics and Speech Signal Processing*, pp. 849–863, 1987.

[44] R. Hut, M. M. Boone, and A. Gisolf, "Cochlear modeling as time-frequency analysis tool," *Acta acustica united with acustica*, vol. 92, pp. 629–636, 2006.

[45] "Recommendation ITU-R BS.1534-1 method for the subjective assessment of intermediate quality level of coding systems."

[46] T. Lokki, L. Savioja, R. Väänänen, R. Huopaniemi, and T. Takala, "Creating interactive virtual acoustic environments," *Journal of the Audio Engineering Society*, vol. 47, pp. 675–705, Sept 1999.

[47] "http://www.muse.demon.co.uk/utils/ambidec.html."

[48] "ITU-R BS.1284-1 general methods for the subjective assessment of sound quality."