Analysis and Re-Synthesis of Directional Spatial Impulse Responses

Diploma Thesis

written by

Klaus Hostniker

Institute of Electronic Music and Acoustics (IEM)
University of Music and Performing Arts
Graz, Austria

University of Technology
Graz, Austria

Supervisor: DI Dr.techn. Alois Sontacchi
Assessor: O.Univ.Prof. Mag.art. DI Dr.techn. Robert Höldrich

Graz, April 2011

Acknowledgements

First I would like to thank Alois Sontacchi for the countless hours of creative and humorous conversations and discussions.

I would like to thank the whole team of the IEM.

I would like to thank Joschi for the help with all my computer problems and upgrades, and James for proofreading this thesis.

Mein größter Dank geht an meine Freundin und Lebensgefährtin Sabine, die mir stehts Rückhalt und Motivation gegeben hat.

Ich widme diese Arbeit meinen Eltern Rosa Maria und Karl-Heinz. Ohne eure Unterstützung wäre dieses Studium und diese Arbeit so nicht möglich gewesen.

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am …… ……..

(Unterschrift)

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, …..…… ……..……..

date (signature)

Abstract

In this thesis the analysis of measured directional spatial impulse responses and their re-synthesis with improved spatial image sharpness is discussed. For both the direct sound and early reflections, the time of arrival is determined by the detection of transients occurring in the impulse responses. Several approaches are compared. Further analysis and re-synthesis is based on the principle of Directional Audio Coding (DirAC). In the analysis phase, the frequency dependent directional information of the transients and the diffuseness of the signals are established by means of energy calculations. The measured results are compared with a room-acoustic simulation. Furthermore, it is shown that the SoundField SPS200 Ambisonics B-Format microphone, used in this thesis, shows deviations in the direction detection of source positions due to differences in recording quality. These deviations occur at higher frequencies, when half the wavelength of the incident sound field coincides with the distance between the microphone capsules, or the sound incidence is 45°. The data obtained from the analysis is applied to the re-synthesis. During re-synthesis, encoding of the impulse responses in higher order Ambisonics is focused on particularly. The non-diffuse parts of the impulse responses are re-synthesized through different weighting of the Ambisonics channels, and the diffuse parts are re-synthesized independently by decorrelation. Furthermore, a possible transfer of room acoustic properties is discussed. In addition to capturing both the acoustic properties of the source room and the reproduction room, additional provisions and assumptions need to be taken into account.

Kurzfassung

In dieser Diplomarbeit wird die Analyse von messtechnisch erfassten gerichteten Raumimpulsantworten und deren Resynthese mit verbesserter räumlicher Abbildungsschärfe behandelt. Die Zeitpunkte des Eintreffens von Direktschall und den frühen Reflexionen in den Impulsantworten werden durch die Detektion von auftretenden Transienten bestimmt. Mehrere Verfahren werden miteinander verglichen. Die weitere Analyse und Resynthese basiert auf dem Prinzip von Directional Audio Coding (DirAC). Frequenzabhängig werden in der Analysephase die Richtungsinformationen der Transienten und die Diffusität in den Signalen durch energetische Berechnungen ermittelt. Die ermittelten Messergebnisse werden mit einer raumakustischen Simulation verglichen. Weiters wird gezeigt, dass das zur Aufnahme verwendete Soundfield SPS200 Ambisonics B-Format Messmikrofon Abweichungen in der Richtungsdetektion von Quellpositionen auf Grund unterschiedlicher Aufnahmequalität zeigt. Diese Abweichungen treten bei höheren Frequenzen auf, wenn die Hälfte der Wellenlänge des einfallenden Schallfeldes dem Abstand zwischen den Mikrofonkapseln ähnlich wird, oder der Schalleinfallswinkel 45° beträgt. Die gewonnen Messdaten aus der Analyse kommen in der Resynthese zur Anwendung. Ein besonderer Schwerpunkt innerhalb der Resynthese liegt in der Kodierung der Impulsantworten in Ambisonics höherer Ordnung. Die Nicht-Diffusanteile der Impulsantworten werden durch unterschiedliche Gewichtung der Ambisonics Kanäle resynthetisiert, die Diffusanteile unabhängig davon durch Dekorrelation. Des Weiteren wird ein möglicher Transfer von raumakustischen Eigenschaften diskutiert. Neben der Erfassung der raumakustischen Eigenschaften des zu übertragenden Raumes, als auch jene des Reproduktionsraumes müssen zusätzliche Vorkehrungen und Annahmen getroffen werden.

Table of Contents

1 Introduction

1.1 Motivation

With regard to visual impulses, humans are limited to the frontal direction. With aural stimulus, it is however possible to detect and localize sources from any direction. Directional hearing plays an unobtrusive but important role in daily life. It is however also becoming increasingly important in the presentation and perception of music in every way imaginable. The ability to discriminate between incoming sound events in left-right direction is quite precise, but also the ability to identify whether a source is located in front, behind, below or above is possible. Hearing in three dimensions works well, and spatial properties and aspects significantly influence the pleasure of listening. The listener is able to roughly estimate the properties of the acoustic environment and can distinguish whether music is performed in a small room, a large reverberant room or even in an anechoic environment.

Researching spatial analysis and spatial re-composition is motivated by practical reasons. In particular, in more modern productions, with a focus on the composition of electronic music, integrate spatial sound design, and spatial effects are attributed to great importance. Sound rendering techniques such as Ambisonics (cf. [1]) have been shown to provide great improvements for such applications. This rendering technique simplifies the adaptation to various loudspeaker layouts at different event locations, by failing to consider any further changes to the defined sound object space-time-trajectories. However, several realizations of such perceivable trajectories depend on the concrete available room acoustics. Often, compositional effects and ideas can only be realized with difficulty, due to the differences in room acoustics of various performance spaces. Special and additional rehearsals are required, which often stress the budget. In many situations this means that an adaptation of the desired compositional effects, to the environment cannot be realized. To overcome this problem a practical solution should be required to provide composers the ability to examine the targeted acoustic sound scene.

1.2 Aim of the Thesis

The aim of this thesis is twofold. On the one hand spatial analysis and re-composition of directional room impulse responses is used to increase and improve the spatial representation of the sound field. On the other hand the same techniques can be used to transfer the natural room acoustics from a given event location to another room (**Figure 1.1**).

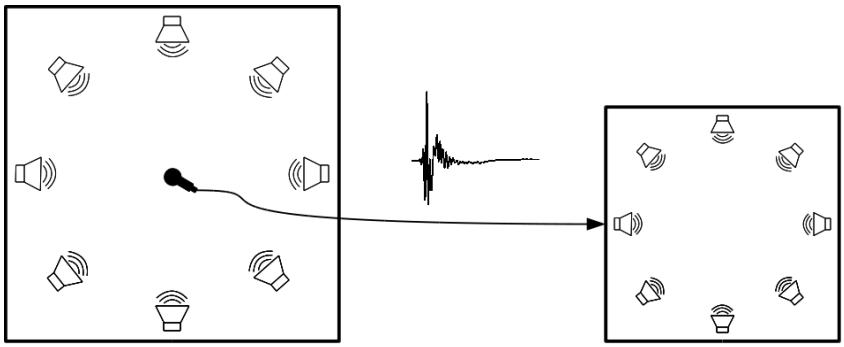


Figure 1.1: Illustration Room Acoustics Transfer

Both the targeted room, and the room from which the acoustic properties are to be captured must measured and analyzed. A loudspeaker array is required to render the results in the reproduction room. The excitation of the location where the acoustic properties are to be captured can be realized with a single loudspeaker, sampling the evaluated enclosing surface at several defined positions [2]. Various different approaches on how to determine and extract the position of a sound source in three dimensional space, based on multi-channel measurements, exist already [3] [4] [5] [6]. In this thesis, the analysis is based on coincident measurements using a B-Format recording from a conventional first-order Ambisonics microphone. Directional room impulse responses are measured from various loudspeaker positions, characteristic of the various reproduction systems. In order to analyze the room's impulse responses, it is essential to know where the direct sound and the discrete reflections occur, and where the diffuse sound prevails. In the case of music information retrieval, onset detection is a well known method to characterize features in audio signals (rhythm, pitch, segmentation, harmony, etc.) [7] [8] [9] and it is used in this thesis to detect and localize the discrete reflections signified by an increases of temporal focused energy. The diffuse part will be modelled by directional, frequency dependent and time-evolving acoustical energy measures. In the case

of (Higher Order) Ambisonics reproduction systems, the results are combined into a decoding procedure for reproduction and as a means to transfer these results to the acoustic space.

1.3 Outline of the Thesis

In chapter 2 the Ambisonics approach towards encoding and decoding of first-order Ambisonics as well as higher order Ambisonics (HOA) is discussed. The spherical coordinate system is established in this thesis.

Chapter 3 provides a brief theoretical insight into the structure and function of room impulse responses. Furthermore, MUMUTH and CUBE, the two analyzed rooms are presented with their geometric properties and their 3D Ambisonics sound rendering systems. The topic of coincident measurements and the procedure used for recording the directional room impulse responses is addressed with regard to the first-order Ambisonics approach.

Chapter 4 describes the analysis phase of the directional room impulse responses recorded in both rooms. This chapter gives a short introduction into Spatial Impulse Response Rendering (SIRR) and Directional Audio Coding (DirAC) approaches, which are the basic approaches used in this thesis. Various onset approaches are applied and compared, to find relevant sound events in the individual impulse responses and hence direct sound and reflections. The direction of arrival is analyzed at these special events and the diffuseness is determined in these regions. Finally the results from the direction and diffuseness analysis are evaluated and presented.

The different steps of the synthesis phase regarding the measured room impulse responses and according to the obtained analysis data are presented in chapter 5. The implementation of the non-diffuse parts to a higher spatial order reproduction and the synthesis of the diffuse parts via a decorrelation method are specified. Finally, an approach for the transfer of acoustic properties from one room to another is discussed.

Chapter 6 gives a summary of what has been achieved and an outlook on further improvements and processing.

The whole algorithm implementation is carried out in MATLAB[1].

[1] MATLAB is a trademark of The MathWork Inc.

2 Ambisonics

The Ambisonics approach describes a multichannel system for encoding and rendering a 3D sound field [10] [11]. The Ambisonics approach was developed primarily by Michael Gerzon [12] and is often mentioned alongside the decomposition of sound fields in spherical harmonics. It provides the ability to record and reproduce a periphonic sound field. It is assumed that both in the recorded and in the synthesized system only plane waves occur and this is generally the case when the sound sources are far enough away from the listening position. This assumption greatly simplifies the calculations in the summation of the recorded sound waves from different loudspeakers at a certain point in the listening area, ideally in the origin of the coordinate system. In case of 3D the sound waves are represented in a spherical coordinate system (see section 2.1). First designs and implementations of Ambisonics encoding and decoding used only first-order Ambisonics. The sound field is encoded into four channels and is known as the B-Format. Regardless of the loudspeaker setup the spatial room information of the recorded sound in the Ambisonics approach is encoded together with the sound itself. In this case the order of Ambisonics indicates the accuracy of encoding. The higher the order, the more precise the direction can be determined. Both first-order Ambisonics and B-Format respectively is described in section 2.2. Higher order Ambisonics is then discussed briefly in section 2.3.

2.1 Spherical Coordinate System

Throughout this thesis and for further calculation a spherical coordinate system as shown in Figure 2.1 is introduced for a sound field description. The origin of the coordinate system is the point from which the sound field is described and analyzed.

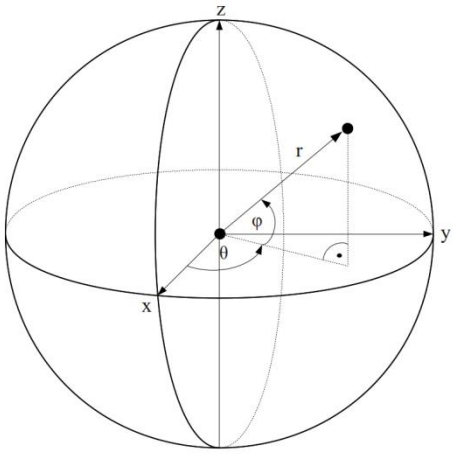


Figure 2.1: Spherical Coordinate System

Each point on a spherical surface is defined by the azimuth angle θ, the elevation angle φ and the radius r in the spherical coordinate system. The transformation from spherical coordinates to Cartesian coordinates can be derived from

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = r \begin{pmatrix} \cos\theta \cos\varphi \\ \sin\theta \cos\varphi \\ \sin\varphi \end{pmatrix} \tag{2.1}$$

where the azimuth angle θ is defined in the range $[-\pi, \pi]$ and the elevation angle φ in the range $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. The reverse transformation in the spherical coordinate system can be performed by the following equations

$$\begin{pmatrix} r \\ \theta \\ \varphi \end{pmatrix} = \begin{pmatrix} \sqrt{x^2 + y^2 + z^2} \\ \mathrm{atan2(y, x)} \\ \mathrm{atan2}\left(z, \sqrt{x^2 + y^2}\right) \end{pmatrix} \tag{2.2}$$

where the two-argument $\mathrm{atan2}$ functions are provided to eliminate quadrant confusion.

2.2 First-Order Ambisonics

Recording a sound field in three dimensions proves to be a demanding undertaking. Ideally the microphones have to be arranged coincidently for spatial recording, but it is not possible to place the capsules at exactly the same point. In Gerzon [13] [14] an approach to solve the geometric problem by developing a microphone with four cardioid microphone capsules in a tetrahedral arrangement is presented. This microphone is composed of four capsules, each with a cardioid directional characteristic, which lie very close together and are arranged on the vertices of a regular tetrahedron pointing outwards. A possible arrangement is shown in Figure 2.2.

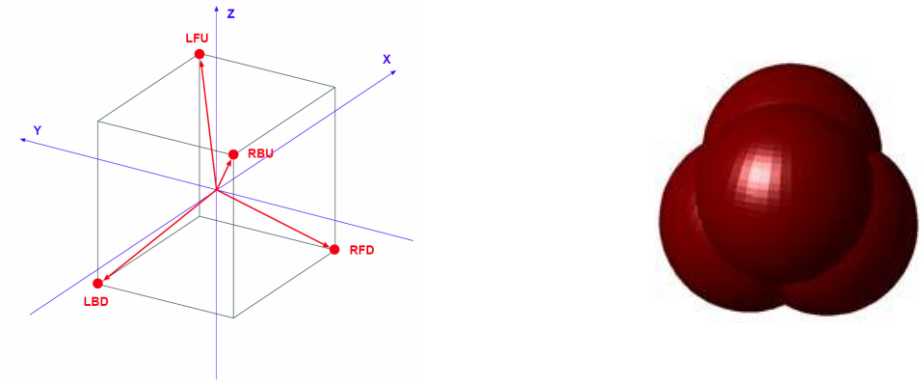


Figure 2.2: A-Format (left: Tetrahedral microphone design [15], right: Coincident cardioid representation in spherical harmonics [16])

In this arrangement, the microphone capsules are not precisely coincident, but they are equally non-coincident in each axis' direction. However, this will simplify the correction of the non-coincident response [16]. The recorded sound field is available in four channels, as required for first-order Ambisonics, the so called A-Format (LFU, LBD, RFD, RBU) [15]. Table 2.1 shows the orientation of the four microphone capsules.

Capsule	Azimuth	Elevation
LFU (left-front-up)	45°	35.3°
LBD (left-back-down)	135°	-35.3°
RFD (right-front-down)	-45°	-35.3°
RBU (right-back-up)	-135°	35.3°

Table 2.1: Capsule orientation in degrees for the A-Format first-order Ambisonics microphone [16]

Since each of the capsules has a cardioids pattern, all the recorded sound will be in phase. Simple calculations can be performed on these A-Format signals to transform them into the common B-Format signals (eq.(2.3)).

$$\begin{aligned} W' &= \mathrm{LFU} + \mathrm{LBD} + \mathrm{RFD} + \mathrm{RBU} \\ X' &= \mathrm{LFU} + \mathrm{RFD} - \mathrm{LBD} - \mathrm{RBU} \\ Y' &= \mathrm{LFU} + \mathrm{LBD} - \mathrm{RFD} - \mathrm{RBU} \\ Z' &= \mathrm{LFU} + \mathrm{RBU} - \mathrm{LBD} - \mathrm{RFD} \end{aligned} \tag{2.3}$$

B-Format

Both A-Format and B-Format Ambisonics are composed of four signals. Thus a first-order sound field representation is given, where a minimum of four channels is needed for a 3D reproduction. The first W-channel and thus the zero-order, refer to the omni-directional signal and represent the pressure of the recorded sound field. To complete the first-order Ambisonics representation, the three other signals have a figure-of-eight characteristic pointing in the direction of the x, y, and z axis proportional to the sound particle velocity components. They are called the X-, Y- and Z-channels and contain the direct information of the recorded sound field according to the three spatial axes. A graphical illustration of these four B-Format signal responses is given in Figure 2.3.

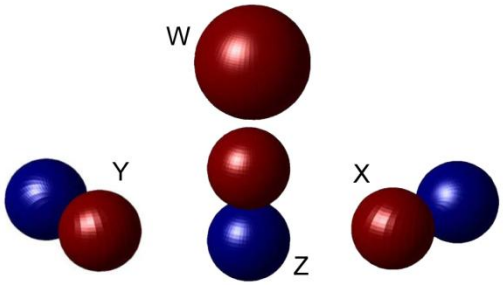


Figure 2.3: B-Format First-Order Ambisonics Patterns (cf. [16])

As mentioned above, ideally, the microphones should be placed at the same position. In the case of B-Format recordings this is hard to realize with omni-directional and figure-of-eight characteristic microphones. Therefore, the natural sound field is recorded in A-Format first, and then transferred into B-Format. The corresponding calculation is given in eq.(2.3). These signals are however not the required output signals and are therefore denoted here with an apostrophe. Further filtering is obligatory. The required complex mathematical workings for this filtering will not be explained in more detail, because the SoundField microphone used for this thesis (see section 3.3) provides a software implementation for the conversion of the recorded A-Format into the B-Format signals. More detailed explanations and remarks concerning the filters can be found in [11] [15] [17]. Eq. (2.4) shows the simple mathematical form for the B-Format Ambisonics channels.

$$W = \frac{1}{\sqrt{2}}$$

$$X = \cos\theta \cos\varphi \tag{2.4}$$

$$Y = \sin\theta \cos\varphi$$

$$Z = \sin\varphi$$

These equations are used to encode a sound source and represent the gains applied to the sound for the channels of the B-Format signal, where the azimuth θ as well as the elevation φ specify the direction. The weighting of the W-channel with the multiplier $\frac{1}{\sqrt{2}}$

is applied to obtain a more even signal level distribution within the Ambisonics channels [18].

2.3 Higher Order Ambisonics (HOA)

Higher Order Ambisonics (HOA) is considered an extension of the first-order Ambisonics approach presented in the previous sections. The notation first-order suggests that even higher order signals can be used in Ambisonics systems. Here the sound field is decomposed into a series of spherical harmonics functions. A higher order entails to a larger sweet area and also results in a more accurate sound field reproduction at higher frequencies. But at the same time more loudspeakers and channels for the transmission and storage are required.

The encoding and decoding process are independent of each other. Thus, portability of the encoded material is possible even without the knowledge of the loudspeaker layout in the decoding phase. With an array of an infinite amount of loudspeakers on a sphere, an accurate sound field reproduction can be achieved. This is however not realizable in real world applications. Using a finite number of loudspeakers, arranged on a spherical surface, a good approximation of the original sound field can be obtained for the sweet spot. The higher the spherical harmonics order, the more stable and more truthfully the sound field can be mapped spatially [19]. The encoding and decoding process are briefly described in the following sections 2.3.1 and 2.3.2.

2.3.1 Encoding

The goal of Ambisonics is to synthesize a copy of a recorded plane wave in a different place [20]. This plane wave is then converted back into a real sound field by reproduction depending on the loudspeaker positions. It is necessary, during analysis, to find a closed description of such a plane wave, being emitted by a mono source. The encoding phase describes the procedure by which a mono source, including direction information, is described as a plane wave.

By writing the 3D wave equation in the spherical coordinate system (see Figure 2.1), and based on the spherical harmonics decomposition of the sound field, the Ambisonics representation is obtained. The mathematical formulation of the wave equation is shown in eq. (2.5),

$$(\Delta + k^2)p(\vec{r}, \omega) = 0 \tag{2.5}$$

where Δ is the Laplace Operator in spherical coordinates, k the wave number with $k = 2\pi f/c = \omega/c$ and p the sound pressure. The speed of sound is $c \approx 340\frac{m}{s}$. Therefore the sound field can be written as the spherical Fourier-Bessel series [11],

$$p(\vec{r}) = \sum_{n=0}^{\infty} j^n j_n(kr) \sum_{0 \le m \le n, \sigma = \pm 1} B_{nm}^{\sigma} Y_{nm}^{\sigma}(\theta, \varphi) \tag{2.6}$$

where $j_n(kr)$ are the n[th] order Bessel functions and Y_{nm}^{σ} are the spherical harmonics functions, which are defined as [11] [21]

$$Y_{nm}^{\sigma}(\theta, \varphi) = N_{nm} \cdot P_{nm}(\sin\varphi) \cdot \begin{cases} \cos m\theta & for\ \sigma = 1 \\ \sin m\theta & for\ \sigma = -1 \end{cases} \tag{2.7}$$

P_{nm} describes the associated Legendre functions of order *n* and mode *m* [11]. Table 2.2 lists a small selection of the associated Legendre functions up to order $n = 3$ and mode $m = 0$. For each order n there are $2n + 1$ modes.

The spherical harmonics Y_{nm}^{σ} form a set of orthogonal basis vectors. By the appropriate choice of N_{nm} the orthogonal basis is normalized. To receive the ortho-normalized Laplace spherical harmonics N_{nm} is defined as

$$N_{nm} = (-1)^m \sqrt{\frac{2n+1}{4\pi} \varepsilon_m \frac{(n-m)!}{(n+m)!}} \tag{2.8}$$

with $\varepsilon_0 = 1$ and $\varepsilon_m = 2\ for\ m \geq \pm 1$ and where ! is the factorial.

n	$P_{nm}(\sin\varphi) \mid m = 0$
0	1
1	$\sin\varphi$
2	$\frac{1}{2}(3\sin^2\varphi - 1)$
3	$\frac{1}{2}(5\sin^3\varphi - 3\sin\varphi)$

Table 2.2: Associated Legendre Polynomials with Order n and Mode $m = 0$ (cf. [11] [21])

The ortho-normalized spherical harmonics Y_{nm}^{σ} up to third order are shown in Figure 2.4 and listed in Table 2.3. For each order n and mode m with $0 \leq m \leq n$, various spherical harmonics functions exist for both indexes $\sigma = \pm 1$ [22].

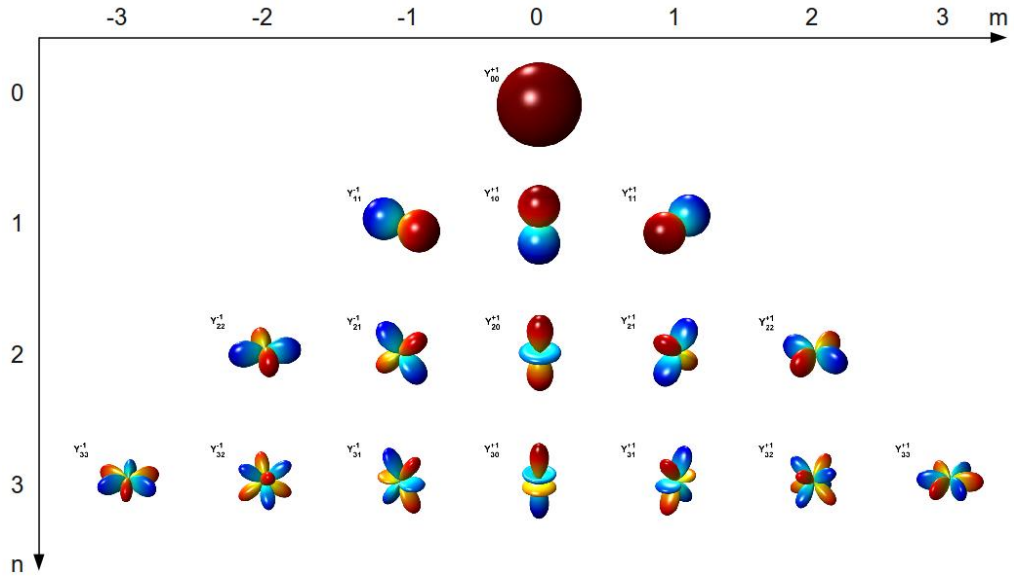


Figure 2.4: Spherical Harmonics up to Third Order (cf. [11])

Order	σ nm	$Y_{nm}^{\sigma(SN3D)^2}(\theta,\varphi)$
0	1 00	$\sqrt{\frac{1}{4\pi}}$
1	1 11	$\sqrt{\frac{3}{4\pi}}\cdot(\cos\theta\cos\varphi)$
	−1 11	$-\sqrt{\frac{3}{4\pi}}\cdot(\sin\theta\cos\varphi)$
	1 10	$\sqrt{\frac{3}{4\pi}}\cdot(\sin\varphi)$
2	1 22	$\sqrt{\frac{5}{4\pi}}\cdot\frac{\sqrt{3}}{2}(\cos 2\theta\cos^2\varphi)$
	−1 22	$-\sqrt{\frac{5}{4\pi}}\cdot\frac{\sqrt{3}}{2}(\sin 2\theta\cos^2\varphi)$
	1 21	$\sqrt{\frac{5}{4\pi}}\cdot\frac{\sqrt{3}}{2}(\cos\theta\sin 2\varphi)$
	−1 21	$-\sqrt{\frac{5}{4\pi}}\cdot\frac{\sqrt{3}}{2}(\sin\theta\sin 2\varphi)$
	1 20	$\sqrt{\frac{5}{4\pi}}\cdot\frac{1}{2}(3\sin^2\varphi-1)$
3	1 33	$\sqrt{\frac{7}{4\pi}}\cdot\frac{\sqrt{10}}{4}(\cos 3\theta\cos^3\varphi)$
	−1 33	$-\sqrt{\frac{7}{4\pi}}\cdot\frac{\sqrt{10}}{4}(\sin 3\theta\cos^3\varphi)$
	1 32	$\sqrt{\frac{7}{4\pi}}\cdot\frac{\sqrt{15}}{2}(\cos 2\theta\sin\varphi\cos^2\varphi)$
	−1 32	$-\sqrt{\frac{7}{4\pi}}\cdot\frac{\sqrt{15}}{2}(\sin 2\theta\sin\varphi\cos^2\varphi)$
	1 31	$\sqrt{\frac{7}{4\pi}}\cdot\frac{\sqrt{6}}{4}(\cos\theta\cos\varphi(5\sin^2\varphi-1))$
	−1 31	$-\sqrt{\frac{7}{4\pi}}\cdot\frac{\sqrt{6}}{4}(\sin\theta\cos\varphi(5\sin^2\varphi-1))$
	1 30	$\sqrt{\frac{7}{4\pi}}\cdot\frac{1}{2}(5\sin^3\varphi-3\sin\varphi)$

Table 2.3: Ortho-Normalized Laplace -Spherical Harmonics up to Third Order (cf. [21])

[2] The exponent tag (SN3D) attached to the spherical harmonic functions Y_{nm}^{σ} means that these are the semi-normalized encoding functions [11].

Looking at the directional information of a plane wavefront at the origin, the desired Ambisonics channels B_{nm}^{σ} (eq. (2.9)) can be represented simply by multiplying the received pressure signal s with real encoding gains of the spherical harmonics Y_{nm}^{σ} in the direction θ_s, φ_s from which the source radiates the plane wave [11]. This equation describes the Ambisonics encoding process of a spatialized signal.

$$B_{nm}^{\sigma} = s \cdot Y_{nm}^{\sigma}(\theta_s, \varphi_s) \tag{2.9}$$

Due to the complete description of the plane wave, it is possible to re-synthesize the original sound field with a reproducing system. The series expansion is truncated after a certain term because the transmission and reproduction of an infinite number of channels is not possible. The *L* transmission channels are determined by the Ambisonics order *N* according to the following equation in the 3D:

$$L^{3D} = (N+1)^2 \tag{2.10}$$

Due to the truncation of the series expansion, a loss of information occurs and the plane wave is not completely representable. In order to avoid reconstruction errors, the reproduction system should exhibit $L \geq (N+1)^2$ loudspeakers.

Combining all the achieved Ambisonics channels, the reduced Ambisonics sound field can be rewrite from eq. (2.9) in a vector representation

$$\mathbf{B} = s_i \cdot \mathbf{Y}(\theta_i, \varphi_i) \tag{2.11}$$

The individual channels of the Ambisonics sound field are denoted by the letters of the alphabet in the following form:

$$\mathbf{B} = \begin{pmatrix} B_{00} \\ B_{11}^{1} \\ B_{11}^{-1} \\ B_{10}^{1} \\ \vdots \end{pmatrix} = \begin{pmatrix} W \\ X \\ Y \\ Z \\ \vdots \end{pmatrix} \tag{2.12}$$

By using superposition, an Ambisonics sound field of a specific order can contain any number of virtual sound sources with consistent properties. By adding several Ambison-

ics sound fields, each with different sound sources at arbitrary positions, the resulting sound field is a representation of several virtual sources.

$$\mathbf{B}' = \sum_{i=1}^{k} s_i \cdot \mathbf{Y}(\theta_i, \varphi_i) \tag{2.13}$$

for *k* sound sources.

2.3.2 Decoding

In a further step, the described plane wave is converted back into a real sound field with a loudspeaker array. The encoded sound field can be reproduced in the sweet spot of the reproduction system by superposing a set of loudspeaker signals. This process is called decoding.

The aim of an Ambisonics reproduction system is to re-synthesize the original sound field *S* as accurately as possible. The following applies:

$$S_{Analyse} \equiv S_{Synthese} \tag{2.14}$$

Accordingly, the relationship between the encoded sound field of a single source and the re-synthesized sound field can be written as

$$s \cdot Y_{nm}^{\sigma}(\theta_s, \varphi_s) = \sum_{l=1}^{L} p_l \cdot Y_{nm}^{\sigma}(\theta_l, \varphi_l) \tag{2.15}$$

where p_l is the signal of the lth loudspeaker at direction θ_l, φ_l. This can be rewritten into the compact matrix form as

$$\mathbf{B} = \mathbf{C} \cdot \mathbf{p} \tag{2.16}$$

where $\mathbf{p}$ represents the loudspeaker signals with $\mathbf{p} = [p_1, p_2, p_3, \ldots, p_L]^{\mathrm{T}}$ and the decoding matrix $\mathbf{C}$, containing the decoded loudspeaker directions θ_l, φ_l, with

$$\mathbf{C} = \begin{bmatrix} Y_{00}^{1}(\theta_1, \varphi_1) & Y_{00}^{1}(\theta_2, \varphi_2) & \dots & Y_{00}^{1}(\theta_L, \varphi_L) \\ Y_{10}^{1}(\theta_1, \varphi_1) & Y_{10}^{1}(\theta_2, \varphi_2) & \dots & Y_{10}^{1}(\theta_L, \varphi_L) \\ Y_{11}^{-1}(\theta_1, \varphi_1) & Y_{11}^{-1}(\theta_3, \varphi_3) & \dots & Y_{11}^{-1}(\theta_L, \varphi_L) \\ \vdots & \vdots & \ddots & \vdots \\ Y_{MM}^{-1}(\theta_1, \varphi_1) & Y_{MM}^{-1}(\theta_2, \varphi_2) & \dots & Y_{MM}^{-1}(\theta_L, \varphi_L) \end{bmatrix} \tag{2.17}$$

By reshaping eq. (2.16), the signals to drive loudspeakers can be obtained by inversion of the matrix $\mathbf{C}$.

$$\mathbf{p} = \mathbf{C^{-1}} \cdot \mathbf{B} = \mathbf{D} \cdot \mathbf{B} \tag{2.18}$$

for $N = L$, where the inverted matrix $\mathbf{C^{-1}}$ is referred to as the decoding matrix $\mathbf{D}$. However generally $N \neq L$, and thus the loudspeaker signals $\mathbf{p}$ are achieved by using the pseudo-inverse of matrix $\mathbf{C}$ as mentioned in eq. (2.19)

$$\mathbf{D} = pinv(\mathbf{C}) = \mathbf{C^{T}} \cdot (\mathbf{C} \cdot \mathbf{C^{T}})^{\mathbf{-1}} \tag{2.19}$$

To ensure that the contained directional information contained in the Ambisonics channels remain after decoding, the number of loudspeakers L must be greater or equal to the number of Ambisonics channels M.

$$L \geq M \tag{2.20}$$

Optimal decoding occurs when M Ambisonics signals are decoded to the $M = L$ loudspeakers (cf. [11] [19]), thus the sound field representation error is minimized.

In general, the higher the Ambisonics order is chosen, the more accurate the reproduction and the smaller the aberrations of sound sources. In [2] and in Figure 2.5, Zotter shows the relationship between the Ambisonics order and the corresponding spherical beam-widths. In the right illustration, the characteristics of the spherical main and side lobes are shown according to the Ambisonics orders. It is clearly visible that the width of the main beam with increasing order is significantly narrower as well and thus increases the sharpness of localization. In the right illustration, the corresponding main lobe widths of certain weighting factors of the source are shown.

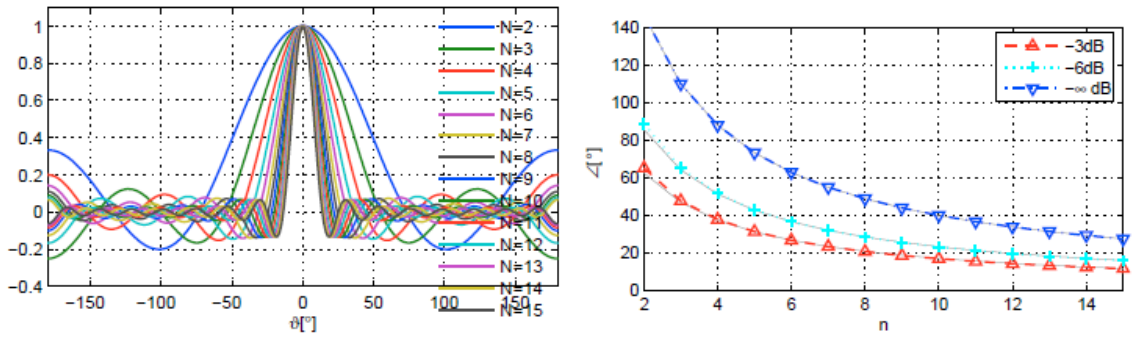


Figure 2.5: Band-Limited Spherical Beam-Widths (left: characteristics of band-limited spherical beams, right: corresponding main lobe width) [2]

The resolution or order of the system plays a major role in terms of localization and is likewise shown in [23]. Here, using the case of a 2D Ambisonics 3rd order system, it is shown that by weighting the orders, a stability of the localization occurs and thus the consistency of the reproduction should be increased. However, weighting the orders leads to expansion of the source which in turn leads to more localization blur. An increase in the Ambisonics order would counteract this effect. [23]

Ambisonics is a coincident recording process where, due to the physical size of microphones, only limited complex microphone characteristics exist. Thus, with this method, the recording of real sound fields with orders greater than 1 is only possible with method with limitations (see section 3.3). However, as deduced from synthetically generated sound fields, the limiting factors are computing power, the number of transmission channels and the physically loudspeakers availeable. The required number of transmission channels is derived from the realized system order. [24]

3 Measuring Directional Room Impulse Responses

In addition to measurement and analyzes of directional room impulse responses in a room, one aim in this thesis is the reproduction and transfer of natural room acoustics from one room to another. Therefore, an acoustic measurement procedure for the transfer of the acoustic properties is applied to both to the source and to the target room. As a concrete example the two rooms selected for the procedure are the concert hall in the MUMUTH (House of Music and Music Theatre) and the IEM CUBE (Institute of Electronic Music and Acoustics) at the University of Music and Performing Arts in Graz, Austria. In both rooms an Ambisonics 3D sound rendering system for the upper hemisphere is installed. An acoustic measurement procedure is applied to retrieve directional room impulse responses, which are measured from every single loudspeaker both in the source and the target room [25].

3.1 Room Impulse Responses

This section gives a brief introduction to room impulse responses. Room impulse responses are closely related to spatial impressions and the perception of sound in a room or hall. In such environments the propagation of sound has a significant influence on the auditory impression. The acoustic properties can be described by many parameters. In [26] [27] [28] [29] several important aspects concerning the spatial impression have been reported and various objective descriptive parameters have been developed. Information from the cited references to describe the characteristics of room impulse responses are given in the following paragraphs.

As soon as a sound source emits a signal in an enclosed room, a sound field is created out of the sound waves reflected from the walls, the floor, the ceiling and the objects in the room. These sound waves come back from different directions, reaching certain positions in the room, such as a listener or a microphone at different times. A schematic il-

lustration of an impulse response with the different regions of direct sound, early reflections and reverberation is shown in **Figure 3.1**.

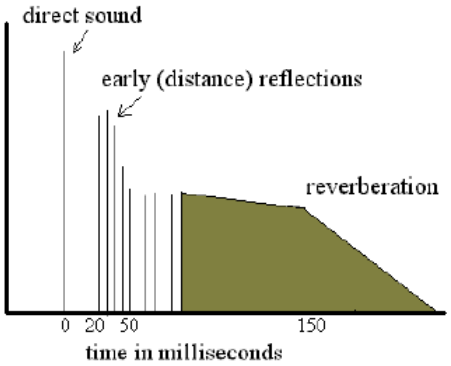


Figure 3.1: Impulse Response Illustration. Direct sound, early reflections and reverberation are shown. [30]

Firstly the direct sound from the sound source occurs, followed by the first reflections. After a certain time, these reflections become denser and build up, creating a reverberation. The acoustic properties of a room are defined by these components, and thus influence the way a sound is perceived. In contrast to the ambience, the direct sound is the first sound wave which arrives at the listening or measuring position without having been reflected. The arriving of the first wave front as a direct sound is important for the determination of direction, while the ambience is essential for the acoustic quality of a room.

The early reflections have different effects on the auditory impression depending on the sound event. The auditory impression depends on the delay time, the intensity, and the directions of early reflections. The time frame in which reflections are considered to be early reflections varies between 50 ms and 80 ms after the direct signal, depending on the referenced literature, and is influenced by the size of the room. The reflections are important for the clarity of the sound impression as well as the perception of spaciousness. In particular, the first reflection plays an important role. The later it arrives after the direct sound the bigger the room is perceived to be.

The late reflections occur about 150 ms after the direct sound and form the reverberation. They are responsible for the enveloping effect. The reverberation is associated with the diffuseness in a sound field. A diffuse field consists of an infinite amount of uncor-

related plane waves from various directions. This means that no energy flows in any direction and the active intensity is zero.

3.2 Exponential Sine Sweep (ESS) Method

Various methods for measuring room impulse response exist and all make use of different excitation signals. It is important that the excitation provide sufficient energy at all desired frequencies, that it be deterministic and reproducible. Established excitation signals are the Maximum Length Sequence (MLS) with a pseudorandom binary sequence similar to white noise, the Time Delay Spectrometry (TDS) with a linear sweep and the Exponential Sine Sweep (ESS) with an exponential sine sweep. There are plenty of studies which deal with the advantages and disadvantages of these different methods [31] [32]. For the measurements of the room impulse responses in this thesis the ESS method according to Farina [33] [34] was selected.

The relation between the excitation signal and hence the recorded signal is given by linear convolution

$$x(t) = s(t) * h(t) \tag{3.1}$$

where $s(t)$ is the excitation signal, $h(t)$ the room impulse response and $*$ the convolution operator.

The ESS excitation signal is calculated as

$$s(t) = \sin\left[\frac{\omega_1 \cdot T}{\ln\left(\frac{\omega_2}{\omega_1}\right)} \cdot \left(e^{\frac{t}{T}\cdot \ln\left(\frac{\omega_2}{\omega_1}\right)} - 1\right)\right] \tag{3.2}$$

with $\omega = 2\pi f$ and with the start frequency f_1 Hz at a time $t = 0$ and the stop frequency f_2 Hz at a time $t = T$.

To obtain the impulse response, deconvolution must be carried out, which is applied by means of the Fast Fourier Transformation (FFT) as follows

$$h(t) = ifft\left(\frac{fft(x(t))}{fft(s(t))}\right) \tag{3.3}$$

Note $s(t)$ is of length N_s and $h(t)$ of length N_h, thus the resulting signal $x(t)$ from eq. (3.1) is of length $N_x = N_s + N_h - 1$ after convolution. To allow the deconvolution calculation in eq. (3.3), $s(t)$ must be zero-padded to N_x samples.

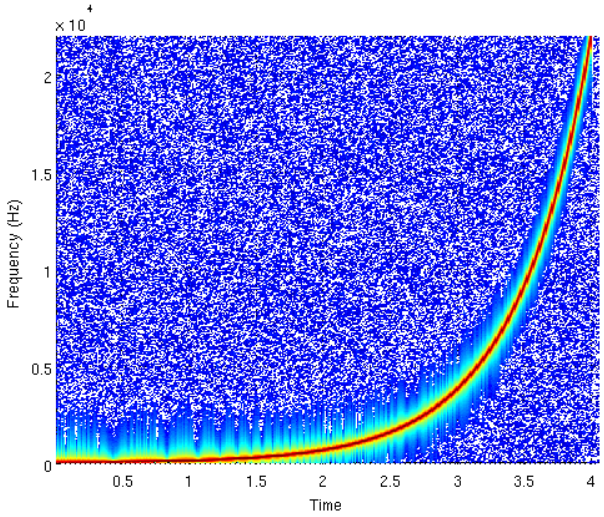


Figure 3.2: Exponential Sine Sweep

The used exponential sine sweep is shown in Figure 3.2 and consists of a length $T = 4\ s$ and a frequency range from $f_1 = 0\ Hz$ to $f_2 = 22\ kHz$.

3.3 Coincident Measurement

To characterize the acoustics of natural rooms, basic standard measurement procedures such as ISO STANDARD 3382 and ISO STANDARD 18233 have been formulated. To measure and analyze room impulse responses with only a single undirected microphone is however insufficient with regard to the direct resolution of incoming sound events. Coincident microphone setups represent a good approach to capture sound from a single source in the same phase by all microphones. A coincidence measurement is obtained with a coincident microphone arrangement in which two or more capsules are placed as close to each other as possible. This has the advantage that there is no time differences

between the measurements of the different capsules. The principle here is to use level differences caused by directive microphones oriented in different directions for the acoustic source localization. As previously mentioned in section 2.2, Ambisonics B-Format is one such option for coincident measurements. However, recording "only" in first-order format, does not provide sufficient directional resolution results, as the directivity patterns are still too broad to record basically non-diffuse sound. The ideal case would be a coincident setup where the microphone orientation corresponds to the loudspeaker configuration. One microphone for each loudspeaker, whereby only the direct and non-diffuse sound would be picked up and reproduced through an adequate loudspeaker setup close to the correct directions. But the higher the number of required microphones, the higher the cost of such a microphone system and thus such an arrangement can hardly be achieved. [35]

Figure 3.3: SoundField SPS200 Microphone[3]

Nevertheless, multi-channel coincident measurements are applied, to determine the location of sound source in the 3D [25]. These provide a basis for the directional room response measurements in this thesis and for the aim of the re-synthesis to higher order reproduction (see chapter 5).

A SoundField SPS200 microphone with its multicapsule microphone unit shown in Figure 3.3 is used as a measuring instrument to capture the spatial impulse responses in this case. This microphone is composed of four capsules, each with cardioid directional characteristic, which lie very close together and are arranged on the vertices of a regular

[3] www.soundfield.com

tetrahedron pointing outwards (cf. Figure 2.2, left). The SoundField SPS200 microphone is able to deliver a recording of the spatial impulse response in A-Format which can be converted, by matrixing, into the first-order Ambisonics B-Format [17] with four channels (see section 2.2). The W-channel represents an omni-directional channel and the X-, Y- and Z-channels have the directional pattern of a dipole, organized orthogonally to deliver the sound information corresponding to axes in the Cartesian coordinate system (see section 2.1). However, several other arrangements providing the possibility to determine the spherical harmonics representation at the origin are adequate [36] [37].

3.4 MUMUTH

The first room to be analyzed, the source room, where the directional room impulse responses are captured is the concert hall in the MUMUTH. An Ambisonics 3D sound rendering system equipped with 29 loudspeakers is installed in the concert hall with the room dimensions from approximately 32x17x11 m^3. The appropriate layout is visualized in a simple shoe box model in Figure 3.4.

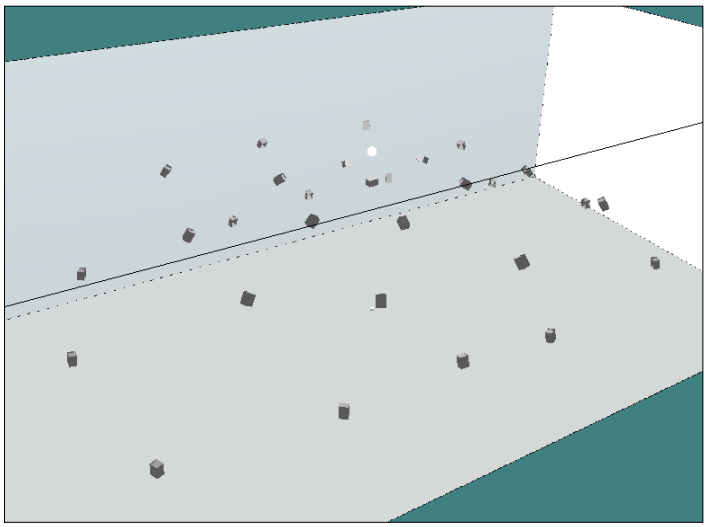


Figure 3.4: Loudspeaker Layout in the MUMUTH. A simple shoe box model of the concert hall showing the 29 installed loudspeakers from the Ambisonics reproduction system.

The coordinates of the 29 loudspeakers at the moment the impulse response measurements are carried out, can be taken from Table 3.1. The positions are denoted by Cartesian coordinates and spherical coordinates. During the measurements, the SoundField

microphone was located at a height of 1,2 meters at the center of the hall and the loudspeakers were directed towards to the sweet spot. Thus the position of the microphone with the center of the loudspeaker arrangement coincide.

Rig	Loud-speaker Number	*Cartesian Coordinates*				*Spherical Coordinates*	
		x[m]	**y[m]**	**z[m]**	**z*[m]**	**Elevation φ [°]**	**Azimuth θ [°]**
upper center	1	0,000	0,000	7,000	5,800	89,3	-
upper	2	-0,358	-2,674	6,594	5,394	63,4	-99,0
	3	2,910	-1,875	7,000	5,800	60,5	-33,9
	4	3,067	1,258	6,747	5,547	59,1	22,8
	5	0,258	2,674	6,594	5,394	63,5	85,5
	6	-2,910	1,875	7,000	5,800	60,5	147,9
	7	-3,067	-1,258	6,747	5,547	59,2	-157,9
middle	8	-3,233	-6,239	6,500	5,300	35,2	-117,7
	9	1,753	-6,588	6,500	5,300	36,0	-75,1
	10	6,298	-5,258	6,758	5,558	33,5	-38,2
	11	6,825	-0,832	6,198	4,998	36,0	-7,1
	12	6,298	3,363	5,885	4,685	32,5	26,5
	13	3,233	6,239	6,500	5,300	35,2	62,9
	14	-1,753	6,588	6,500	5,300	36,0	105,3
	15	-6,548	5,258	6,758	5,558	33,5	141,6
	16	-6,825	0,832	6,198	4,998	36,0	173,1
	17	-6,548	-3,363	5,885	4,685	32,5	-152,7
lower	18	-4,738	-6,675	3,265	2,065	8,0	-125,4
	19	-0,389	-6,675	3,418	2,218	8,0	-93,7
	20	3,381	-6,675	3,015	1,815	8,0	-63,5
	21	10,125	-5,258	3,165	1,965	8,0	-27,7
	22	11,400	0,000	3,181	1,981	8,0	-0,1
	23	10,125	5,258	2,362	1,162	8,0	27,4
	24	4,738	6,675	3,265	2,065	8,0	54,5
	25	0,389	6,675	3,418	2,218	8,0	87,0
	26	-3,381	6,675	3,015	1,815	8,0	117,5
	27	-10,125	5,258	3,165	1,965	8,0	152,7
	28	-11,400	0,000	3,181	1,981	8,0	-180,0
	29	-10,125	-5,258	2,362	1,162	8,0	-152,6

Table 3.1: Loudspeaker Positions in Cartesian and Spherical Coordinates, MUMUTH

Table 3.1 ranks the loudspeakers according to their arrangement. The setup is divided into three rigs at different heights plus one loudspeaker in the center, located exactly above the microphone. This loudspeaker is denoted as number one, followed by the upper rig with six loudspeakers, then the middle rig with ten loudspeakers and finally the lower rig with twelve loudspeakers. The positions are given in Cartesian and spherical coordinates. Note, the z-coordinates without the * are the distance from the ground and the z-coordinates marked with the * are the distance from the microphone position, 1,2 meters above the ground.

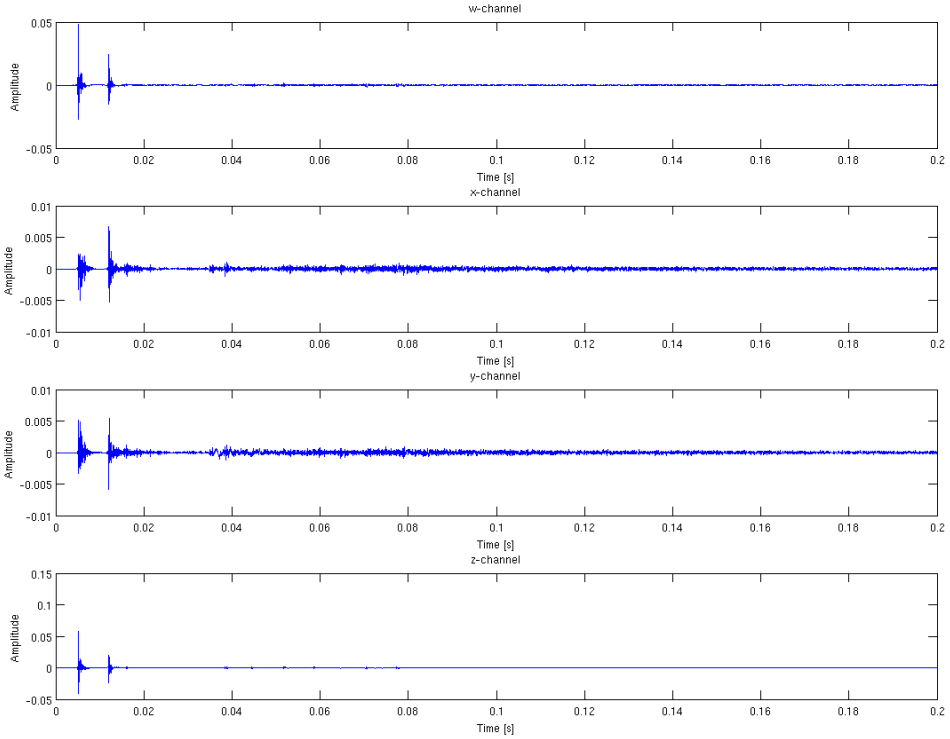


Figure 3.5: B-Format Room Impulse Responses after Deconvolution, Loudspeaker Number 1, MUMUTH. (first: W-channel, second: X-channel, third: Y-channel, fourth: Z-channel)

To obtain the acoustic properties of the room, the spatial impulse responses from each loudspeaker was measured. The exponential sine sweep method mentioned in section 3.2 was chosen and the excitation signal used is shown in Figure 3.2. The resulting signal, which was recorded at the measuring point, represents the sweep that was played, convolved with the impulse response of the hall. To determine the impulse responses, the measured sound pressure distributions were deconvolved with the known excitation signal in frequency domain. Figure 3.5 shows an example of the recorded room impulse response in Ambisonics B-Format from the loudspeaker number 1 in the MUMUTH. In the first illustration, the W-signal followed by the X-, Y- and Z-signals for approximately the first 200 ms is plotted. Taking the different scaling of the amplitudes into account, one can see that the most of the energy is found in the first sound event and thus the direct sound at the Z-channel. This is obviously due to the fact that the loudspeaker position is exactly above the microphone.

3.5 CUBE

The second room, the reproduction room where the acoustic properties from the MU-MUTH are to be transferred to, must be measured and analyzed. The CUBE is a medium-sized concert hall used mainly for the reproduction of electro-acoustic music [38]. The room is approximately 10x11x5 m^3 and is equipped with an Ambisonics 3D sound rendering system consisting of 24 loudspeaker. The appropriate layout is visualized in a simple shoe box model in Figure 3.6.

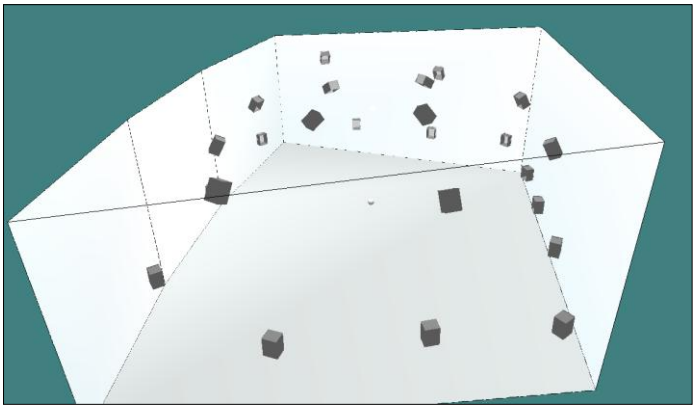


Figure 3.6: Loudspeaker Layout in the CUBE. A simple shoe box model of the medium-sized concert hall showing the 24 installed loudspeakers from the Ambisonics reproduction system.

Again the single positions of the 24 loudspeakers are listed in Table 3.2 in Cartesian and spherical coordinates. The 24 loudspeakers are arranged in three rigs where the lower rig consists of twelve loudspeakers, the middle rig of eight and the upper rig consists of four loudspeakers. The z-coordinates in the table denoted with the * are the distance from the microphone position 1,2 meters above the ground during the measurements.

Rig	Loudspeaker Number	*Cartesian Coordinates* x[m]	y[m]	z[m]	z*[m]	*Spherical Coordinates* Elevation φ [°]	Azimuth θ [°]
lower	1	4,635	0,000	1,341	0,141	1,7	0,0
	2	4,600	2,023	1,381	0,181	2,1	-23,7
	3	4,113	4,596	1,401	0,201	1,9	-48,2
	4	1,574	5,028	1,405	0,205	2,2	-72,6
	5	-1,289	5,553	1,406	0,206	2,1	-103,1
	6	-4,376	3,873	1,358	0,158	1,5	-138,5
	7	-4,636	0,016	1,371	0,171	2,1	-179,8
	8	-4,331	-3,860	1,353	0,153	1,5	138,3
	9	-1,068	-5,533	1,400	0,200	2,0	100,9
	10	1,821	-4,943	1,376	0,176	1,9	69,8
	11	4,481	-4,456	1,387	0,187	1,7	44,8
	12	4,711	-1,850	1,385	0,185	2,1	21,4
middle	13	4,230	1,766	3,828	2,628	29,8	-22,7
	14	1,806	4,441	3,938	2,738	29,7	-67,9
	15	-2,189	4,873	4,173	2,973	29,1	-114,2
	16	-3,624	1,476	3,478	2,278	30,2	-157,8
	17	-3,602	-1,577	3,493	2,293	30,2	156,4
	18	-2,055	-4,782	4,160	2,960	29,6	113,3
	19	1,925	-4,210	3,854	2,654	29,8	65,4
	20	4,223	-1,767	3,771	2,571	29,3	22,7
upper	21	1,368	1,456	4,423	3,223	58,2	-46,8
	22	-1,252	1,324	4,153	2,953	58,3	-133,4
	23	-1,267	-1,342	4,142	2,942	57,9	133,4
	24	1,399	-1,325	4,392	3,192	58,9	43,4

Table 3.2: Loudspeaker Positions in Cartesian and Spherical Coordinates, CUBE

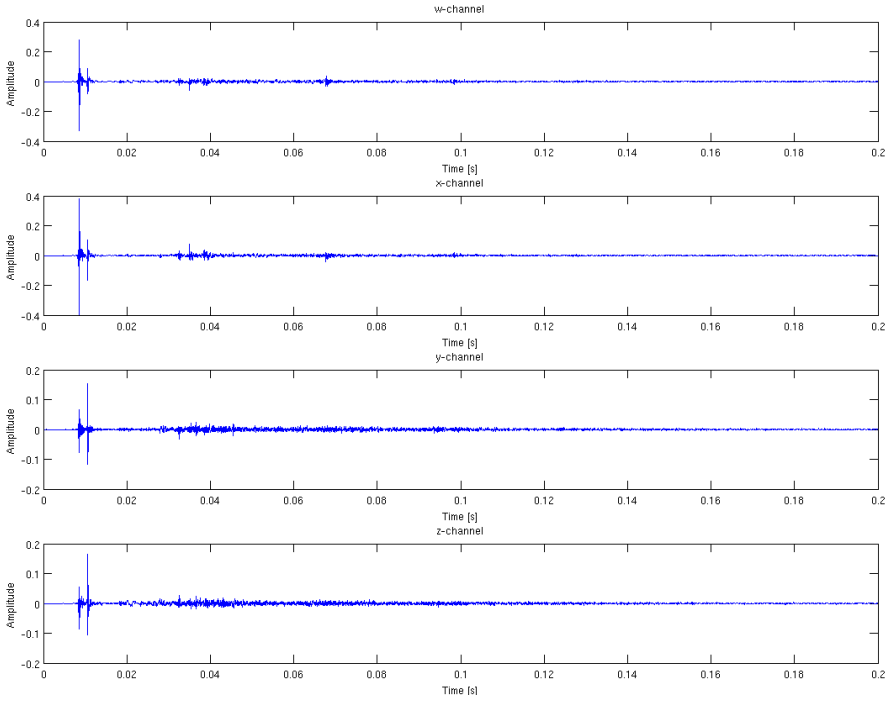


Figure 3.7: B-Format Room Impulse Responses after Deconvolution, Loudspeaker Number 1, CUBE. (first: W-channel, second: X-channel, third: Y-channel, fourth: Z-channel)

The same impulse response measurement technique as was used in the MUMUTH was applied. Figure 3.7 shows an example of the measured B-Format room impulse response from the loudspeaker number 1 in the CUBE. This loudspeaker is positioned per definition in alignment with the x-axis. Once again when the different scaling of the amplitudes is taken into account, one can see that the X-channel exhibits the most energy of the direct sound measured. For further analysis of the directions of incoming sounds see section 4.3.

4 Analysis Phase

4.1 Directional Audio Coding / Spatial Impulse Response Rendering

Spatial Impulse Response Rendering (SIRR) [39] [40] and Directional Audio Coding (DirAC) [3] [37] [41] gives an insight into the recent techniques for the reproduction and transfer of natural room acoustics with a multichannel loudspeaker system. The spatial perception of human hearing is determined by the direction of arrival, the diffuseness and the spectral and temporal properties of the sound field [42]. The Dirac approach takes these aspects into account and is based on an energy analysis of the sound field and is a well-proved technique in spatial sound reproduction [5] [36].

The DirAC process is divided into two parts. First, in an analysis phase, the direction of arrival and the diffuseness of a sound field are estimated in auditory frequency bands as a function of time, measured at a single point. The resulting metadata of the analyzed room is transmitted as an additional information together with one or more audio channels for a reproduction and re-synthesis of the sound field. This data can then used in another room with a multichannel loudspeaker system for example. Second, in the following synthesis phase the recorded sound is split up in two parts, the non-diffuse stream and the diffuse stream. The non-diffuse stream contains directional information of the incident sound and is used to create point-like virtual sources. The diffuse part consists of the diffuse sound or possible noise in the measured place and is synthesized by decorrelation.

The SIRR technique discussed in [39] [40] follows mainly the same methodology and is based on DirAC. Whereas the DirAC approach is particularly construed to analyze and re-compose high quality applications such as music reproduction, the SIRR approach is especially scaled for the analysis and reproduction of spatial impulse responses characteristic to short transient sound events. The signals are also subjected to a time-frequency analysis in order to retrieve information of the direction of arrival and the dif-

fuseness of sound. This is followed the extraction of spatial multichannel impulse responses, which can be tailored to any loudspeaker systems by means of synthesis. SIRR is primarily a technique to reproduce such room impulse responses for convolving reverberators. However, there is a great similarity between the two methods mentioned above.

Based on the SIRR method the following sections of chapter 4 describe the analysis phase and chapter 5 the synthesis phase from the measured directional spatial room responses in the MUMUTH and the CUBE (see section 3.4 and section 3.5).

The flow chart in Figure 4.1 displays a general overview of the performed steps in this analysis phase.

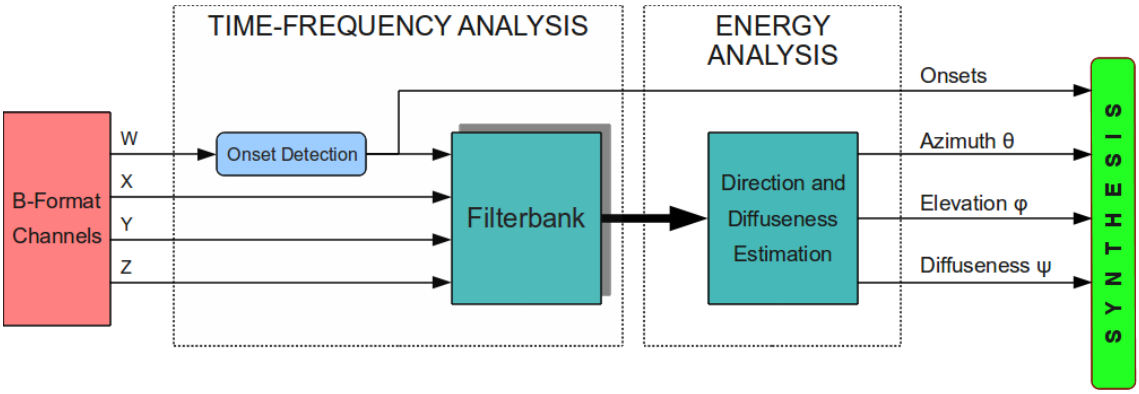


Figure 4.1: Flow Chart of the Analysis Phase. A time-frequency analysis of the directional impulse responses recorded in B-Format is performed, followed by an energy analysis to obtain the required metadata for the synthesis phase.

First of all, the measured directional room impulse responses, presented in Ambisonics B-Format, go through time-frequency processing. The sound pressure and particle velocity measured at one single point of the sound field is needed as a function of time and frequency. Transients and thus the arrival of direct sound and its reflections are located in the impulse responses using an onset detection procedure. Several approaches were reviewed with regard to the efficiency and usefulness in this specific case and the most appropriate method is selected. The implementation is described in section 4.2 and its corresponding subsections. Based on the equivalent rectangular bandwidth (ERB) scale, all microphone signals are divided into frequency bands for analysis, which is explained in section 4.3.2. Following this, the conditions for obtaining the direction of arrival of

the incident sound events and the diffuseness of the sound field via an energy analysis can be designed. The directional analysis and the diffuseness analysis with the results are shown in the sections 4.3 and 4.4. The collected information and metadata (onsets, the angles azimuth and elevation and the diffuseness parameters) are required for the following spatial re-synthesis process in chapter 5. All processes described in the following steps are performed on all spatial impulse responses measured from every loudspeaker in the MUMUTH and the CUBE (see section 3.4 and section 3.5).

4.2 Onset Detection

Onset detection is used to detect the beginning of discrete events in acoustical signals [7]. In this case the short transient events in the spherical harmonics representation responses must be detected. Figure 4.2 depicts the definition of onset, transient and attack. An onset is defined as the point in time, where a significant change happens in the signal characteristics. This point occurs at the beginning of an event or a transient. The attack refers to the time period in which the amplitude envelope increases, while the transient refers to short intervals in which the signal evolves quickly in a non-trivial and unpredictable way.

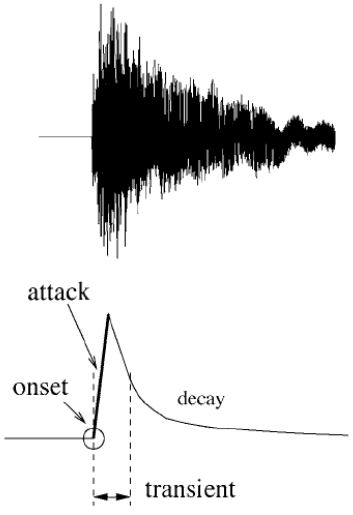


Figure 4.2: Onset Illustration [7]

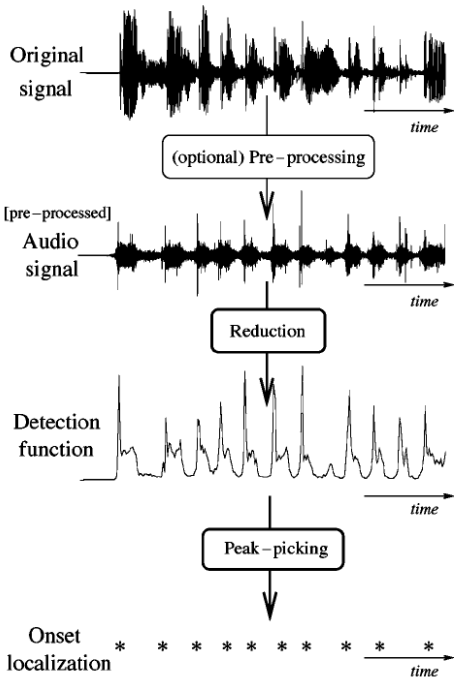


Figure 4.3: Onset Detection Procedure [7]

The general structure of the onset detection procedure is shown in Figure 4.3. In an optional pre-processing step it is possible to accentuate important aspects of the original signal for further analysis. The signal is transformed into the so called detection function, an intermediate representation of the signal itself in which the following peak-picking algorithm locates the most likely onset candidates.

In this thesis the direct sound and the reflections from the measured spatial room responses are determined by an onset analysis of the undirected W-channel for each loudspeaker. Several onset approaches (see section 4.2.2) were applied and evaluated to get an appropriate method for detecting the short transient events in the room impulse responses.

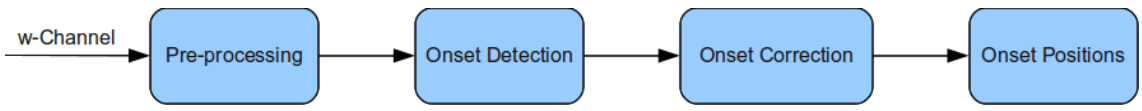


Figure 4.4: Implemented Onset Detection Procedure

The implemented onset detection process is shown in the flowchart in Figure 4.4. As a first step the measured W-channel is pre-processed (see section 4.2.1) and prepared for calculating the detection function in the actual onset detection procedure. The process to receive a suitable representation of the detection function depends on several factors and variables. These variables can vary from method to method and must be carefully set by hand. The exact settings in the different implemented onset detection methods for obtaining an appropriate detection function of the signal are discussed in the corresponding subsections to section 4.2.2. The peak-picking algorithm described in section 4.2.3 chooses the peaks according to the local maxima of the detection functions that are most likely to be onset events. Finally incorrectly detected onsets will be removed (see section 4.2.4) to get the correct onset positions for further analysis and synthesis of the room impulse responses.

4.2.1 Pre-Processing

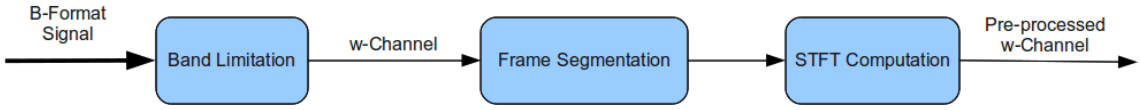


Figure 4.5: Pre-Processing for Onset Detection

In a general pre-processing phase (Figure 4.5) the loaded B-Format signals are filtered with an ordinary 2nd-order Chebyshev band-pass filter with a low cut-off frequency of 50 Hz and a high cut-off frequency of 4 kHz. Only approximately the first 200 ms of the measured impulse responses are used in the further calculations during this analysis phase (cf. Figure 3.5 and Figure 3.7). In both the MUMUTH and CUBE, no significant singular detectable reflections will occur after this period and thus the impulse response after 200 ms is irrelevant for the calculation of onset events. Thereafter, the omni-direct W-channel will be zero-padded to ensure correct onset detection at the start and the end, before dividing it into time frames for the following STFT[4] (Short-Time Fourier Transform) computation. The time resolution must be kept very small in order to detect tran-

[4] Note that for the time domain approach (cf. section 4.2.2.1) no STFT computation is needed in the pre-processing phase.

sient events, which entail a reduction in frequency resolution, but this is not an issue here. For the time fragmentation a window length of approximately 1,45 ms (64 samples, f_s[5] = 44100 Hz), a hop-size[6] from 0,36 ms (16 samples, f_s = 44100 Hz) and windowing with a Hann window function to minimize spectral leakage effects were elected. The coefficients from a Hann window are computed from the equation

$$w(m) = \frac{1}{2}\left(1 - \cos\left(2\pi\frac{m}{M}\right)\right), \qquad 0 \leq m \leq M \tag{4.1}$$

with a window length $L = M + 1$.

4.2.2 Onset Approaches

4.2.2.1 Time Domain Approach

When a transient event in the signal occurs, the signal amplitude increases significantly within a short time period [7]. For percussive and non-complex signals the onsets can be reliably detected by following the amplitude envelope. Such an envelope follower can be constructed by rectifying and low-pass filtering the signal *x(n)* as follows

$$E(n) = \frac{1}{N}\sum_{m=-\frac{N}{2}}^{\frac{N}{2}} |x(n+m)|w(m) \tag{4.2}$$

where *w(m)* is a *N*-point smoothing window. Eq. (4.3) describes a variation of eq. (4.2) where the signal is not rectified but squared to follow the local energy.

$$E(n) = \frac{1}{N}\sum_{m=-\frac{N}{2}}^{\frac{N}{2}} [x(n+m)]^2 w(m) \tag{4.3}$$

[5] f_s … sampling rate

[6] The hop-size is the duration between the start of successive time windows.

However the energy envelope is not really suitable for the following onset detection process, because the located maximum peak does not occurs until a certain time after the real onset. Especially low frequencies need some time until they reach the point where the increase in amplitude is at its sharpest. The use of the time derivative (eq. (4.4)) or the logarithm of the signal energy or a combination of both (eq. (4.5)), provides better results [8].

$$d(n) = \frac{d\big(E(n)\big)}{dt} \tag{4.4}$$

$$d_{log}(n) = \frac{d(\log E(n))}{dt} \tag{4.5}$$

Additionally, sharper peaks and a further refinement of the detection function result (Figure 4.6 and Figure 4.7).

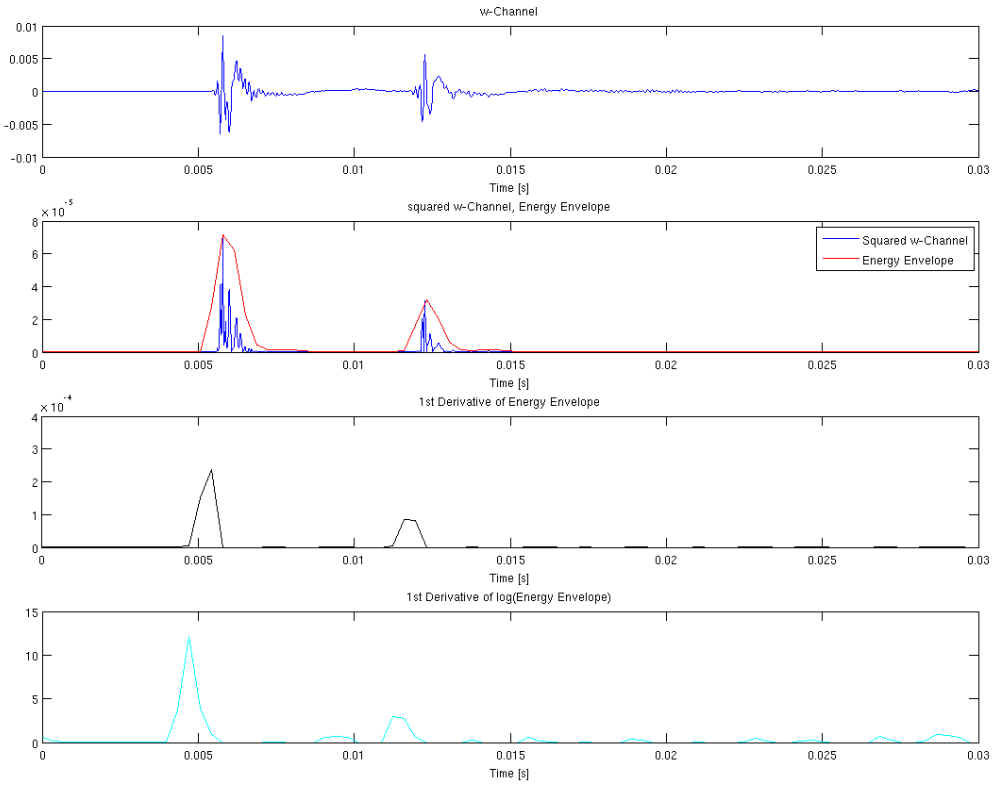


Figure 4.6: Energy Envelope Detection Functions (first: measured W-channel room impulse response including the direct sound and the first reflection from the loudspeaker number 1, MUMUTH; second: squared W-channel and the energy envelope (red); third: 1st derivative of the energy envelope (black); fourth: 1st derivative of the logarithm of the energy envelope (cyan))

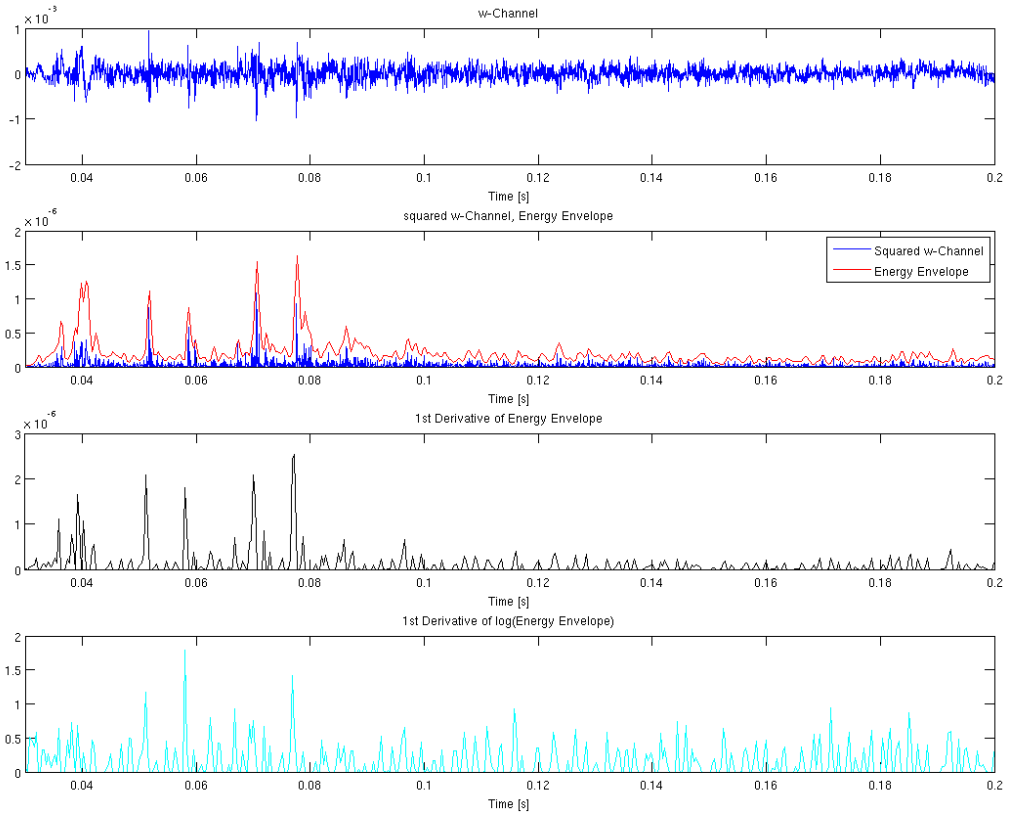


Figure 4.7: Energy Envelope Detection Functions (first: measured W-channel room impulse response including further reflections from the loudspeaker number 1, MUMUTH; second: squared W-channel and the energy envelope (red); third: 1st derivative of the energy envelope (black); fourth: 1st derivative of the logarithm of the energy envelope (cyan))

Using the impulse response signal measured for loudspeaker number 1 in the MUMUTH, Figure 4.6 and Figure 4.7 show the calculations of the onset detection functions. These functions are derived from the time domain onset approach described in this section. For the sake of clarity, Figure 4.6 shows approximately the first 30 ms of the recorded impulse response, where the more pronounced events of the direct sound and the first reflection can be seen. Figure 4.7 displays the following range from about 30 ms to 200 ms, where further reflections occur. In both figures the recorded omnidirected W-channel of the loudspeaker is seen in the top representation. The squared W-channel of the impulse response (blue) with the calculated energy envelope from eq. (4.3) is the second from the top. The first derivative of the energy envelope (black) from eq. (4.4) is below that. The final illustration represents the first derivative of the logarithm of the energy envelope (cyan) from eq. (4.5). Pre-processing before calculation is done as described in section 4.2.1 with band limitation and a 64 samples moving aver-

age window (cf. eq. (4.3)) with a hop-size from 16 samples at 44100 Hz sampling rate. A Hann window function and zero-padding were chosen.

In the case of the energy envelope of the measured impulse response (red line), it is clearly visible that the shape is significantly broader compared to the other two calculations of the first derivative of the energy envelope (black and cyan line). Furthermore it is worth noting that the maximum of the energy envelope takes place at the maximum of the incoming events in the signal and thus does not define the actual onset (cf. Figure 4.2) but rather the time where the most energy is present. Peaks in the functions of the first derivative of the energy envelope (third and last representation) occur earlier in time and define the true onset time better (cf. [8]).

4.2.2.2 Frequency Domain Approach

In contrast to the previous section, in the following an onset approach in the frequency domain is considered. There the energy function is given as the sum of the magnitude squared of each frequency bin

$$E(n) = \frac{1}{N}\sum_{k=0}^{\frac{N}{2}} |X_k(n)|^2 \tag{4.6}$$

where N is the FFT array length and $X_k(n)$ is the k^{th} bin of the FFT.

The increase of energy in the transients tends to be distributed in a broad band. The majority of this energy resides in general in the lower frequencies of the spectrum, therefore the energy increases in higher frequencies are more indicative of onsets. A linear weighting k (cf. eq. (4.7)) toward the high frequencies is applied to make these changes even clearer. In [43] Masri found out that the High Frequency Content (HFC) approach produces sharp peaks during attack transients and performs well when strongly percussive signals are analyzed. The mathematical expression is given by

$$E_{HFC}(n) = \frac{1}{N}\sum_{k=0}^{\frac{N}{2}} |X_k(n)|^2 k \tag{4.7}$$

The process of the HFC implementation is shown in Figure 4.8. As described in section 4.2.1, the W-channel signal first has to pass through the pre-processing phase and is transformed into frequency domain, before the detection function can be calculated. The post-processing and subsequent peak picking algorithm, to detect the onset positions is explained in section 4.2.3.

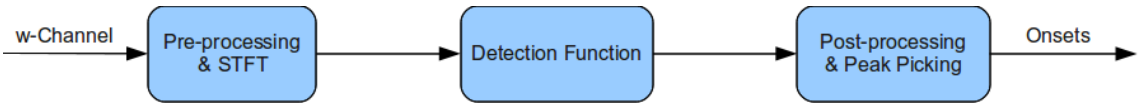


Figure 4.8: HFC Onset Detection

The comparison of the results from the HFC approach and the following Complex Domain approach is given in Figure 4.11 and Figure 4.12.

4.2.2.3 Complex Domain Approach

In the method mentioned in section 4.2.2.2, onset detection is achieved by looking at energy from a signal in the frequency domain. The phase data also provides information about the onsets in a signal spectrum. If the difference between the actual phase and the expected phase value is high or not zero, this is an indication of possible onset points.

The Complex Domain Approach method creates an onset detection function by combining information of the actual and intended propagation of magnitude and phase of a signal in the complex frequency domain [7] [44] [45] [46]. A phase detection onset approach is more robust at the lower part of the spectrum and works well to find soft onsets whereas the magnitude approach is more reliable when using higher frequency information (see section 4.2.2.2) and more efficient for percussive onsets. Thus the advantages of both are combined.

The polar form of the observed value for the k^{th} bin from the Short-Time Fourier Transform (STFT) in the complex domain is given by

$$X_k(n) = |X_k(n)|e^{j\phi_k(n)} \tag{4.8}$$

where $X_k(n)$ is the magnitude and φ_k the phase of the current STFT frame. It can be considered, that the expected combination of spectral magnitude and phase for the k^{th} bin of the STFT of a signal is given by

$$\hat{X}_k(n) = \left|\hat{X}_k(n)\right| e^{j\hat{\phi}_k(n)} \tag{4.9}$$

where the target amplitude $\hat{X}_k(n)$ is the expected magnitude value and should be equal to the magnitude of the previous frame

$$\left|\hat{X}_k(n)\right| = |X_k(n-1)| \tag{4.10}$$

and the target phase $\hat{\phi}_k$ is the expected phase value as the sum phase difference between preceding frames and the previous phase.

$$\hat{\phi}_k = princarg[2\varphi_k(n-1) + \varphi_k(n-2)] \tag{4.11}$$

The *princarg* function maps the phase to the $[-\pi, \pi]$ region.

By rotating $\hat{X}_k(n)$ onto the real axis (cf. Figure 4.9) and consequent setting of the expected phase value $\hat{\phi}_k$ to zero, the Euclidean distance[7] between the expected complex vectors and the actual vectors is calculated:

$$\Gamma_k(n) = \sqrt{\left\{\left|\hat{X}_k(n)\right|^2 + |X_k(n)|^2 - 2\left|\hat{X}_k(n)\right||X_k(n)|\cos d_{\varphi_k}(n)\right\}} \tag{4.12}$$

[7] The Euclidean distance defines the distance between two points on a plane or in space. It examines the root of square differences between coordinates of a pair of objects.

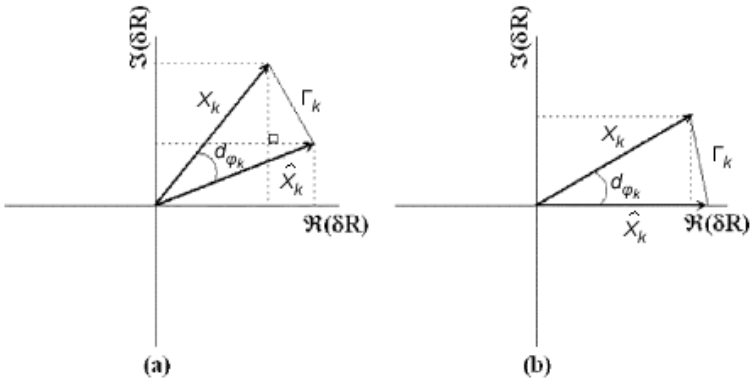


Figure 4.9: Complex frequency domain phasor diagram illustrating the phase deviation (d_{φ_k}) between current vector (X_k) and target vector ($\hat{X}_k$) and the Euclidean distance (Γ_k). (a) normal and (b) rotated diagram [44]

The phase deviation between the real phase values and the target values in a given frame can be expressed as

$$d_{\varphi_k} = princarg[\varphi_k(n) - 2\varphi_k(n-1) + \varphi_k(n-2)]. \tag{4.13}$$

d_{φ_k} is equal to zero if the difference between the current phase value and the estimated phase value remains constant. Thus the detection function is only created by the energy difference and otherwise, if the difference deviate from zero significantly, the detection function also affected by the phase value.

The final detection function is achieved by summing the complex distance measures for each frame *k* defined by

$$\eta_c(n) = \sum_{k=0}^{\frac{N}{2}} \Gamma_k(n) \tag{4.14}$$

Figure 4.10 shows the complex domain implementation process.

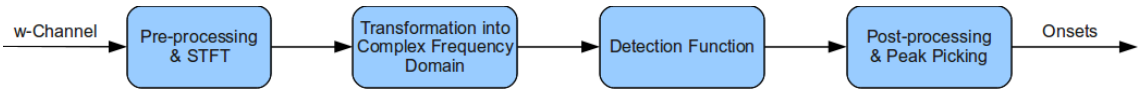


Figure 4.10: Complex Domain Onset Detection

For the implementation the same pre-processing stage as in the HFC approach is used (cf. section 4.2.1) to standardize the W-channel signal representation. Band limitation, zero-padding and fragmentation with a Hann window function, with a 64 samples moving average window and a hop-size of 16 samples at 44100 Hz sampling rate, were chosen before calculating the STFT for every windowed frame. From the STFT values of the pre-processed signal, the magnitude and the phase is extracted to get a signal representation in the complex frequency domain. A detection function (cf. eq. (4.14)) is created using the deviation between the expected and detected magnitude and phase values of the spectrum.

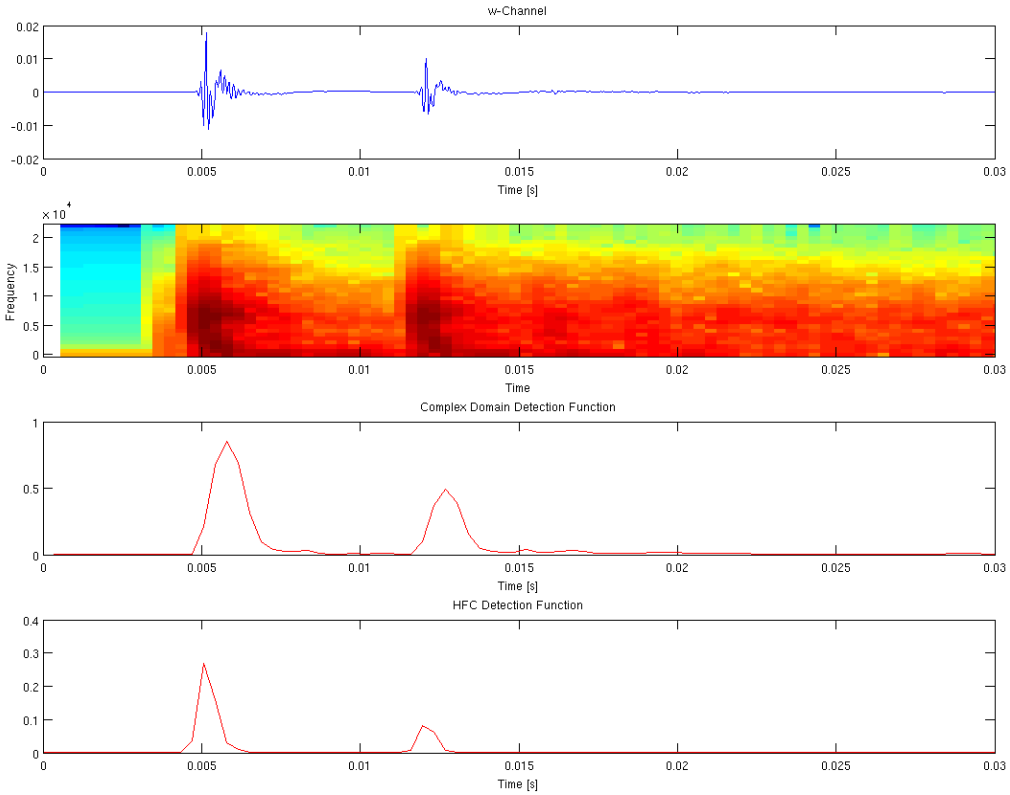


Figure 4.11: Complex Domain and HFC Detection Functions (first: measured W-channel room impulse response including the direct sound and the first reflection from the loudspeaker number 1, MUMUTH; second: spectrogram; third: complex domain detection function; fourth: HFC detection function)

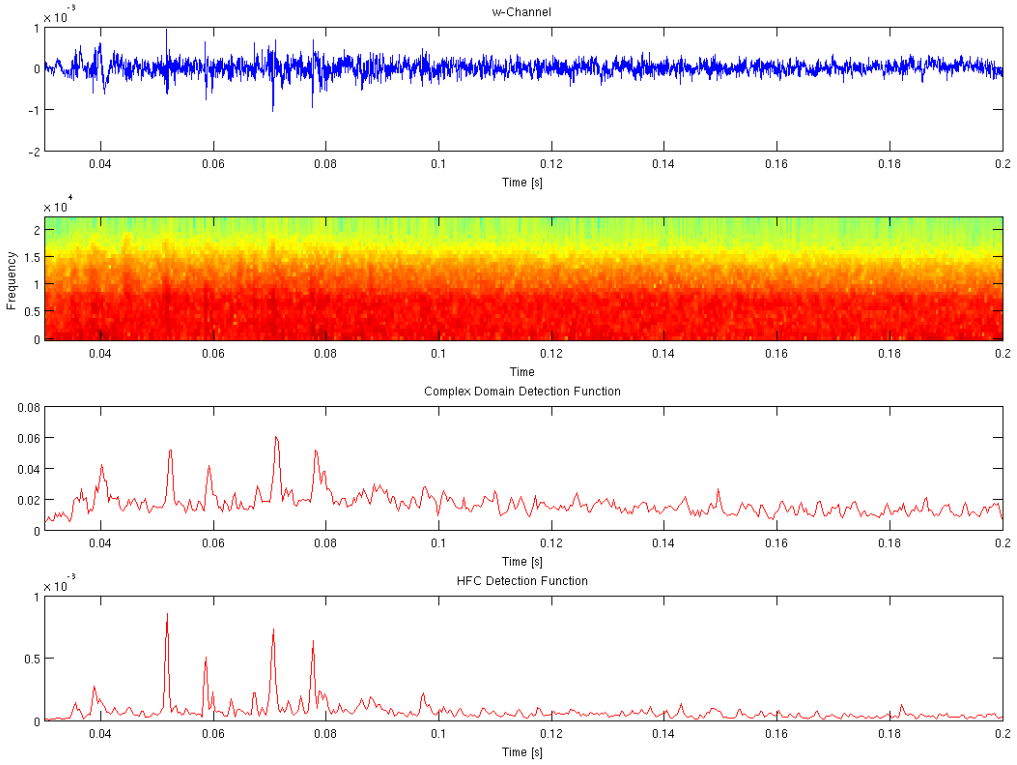


Figure 4.12: Complex Domain and HFC Detection Functions (first: measured W-channel room impulse response including further reflections from the loudspeaker number 1, MUMUTH; second: spectrogram; third: complex domain detection function; fourth: HFC detection function)

Figure 4.11 and Figure 4.12 provide a comparison of the onset detection functions of both the HFC onset approach and the complex domain onset approach. This shown using the impulse response signal measured for loudspeaker number 1 in the MUMUTH. For the sake of clarity, Figure 4.11 shows approximately the first 30 ms of the recorded impulse response, where the more pronounced events of the direct sound and the first reflection can be seen. Figure 4.12 displays range about between 30 ms to 200 ms, during which further reflections occur. In both figures the recorded omni-directed W-channel of the loudspeaker is seen in the top representation. Second the spectral representation, showing the spectral density of the W-channel in reference to the time representation on the x-axis and the frequency representation on the y-axis. Third, the onset detection function from the complex domain approach is depicted while at the bottom the onset detection function from the HFC approach can be seen.

In a further step, the onset events are collected by post-processing and by finding local maxima from the detection function. Peak picking is performed as described in section 4.2.3. This is then and applied to the different detection functions. A comparison can be seen in Figure 4.14 and Figure 4.15.

4.2.3 Peak Picking

The previous sections show the different approaches for deriving the onset detection functions from the impulse responses of the rooms which were measured. Based on these intermediate representations of the signal, the main goal in the peak picking stage, is to detect significant peaks and to eliminate spurious peaks and noise. This algorithm is used with all onset detection function implementations described in the previous sections. The single steps for the peak picking process are displayed in Figure 4.13.

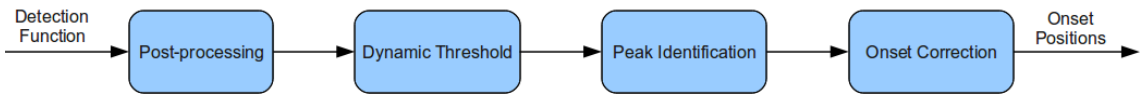


Figure 4.13: Peak Picking

Before the onsets with local maxima can be found, a post-processing of the detection function is necessary [7]. The intention of post-processing is to increase the consistency of event-related features in the detection function in order to facilitate the following processes of thresholding and peak-picking. In a first step the detection functions were normalized by subtracting the mean of the function from the function itself and then dividing it by the maximum absolute deviation. Subsequently low-pass filtering was employed to smooth the curve and eliminate spurious peaks. To pick out the relevant local maxima and subsequently find the onsets, a dynamic threshold curve $\tilde{\delta}(n)$ is used (eq. (4.15)). This is a transformed version of the detection function signal itself, where a fixed threshold is combined with an adaptive threshold obtained by using a median filter [44] [47].

$$\tilde{\delta}(n) = \delta + \lambda \cdot median\big(\eta(k_n)\big), \qquad k_n \in \left[n - \frac{H}{2}, n + \frac{H}{2}\right] \tag{4.15}$$

With the constant value δ representing the fixed threshold and the scaling factor λ, which dictates the influence of the nearby frames on the threshold of the current frame. H stands for the window length of the median calculation. Compared to δ, λ has far less influence on the correct and incorrect detection of onsets. Therefore, δ should be chosen with care. The range of these values depends on a number of factors such as the type of the detection function, the signal level, STFT size or the window length. Due to the fact that the directional room impulse responses measured at the MUMUTH and the CUBE have different signal levels, different parameter settings are necessary. The settings were set by hand for the different detection functions and for both rooms based on experimentation. The constant δ was always calculated from the median of the post-processed detection function. Thus δ is given a different value for each measured impulse response of each loudspeaker. The scaling factor λ is chosen differently for each case, depending on the room analyzed and onset approach. The selected λ-value for the respective detection functions are listed in **Table 4.1**.

	Complex Domain Approach	HFC Approach
MUMUTH	$\lambda = 0.3$	$\lambda = 0.75$
CUBE	$\lambda = 0.5$	$\lambda = 1$

Table 4.1: Selected values for the scaling factor λ in the dynamic threshold curve in eq. (4.15) for the analyzed rooms (MUMUTH and CUBE) according to the onset approach.

In Figure 4.14 and Figure 4.15 the calculated onsets for the measured impulse response from loudspeaker number 1 in the MUMUTH and the CUBE using the HFC and complex domain approaches are shown. All values beneath the threshold are set to zero and every local maximum above zero was counted as an onset.

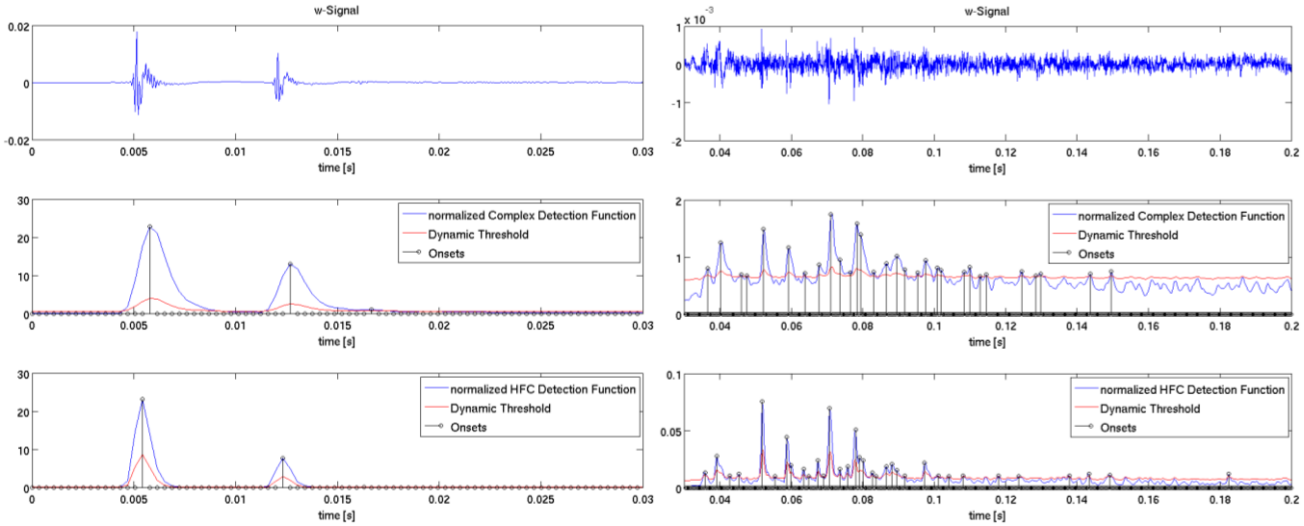


Figure 4.14: Normalized Detection Functions with Dynamic Threshold Curves and Detected Onsets from the Measured Directional Room Impulse Response from Loudspeaker Number 1, MUMUTH. (left: direct sound and 1st reflection, right: further reflections, top: the original omni-directed W-channel, middle: complex domain approach, bottom: HFC approach)

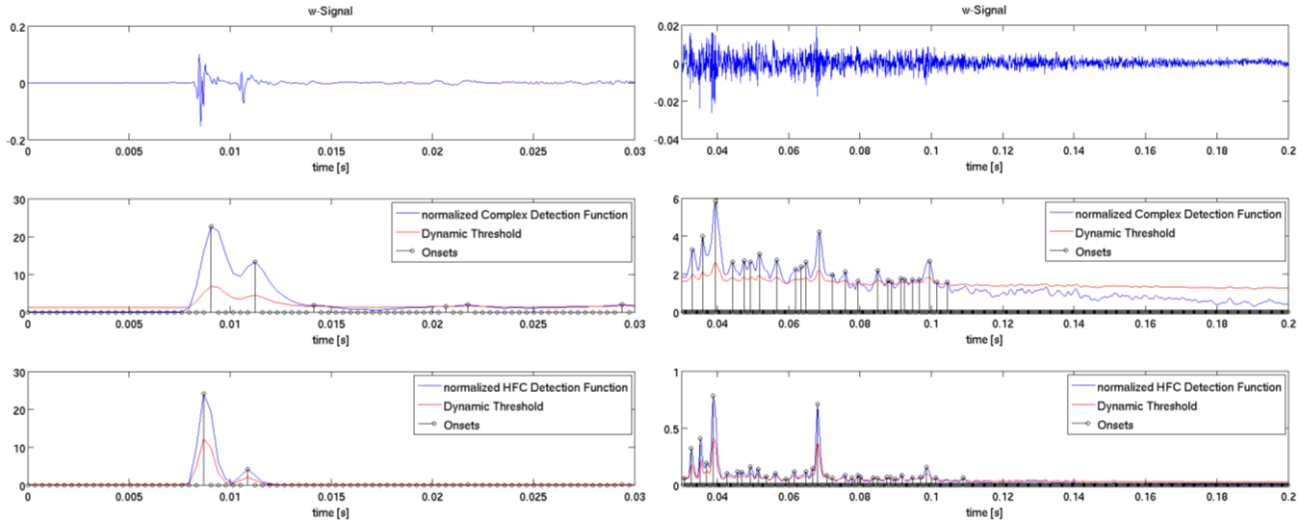


Figure 4.15: Normalized Detection Functions with Dynamic Threshold Curves and Detected Onsets from the Measured Directional Room Impulse Response from Loudspeaker Number 1, Cube. (left: direct sound and 1st reflection, right: further reflections, top: the original omni-directed W-channel, middle: complex domain approach, bottom: HFC approach)

A comparison of the two onset detection approaches shows that, unlike the detection function for complex domain processes, the detection function of the HFC procedure provides narrow and more differentiated peaks. Roughly speaking, the detected onsets are closely related to the expected raise in the amplitude due the incoming sound event for the HFC approach. Thus, this approach may delivers better results in the onset localization. Therefore, all further calculations in this thesis are based on the onset calculation from the HFC approach. However, it is possible to select the two approaches in the software implementation.

As previously described, the localization of correct and incorrect onsets depends on the choice of different parameters in the onset detection process. Hence, it cannot be ruled out completely, that some maxima above the threshold curve are falsely categorized as onsets. For example, in the middle representation out of Figure 4.14 one can see incorrect onset detection at approximately 10 ms after the direct sound occurs, resulting from the first reflection post-oscillation. Section 4.2.4 shows how to eliminate incorrect onset detections.

4.2.4 Onset Correction

While the above mentioned approaches work quite well for detecting probable onsets by a detection function, additional steps have been taken to improve the detection results. After the adaptive threshold has elicited potential onset candidates, the following criteria were used to eliminate false positives.

Global Diffuseness

- If onsets were detected in regions where the global diffuseness percentage of the W-signal is very high, the onsets are regarded as incorrect and are thus removed. The greater the value of diffuseness, the lower the proportion of an incoming direct sound event.

- In addition, onsets which are not in the immediate vicinity of a minimum in the diffuseness curve are eliminated. A minimum in the diffuseness curve is an indication of an incoming sound event in the signal.

- If adjacent onsets are located closer together than 1 ms, the onset with the position next to the lower diffuseness value is retained and the other one is removed. If onsets are located to close together, a separate analysis is no longer possible.

Furthermore, detected onsets are eliminated if they arrive a certain period after the direct sound. After this period of time no reflections other than diffuse sound can be expected, due to the size of the rooms (cf. section 3.4 and section 3.5). For the measured impulse responses in the MUMUTH a time of 150 ms after the incoming direct sound was chosen and for the impulse responses in the CUBE a time of 100 ms was chosen.

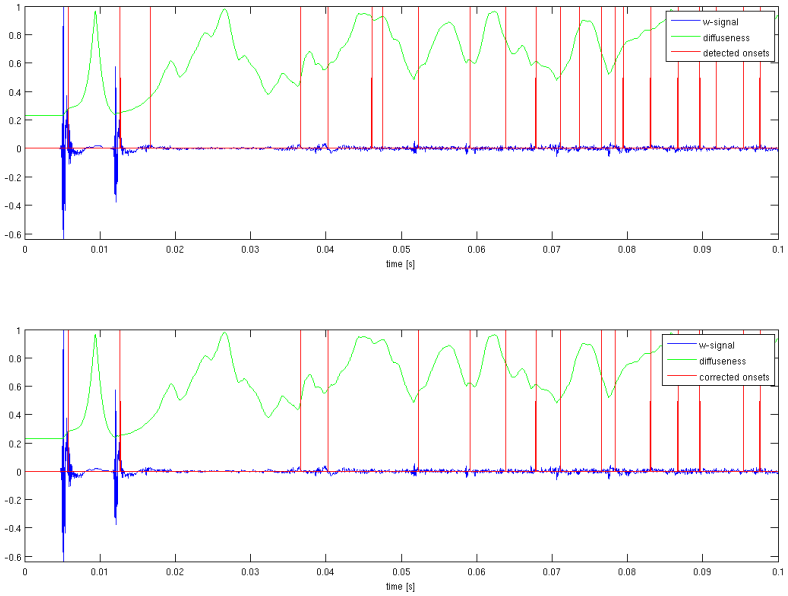


Figure 4.16: Onset Correction at Measured Impulse Response from Loudspeaker Number 1, MUMUTH. (blue: W-channel, green: global diffuseness, red: detected onsets (top) respectively corrected onsets (bottom))

Figure 4.16 shows an example of an onset correction according to the criteria listed above. The blue curve represents the W-channel of the measured impulse response from loudspeaker number 1 in the MUMUTH, the green line the global diffuseness and in red line the detected onsets in the illustration on the top. The respective corrected onsets can be seen in the illustration at the bottom. In this case incorrect detected onsets can be located at approximately 17 ms, 46 ms or 48 ms and will be sorted out. The calculation of the diffuseness is described in more detail in section 4.4.

4.3 Directional Analysis

After the onsets of the direct sound and the reflections are determined for the measured spatial directional impulse responses, this section deals with the directional analysis of these detected onset positions. Figure 4.17 shows the directional analysis process which was implemented, leading up to the determination of the azimuth and elevation angle information.

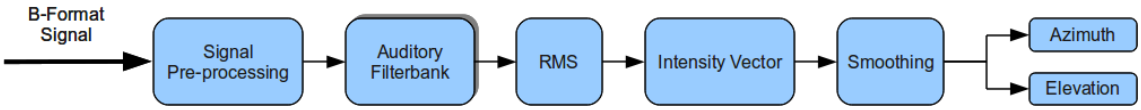


Figure 4.17: Directional Analysis Procedure

The recorded B-Format signals are subjected to a pre-processing phase. This course of action is described and presented in more detail in section 4.3.1. Following this, a division into separate frequency bands is achieved by filtering with an auditory filterbank (see section 4.3.2). Later intensity vectors are calculated (see section 4.3.3) to finally determination the direction angles of the incoming sound events.

4.3.1 Signal Pre-Processing

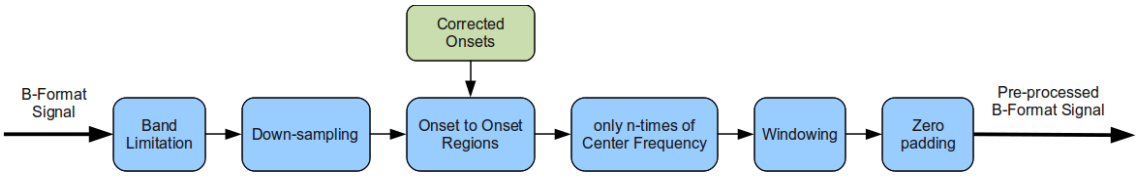


Figure 4.18: Signal Pre-Processing for the Analysis Phase

The following list describes the single steps of the signal pre-processing stage, shown in Figure 4.18:

- A general band limitation from 50 Hz to 4 kHz is performed. The same ordinary 2^{nd}-order Chebyshev band-pass filter as in the pre-processing stage is used before the onset detection is calculated, as described in section 4.2.1.

- A down-sampling to a lower sample rate (11025 Hz) is performed. One reason for this is the reduction in data volume and thus of computing time. The primarily it is done to guarantee the stability of filtering the signals with the auditory filterbank as described in section 4.3.2. This is especially relevant for lower frequency bands.

- The analysis of the transient regions starts 1 ms before an onset occurs, and extends to 1 ms before the next onset occurs. The maximum overall duration of these regions is restricted to 10 ms. Similarly, a maximum of 10 ms is allowed after the final onset. The range of 10 ms is based on the fact that the lowest frequency band to be analyzed has a center frequency of 100 Hz (see section 4.3.2). Thus, a whole wavelength is used for analysis.

- Depending on the frequency bands defined by the following band filtering, the regions are adjusted to n-times periods of the examined center frequency. Primarily, this task has an effect on the higher frequency bands, where a smaller wavelength is found. Thus, it is guaranteed that only the first n-time periods with the greatest energy resources are analyzed.

- Windowing [48], using a Tukey window function (eq. (4.16)), as well as additional zero-padding is applied to reduce artifacts at the beginning and at the end of the selected regions. The Tukey window is defined as

$$w(m) = \begin{cases} \frac{1}{2}\left\{1 + \cos\left(\frac{2\pi}{\alpha}\left[m - \frac{\alpha}{2}\right]\right)\right\}, & 0 \le m < \frac{\alpha}{2} \\ 1, & \frac{\alpha}{2} \le m < 1 - \frac{\alpha}{2} \\ \frac{1}{2}\left\{1 + \cos\left(\frac{2\pi}{\alpha}\left[m - 1 + \frac{\alpha}{2}\right]\right)\right\}, & 1 - \frac{\alpha}{2} \le m < 1 \end{cases} \qquad (4.16)$$

where m is a N-point linearly spaced vector and α denotes the ratio tapered section to constant section with $0 \le \alpha \le 1$. In this case, the ratio of the Tukey window is set to $\alpha = 0.1$, in order to window the regions.

4.3.2 Auditory Filterbank

The pre-processed signals are divided into several frequency bands for further analysis. For this purpose, a filterbank according to the gammatone filter model [49] is used. This model is based on the Equivalent Rectangular Bandwidth (ERB) scale of Glasberg and Moore [50]. A set of overlapping band-pass filters is designed for the filterbank. The bandwidth of one filter is derived from the formula

$$\Delta f_{ERB} = 24.7\left(\frac{4.37 f_c}{1000} + 1\right) \tag{4.17}$$

where f_c is the center frequency of the band in Hz and Δf_{ERB} is the bandwidth. In Patterson's filter model [51] each designed band-pass filter is designed one ERB wide, while the frequency spacing between the respective frequency bands is not specified. In this thesis, a subdivision of the frequency range in 50 frequency bands was chosen, starting at a center frequency of 100 Hz and ending at a center frequency of 4 kHz. The resulting filters are illustrated in Figure 4.19 and the different ERB bands with the corresponding center frequencies are shown in Table 4.2.

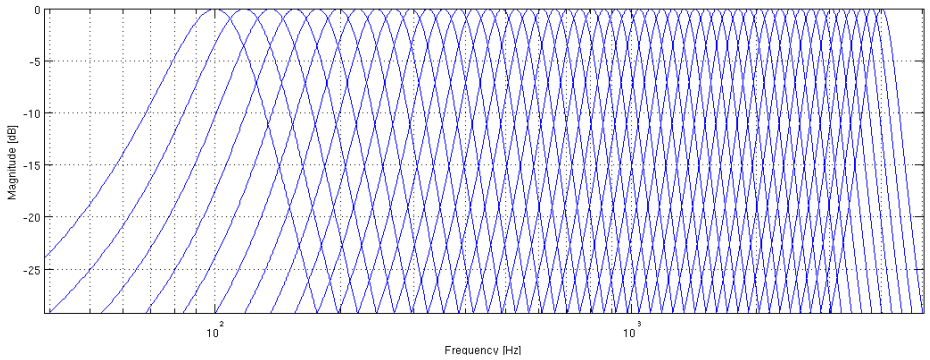


Figure 4.19: Frequency Responses of the Used Gammatone Filterbank with 50 Frequency Bands, covering 100-4000 Hz Band.

ERB band	Center-frequency [Hz]	ERB band	Center-frequency [Hz]	ERB band	Center-frequency [Hz]	ERB band	Center-frequency [Hz]	ERB band	Center-frequency [Hz]
1	100.0	11	325.0	21	703.8	31	1341.9	41	2416.5
2	117.6	12	354.6	22	753.7	32	1426.0	42	2558.1
3	136.1	13	385.8	23	806.3	33	1514.5	43	2707.2
4	155.7	14	418.7	24	861.7	34	1607.8	44	2864.3
5	176.2	15	453.4	25	920.1	35	1706.1	45	3029.8
6	197.9	16	489.9	26	981.5	36	1809.6	46	3204.1
7	220.7	17	528.3	27	1046.3	37	1918.7	47	3387.8
8	244.8	18	568.8	28	1114.5	38	2033.6	48	3581.3
9	270.1	19	611.5	29	1186.4	39	2154.6	49	3785.2
10	296.8	20	656.5	30	1262.1	40	2282.2	50	4000.0

Table 4.2: ERB Bands with the Corresponding Center Frequencies

For further illustration Figure 4.20 shows the gammatone filtered W-channel of the room impulse response measured from loudspeaker number 1 in the MUMUTH (Figure 3.5, top) in the divided frequency bands. Note that this illustration display the global signal to the extent of an elapsed time from approximately 200 ms and not the previously described detected regions according to the onsets.

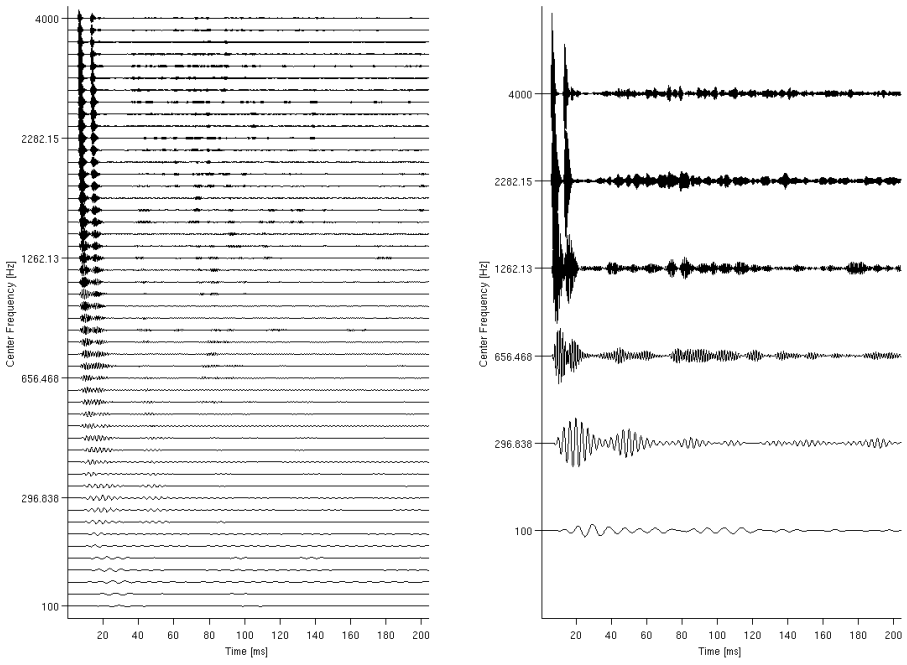


Figure 4.20: W-Channel Room Impulse Response Filtered with the Gammatone Filterbank from the Loudspeaker Number 1, MUMUTH. (left: all 50 frequency bands, right: 6 selected frequency bands for better representation)

4.3.3 Intensity Vectors

Using the concept of sound intensity, the transfer of energy within a sound field at the measurement position can be defined as the instantaneous intensity $\boldsymbol{I}(n)$. It is calculated as the product of the sound pressure $p(t)$ and the particle velocity vector $\boldsymbol{u}(t)$ [52].

$$\boldsymbol{I}(n) = p(n)\boldsymbol{u}(n) \tag{4.18}$$

The sound pressure $p(n)$ is proportional to the omni-directional signal $W(n)$ (cf. eq. (4.19)) and the particle velocity $\boldsymbol{u}(n)$ corresponds to the figure-of-eight signals $X(n)$, $Y(n)$ and $Z(n)$ (cf. eq. (4.20)).

$$p(n) = W(n) \tag{4.19}$$

$$\boldsymbol{u}(n) = \frac{1}{\sqrt{2}\mathrm{Z}_0}[X(n), Y(n), Z(n)]^T \tag{4.20}$$

The factor $\frac{1}{\sqrt{2}}$ applies to the scaling of the dipole X-, Y- and Z-signals of the B-Format microphone (see section 3.3) and Z_0 refers to the characteristic acoustic impedance with

$$Z_0 = \rho_0 c \tag{4.21}$$

where ρ_0 is the mean density of air and c the speed of sound [52].

Based on the physics, low and high frequency components of transient events evolve differently over time and neighboring onsets can overlap in time [25]. Thus the calculation of intensity vectors is split into several frequency bands (see section 4.3.2). The instantaneous intensity vector can be written as

$$\boldsymbol{I}(n, f_c) = \frac{1}{\sqrt{2}\mathrm{Z}_0} W(n, f_c)\left[X(n, f_c)\boldsymbol{e}_x + Y(n, f_c)\boldsymbol{e}_y + Z(n, f_c)\boldsymbol{e}_z\right] \tag{4.22}$$

where $\boldsymbol{e}_x$, $\boldsymbol{e}_y$ and $\boldsymbol{e}_z$ represent Cartesian unit vectors, n is the time index and f_c denotesthe frequency channel. The instantaneous intensity vector consists of the components

in accordance with the coordinate axis (eq. (4.23)). In this analysis the multiplier $\frac{1}{\sqrt{2}Z_0}$ implemented in eq. (4.20) and eq. (4.22) can be discarded because this constant gain factor does not affect the analyzed direction.

$$\boldsymbol{I}(n, f_c) = \left[I_x(n, f_c), I_y(n, f_c), I_z(n, f_c)\right]^T \tag{4.23}$$

Before using the calculated intensity vectors to estimate the direction angles, they are passed through an adaptive 1-pole low-pass filter (eq. (4.24)) for smoothing reasons.

$$s(n) = (1 - \alpha) \cdot r(n) + \alpha \cdot s(n - 1) \tag{4.24}$$

The pole α is a constant factor and for this application it was found out, in practical experiments, that a value $\alpha = 0.9$ performed well.

Afterwards the azimuth angle $\theta(n, f_c)$ and the elevation angle $\varphi(n, f_c)$ can be written in the form

$$\begin{aligned} \theta(n, f_c) &= -\operatorname{atan2}\left(I_y(n, f_c), I_x(n, f_c)\right) \\ \varphi(n, f_c) &= -\operatorname{atan2}\left(I_z(n, f_c), \sqrt{{I_x}^2(n, f_c) + {I_y}^2(n, f_c)}\right) \end{aligned} \tag{4.25}$$

where the two-argument atan2 function is provided to eliminate quadrant confusion. The direction of arrival can be expressed as the opposite direction to the calculated intensity vector, which is carried out with the minus sign in eq. (4.25).

In addition to the calculation of the intensity vectors an energy criterion is introduced to define relevant sub-regions within the selected regions. A threshold is defined by the Root Mean Square (RMS) of the undirected W-channel, shown in eq. (4.26), within the region to be analyzed.

$$W_{RMS}(f_c) = \sqrt{\frac{1}{N}\sum_{n=1}^{N} W_n^2(f_c)} \tag{4.26}$$

The energy measure ensures that signal components that lie above the RMS value, and therefore where relevant information occurs, are used for the directional analysis. For the remaining selected regions the intensity vector $\boldsymbol{I}(n, f_c)$ is calculated for each point in time using eq. (4.22). From which the angle information for azimuth $\theta(n, f_c)$ and elevation $\varphi(n, f_c)$ can be derived according to eq. (4.25).

The median values of the angles are determined by using circular statistics to obtain a meaningful value for each detected sound event in each frequency band (eq. (4.27)). Circular statistics can be considered a subfield of statistics and is relevant for the development of statistical techniques using angle data. The MATLAB circular statistics toolbox CircStat provides corresponding computational functions and is used in this thesis [53].

$$\begin{aligned} \theta_c &= cmedian\big(\theta(n, f_c)\big) \\ \varphi_c &= cmedian\big(\varphi(n, f_c)\big) \end{aligned} \tag{4.27}$$

Each frequency band is then analyzed with regard to the existing local diffuseness. Thus, a further condition is introduced to any frequency bands in which the diffuse component relative to other frequency bands is too high to be sidelined. If the mean of the diffuseness within a frequency band lies above the mean of the diffuseness of the entire frequency range, they are discarded. By removing this data the impact of errors can be minimized further.

Figure 4.21 shows an example of the local diffuseness in the region where the direct sound occurs. The left shows a measurement taken from loudspeaker number 1, while the right shows loudspeaker number 2 at the MUMUTH (see section 3.4).

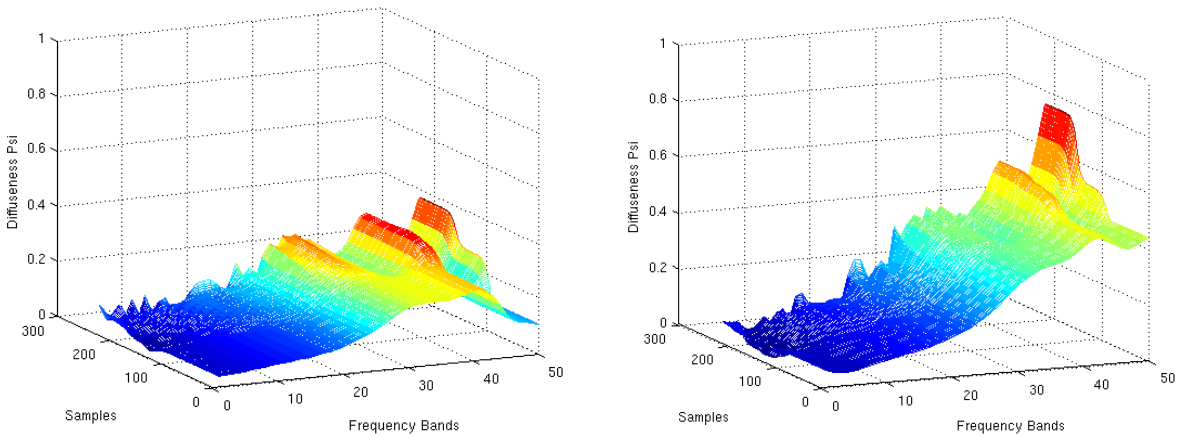


Figure 4.21: Local Diffuseness Estimation from the Direct Sound Region Over Frequency Bands (left: loudspeaker number 1, MUMUTH, right: loudspeaker number 2, MUMUTH)

Both figures demonstrate the proportion of the diffuseness at higher frequencies is larger than at lower frequencies. Especially in the higher frequency bands, for longer distances from the beginning of the localized sound event, the diffuseness increases much more than at lower frequencies. In such cases the information derived from directional analysis will be meaningless. The images show that it is useful to adjust the regions of high frequency bands prior to analysis. Such a pre-processing to n-times periods of the examined center frequency when analyzing the direction of the sound is described in section 4.3.1. The detailed procedure for calculating the diffuseness analysis is described in section 4.4.

4.3.4 Directional Results

Within this section, the results from the analysis of the direct sound directions gained from the Intensity Vector Approach introduced in section 4.3.3 are presented. Various evaluations concerning the directional analysis and code testing were performed and are briefly outlined at the beginning.

Before the results of the analysis of the measured room impulse responses are evaluated, a test of the implemented analysis algorithm was conducted. The efficiency of the algorithm is tested using virtually generated spatial impulse responses. Testing is carried out both with pure impulses and with noisy test signals. The influence of the intensity of the added noise is investigated with respect to the resulting effects on the correct detection of onsets and the deviations in the direction detection. This is followed by a directional analysis evaluation of the direct sound of the measured room impulse responses. Finally, simple simulations of the analyzed rooms with a room acoustic simulation program are implemented. Both the simulated and real measured impulse responses are compared and the correctness of the onset detection is verified.

Testing the implemented analysis procedure using virtual generated sound sources as a test signal

In order to test the algorithm, a signal with five simple impulse responses from five virtual sound sources in a virtual space were generated based on the model found in [54]. The predefined spherical coordinates, for elevation and azimuth from the virtual loudspeakers, seen from a supposed origin measurement point can be taken from Table 4.3. The sound source positions were chosen at random. The signal was generated in Ambisonics B-Format (cf. section 2.2) by calculating the first-order spherical harmonics according to eq. (2.7) and multiplying each channel with the generated signal (cf. eq. (2.11)).

	Source 1	Source 2	Source 3	Source 4	Source 5
Azimuth θ in [°]	0	-90	180	-90	-135
Elevation φ in [°]	26,6	53,1	26,6	39,8	43,3

Table 4.3: Position of Virtual Sound Sources in Spherical Coordinates for Code Testing.

First, the implemented onset detection analysis code was tested with the pure impulse signals. No additional noise or disturbing influences were present. Figure 4.22 shows the generated W-channel with the generated impulse responses in the upper illustration and the results from the onset detection (see section 4.2) with the complex domain approach and the HFC approach in the two lower illustrations. It is clearly visible in both approaches that the transients are definitely recognized as onsets and there are no false detections.

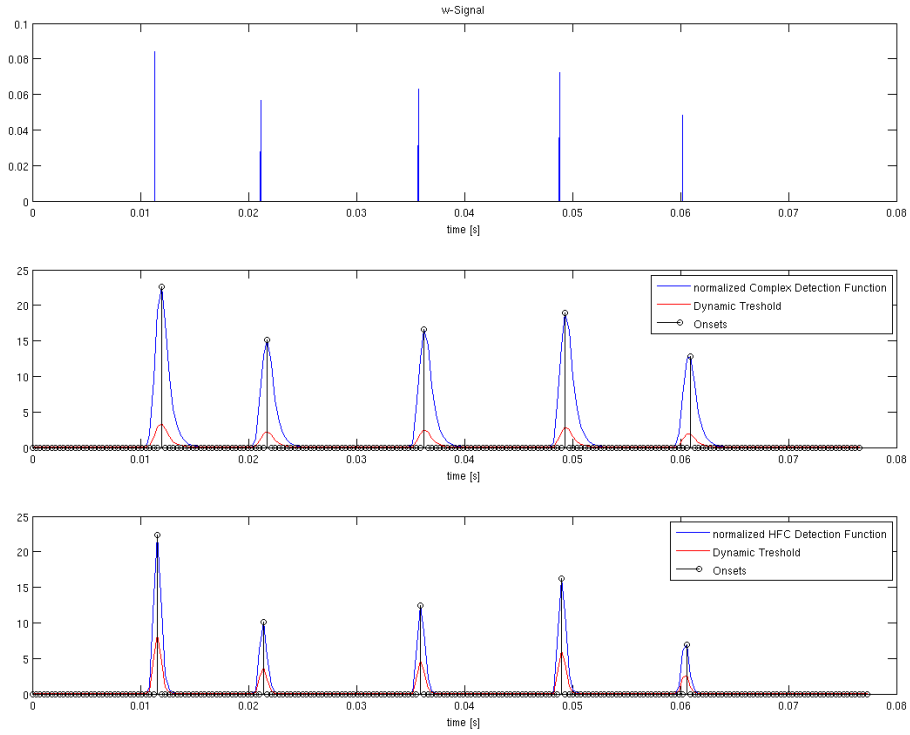


Figure 4.22: Results from the Onset Detection for the Impulse Response Test Signal without Noise. (top: W-channel from the generated impulse response test signal, middle: onset detection with the complex domain approach, bottom: onset detection with the HFC approach)

In a second step, the same generated impulse responses were used with a broad-band white noise signal was added to the signal. Figure 4.23 shows the generated W-channel impulse responses with additional noise. Beneath the results from the onset detection are shown after the complex domain approach and the HFC approach. Here it is visible, that the transients are detected as clear onsets, there are however areas where false onsets are detected. Such false detections can be seen at about 51 ms in the complex domain approach calculation in the middle illustration and at about 69 ms in the HFC approach calculation in the lower illustration. These false detected onsets are eliminated by an onset correction (cf. section 4.2.4). Figure 4.24 shows the W-channel input signal (blue) with the detected onsets and the incorrectly detected onset marked in magenta above and below the corrected version. The global diffuseness curve (green) also shows that at the time where the right onsets occur, the diffuseness is clearly lower compared to the noisy areas.

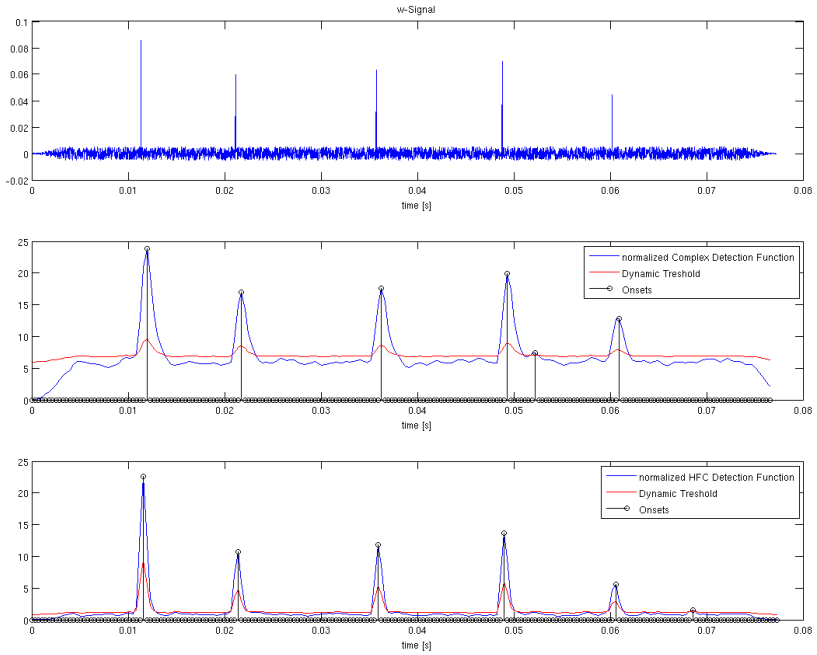


Figure 4.23: Results from the Onset Detection for the Impulse Response Test Signal with Added Noise. (top: W-channel from the generated impulse response test signal, middle: onset detection with the complex domain approach, bottom: onset detection with the HFC approach)

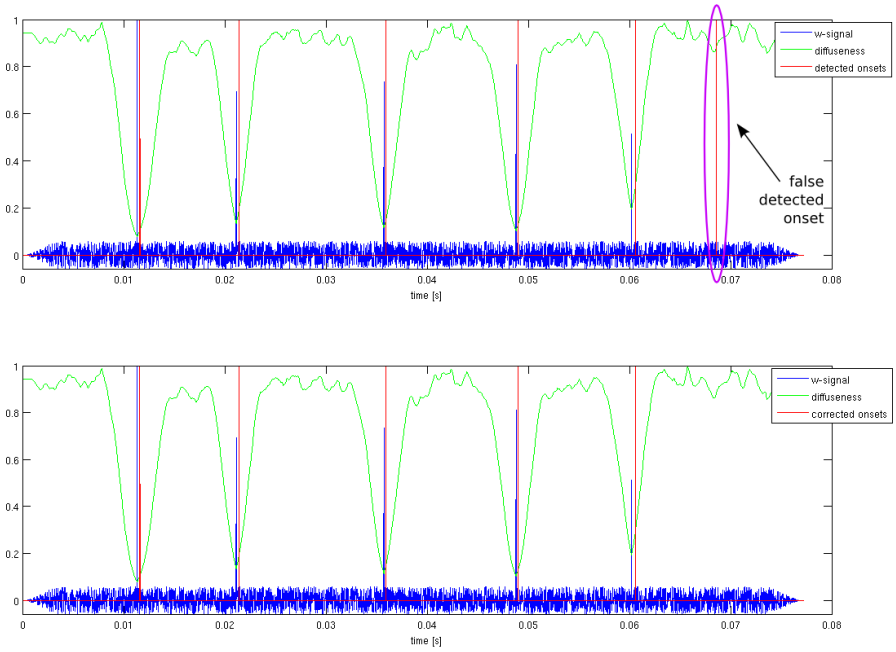


Figure 4.24: Onset Correction of False Detected Onset. (top: W-channel (blue) with detected onsets (red), global diffuseness line (green) and false detected onset (magenta), bottom: W-channel (blue) with corrected onsets (red) and global diffuseness line (green))

Once the correct localization of onsets for both the pure and the noisy signal had been achieved, the accuracy of the directional analysis is tested with regard to the signal-to-noise ratio (SNR) in the signal. The SNR is defined as the ratio of the available average signal power P_{signal} to the existing average noise power P_{noise} and is often expressed in decibels [dB].

$$SNR_{dB} = 10 \log_{10} \left(\frac{P_{signal}}{P_{noise}} \right) \tag{4.28}$$

In addition, for further evaluation the results of the deviations in the directional analysis are specified in the solid angle Ω. The solid angle is a geometrical quantity of the three-dimensional space and is defined on a unit sphere as part of an area surface S [55]. The maximum solid angle is equal to the surface of the unit sphere and is 4π. The solid angle, although dimensionless, is given in units of steradian [sr]. The relationship between solid angle Ω and the spherical coordinates elevation φ and azimuth θ is defined as follows

$$\Omega = \iint_{\varphi_1\theta_1}^{\varphi_2\theta_2} \sin\varphi \, d\varphi \, d\theta \tag{4.29}$$

with φ_1 and θ_1 which define the positions of the virtual sound sources and φ_2 and θ_2 which define the angle from the direction detection, which determine a surface element with the corresponding solid angle Ω. The following correlation exists between the solid angle Ω in steradians [sr] and the apex angle α in degrees [°]:

$$\Omega = 2\pi\left(1 - \cos\left(\frac{\alpha}{2}\right)\right) \tag{4.30}$$

For illustration and a better understanding a comparison of some pairs of values between the solid angle in steradian and the apex angle α is given in Table 4.4.

Solid Angle [sr]	0,000239	0,0239	0,0538	0,214	0,478	0,842	1	1,84	3,14	6,28	12,6
Apex Angle [°]	1	10	15	30	45	60	65,5	90	120	180	360

Table 4.4: Comparison of Solid Angle in Steradian [sr] and Apex Angle in Degree [°]

The evaluation is performed in MATLAB using the statistical method of analysis of variance (ANOVA). The ANOVA examines how the variances between different groups of measured values behave compared to the variances within each group [56]. Thus it can be determined whether the mean values of the groups differ significantly. As a hypothesis it is claimed that no differences between groups are observed, so all measured values can be assigned to a single group. At the end of the calculation a statement is obtained with which significance this hypothesis is accepted or rejected. More detailed description for calculating the ANOVA with a simple example is discussed in [56].

Several noisy test signals, as described above, with different SNR were used to verify the analysis direction. For each test signal, the directional analysis was performed in 5 runs. For each run a new noisy test signal was generated, but with the same 5 virtual sources and the same SNR. Thus, 25 values with the measured deviations in the solid angle are obtained for a group. The SNR in the test signal was gradually reduced until

the noise component is so large that no accurate detection of onsets was possible anymore.

The added white noise signal is distributed spherically and contains no information about direction. Thus it has an influence on the direction detection of the transients, because the transients cannot be extracted by sample exactly. The lower the SNR in the signal the more deviations in the directional analysis will occur. Figure 4.25 shows the results of the ANOVA for this test run.

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	0.13585	7	0.01941	4.84	4.7758e-05
Error	0.76938	192	0.00401		
Total	0.90524	199			

Figure 4.25: ANOVA of the SNR Test

The resulting P-value of the ANOVA is approximately 0.005%, thus very small and lies below the significance level of 5%. This indicates that differences between group means are highly significant. The probability that this result meets hypothesis is equal to the P-value. The boxplot in Figure 4.26 shows a graphical representation of the distribution of the results from the SNR test run. The different groups are compared.

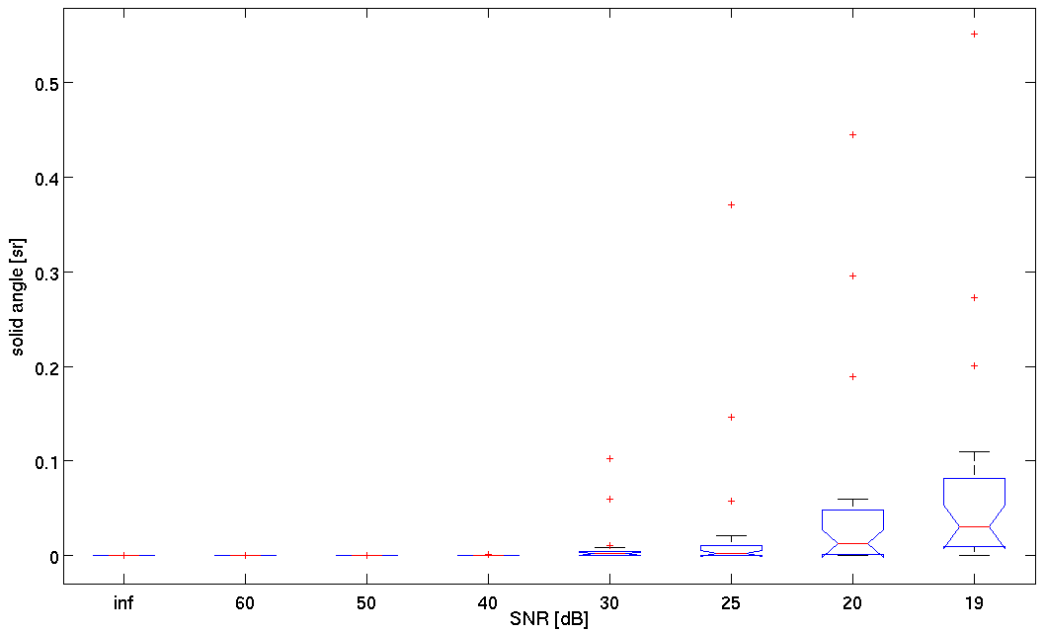


Figure 4.26: Deviations in the Direction Detection in Steradian [sr] in Relation to the SNR in the Test Signal.

The median value as the central value of each group is characterized as a horizontal red line in the boxplot. The interquartile range (IQR) is a measure of the deviation of the measured values from the median value which are marked by horizontal blue lines above and below the median. Half of all measured values are located between these two values. The notches within the IQR show those areas where the true median lies with a probability of 95%. Comparing the boxplots of two groups of values, the overlap of these areas is of great importance. If they do not overlap, the true median values of the two groups are not equal with a probability of at least 95%. The black whiskers outside of the IQR mark areas where other measured values lie and the red crosses mark the outlier. [56]

The first "box" represents the result of the directional analysis from a test signal with no noise, where no deviations occur, and can be supposed as a reference. The tests at a SNR of 60 dB, 50 dB and 40 dB have no significant difference to the reference. The analysis based on the measurements with a SNR of 30 dB, 25 dB and 20 dB have no significant difference among each other. Results with an SNR below 30 dB are significantly different to those with a higher SNR. From the measurement with a SNR of 19 dB a clearly significant difference to the measurement groups above 20 dB exists. With a probability of at least 95% the true median value is not equal to the median value in this groups, thus the hypothesis that no differences are observed between the groups is declined. In the last test with the SNR of 19 dB, the onset algorithm has increasing difficulty to detect onsets such as onsets, due the increased noise.

Evaluation of the direction results from the measured directional room impulse responses

Due to the fact that the positions of the various loudspeakers in the analyzed rooms are known (cf. Table 3.1 and Table 3.2) it is useful to evaluate the directional analysis of the direct sounds from each measured room impulse response. Figure 4.27 and Figure 4.28 show a comparison between the measured and the real loudspeaker positions presented on a spherical surface with top view in the MUMUTH and the CUBE respectively. In the left illustration, the measured directions of the direct sounds are shown and in the right illustration, the real positions of the respective loudspeakers are shown.

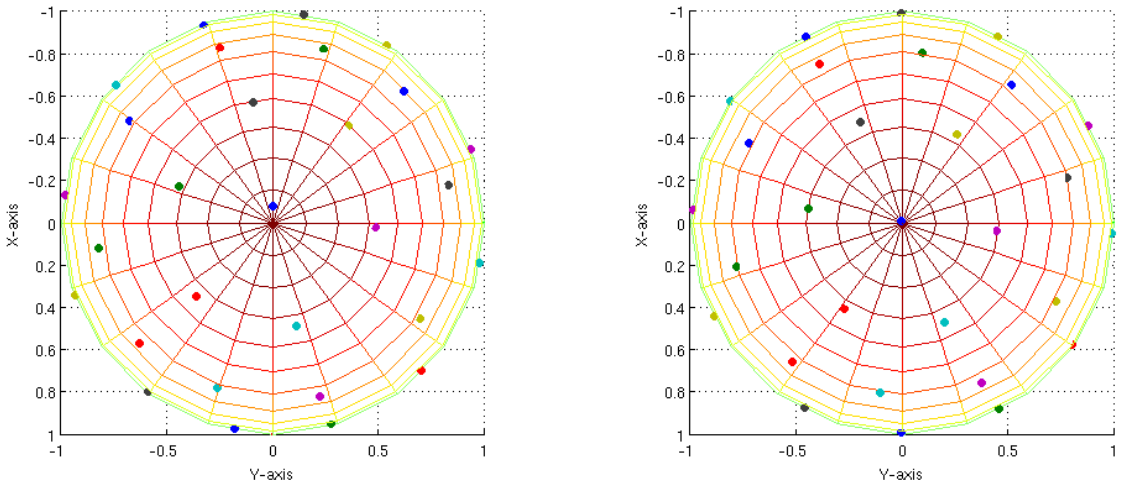


Figure 4.27: Directions from the Direct Sound Mapped on a Spherical Surface with Top View, 29 Loudspeaker, MUMUTH. (left: analyzed direct sound directions from the measured impulse responses, right: real directions)

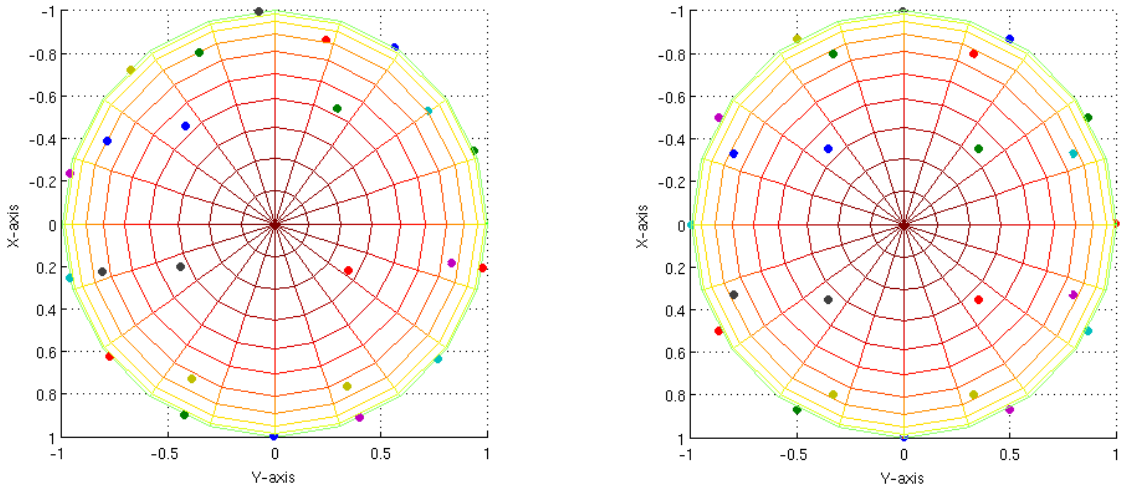
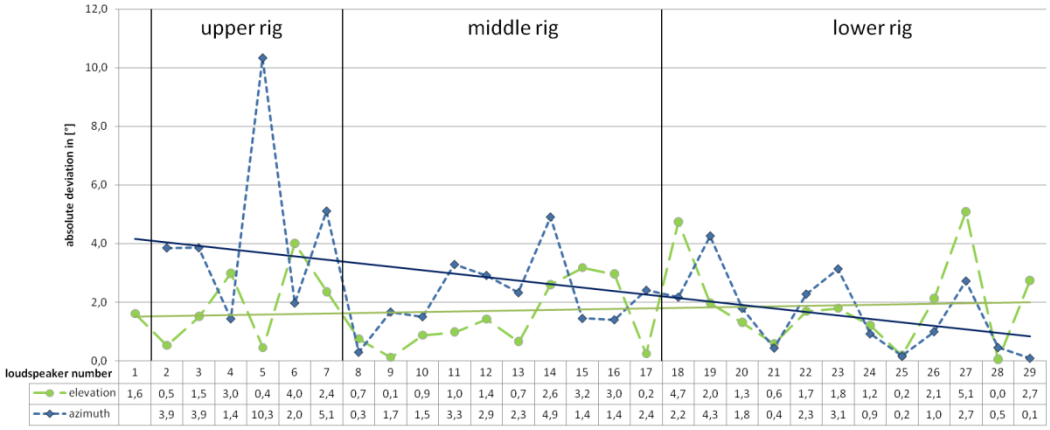


Figure 4.28: Directions from the Direct Sound Mapped on a Spherical Surface with Top View, 24 Loudspeaker, CUBE. (left: analyzed direct sound directions from the measured impulse responses, right: real directions)

	Elevation φ in [°]				Azimuth θ in [°]			
			abs. deviation				abs. deviation	
	µ	**std**	**µ**	**std**	**µ**	**std**	**µ**	**std**
MUMUTH	**2,1**	**2,2**	**1,8**	**1,4**	**8,1**	**3,2**	**2,4**	**2,1**
lower rig @ 8°	1,5	2,5	2,0	1,6	7,8	2,1	1,6	1,3
middle rig @ 32°-36°	2,6	1,8	1,4	1,1	8,3	2,7	2,2	1,3
upper rig @ 59°-63°	2,5	2,6	2,0	1,4	8,5	5,8	4,4	3,2
CUBE	**-1,2**	**3,7**	**3,2**	**1,9**	**-2,6**	**8,0**	**6,2**	**4,9**
lower rig @ 2°	-2,2	2,7	2,7	1,4	-2,5	4,3	3,7	2,1
middle rig @ 30°	-0,6	3,3	2,8	1,4	-2,7	7,6	6,2	3,7
upper rig @ 58°	0,7	6,7	5,6	2,4	-2,4	16,9	13,6	6,1

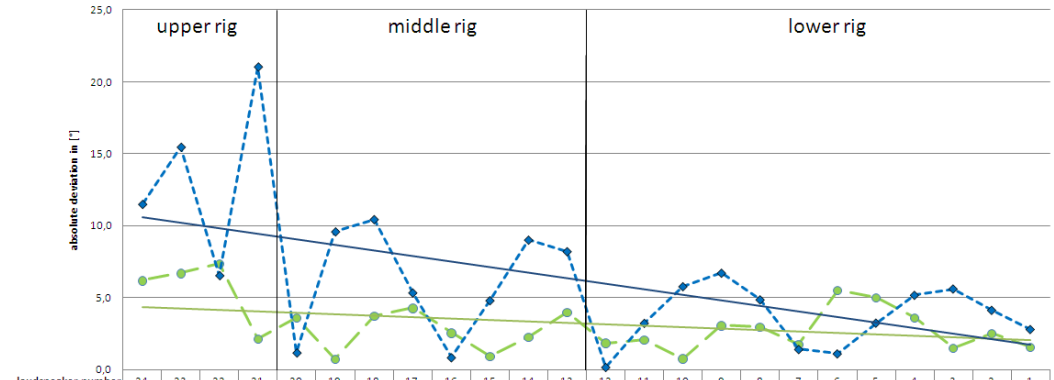
Table 4.5: Deviation and Absolute Deviation Comparing Measured and Real Loudspeaker Positions in the Two Rooms MUMUTH and CUBE. All data are given in degrees [°]. The rigs are specified by elevation.

In Table 4.5 the deviation and the absolute deviation comparing measured and real loudspeaker positions in the two rooms are given. The evaluation of the results was performed for each of the rigs from the Ambisonics systems. The mean and standard deviation of the deviations of the measured data from real data was calculated for each rig. In both rooms the differentiated examination of the rigs shows an increased deviation in determined azimuth with increasing height of the rigs, whereas the absolute deviations in the elevation are more even. Generally the deviation in the azimuth is greater than the deviation in the elevation. Furthermore, the deviation dependencies are similar in both rooms.



loudspeaker number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
elevation	1,6	0,5	1,5	3,0	0,4	4,0	2,4	0,7	0,1	0,9	1,0	1,4	0,7	2,6	3,2	3,0	0,2	4,7	2,0	1,3	0,6	1,7	1,8	1,2	0,2	2,1	5,1	0,0	2,7
azimuth		3,9	3,9	1,4	10,3	2,0	5,1	0,3	1,7	1,5	3,3	2,9	2,3	4,9	1,4	1,4	2,4	2,2	4,3	1,8	0,4	2,3	3,1	0,9	0,2	1,0	2,7	0,5	0,1

Figure 4.29: Absolute Deviation from the Measured Angles for the Loudspeaker Positions in the MUMUTH. (elevation (green dashed line) with trendline (green solid line) and azimuth (blue dashed line) with trendline (blue solid line))



loudspeaker number	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
elevation	6,2	6,7	7,4	2,1	3,6	0,7	3,7	4,3	2,5	0,9	2,3	4,0	1,8	2,1	0,7	3,1	2,9	1,7	5,5	5,0	3,6	1,5	2,5	1,5
azimuth	11,5	15,5	6,5	21,0	1,2	9,6	10,4	5,3	0,8	4,8	9,0	8,2	0,1	3,2	5,8	6,7	4,8	1,4	1,1	3,2	5,2	5,6	4,1	2,8

Figure 4.30: Absolute Deviation from the Measured Angles for the Loudspeaker Positions in the CUBE. (elevation (green dashed line) with trendline (green solid line) and azimuth (blue dashed line) with trendline (blue solid line))

The graphic representations in Figure 4.29 and Figure 4.30 show the absolute deviations of angles for each individual loudspeaker measured. The green dashed line shows the absolute deviations in elevation and the blue dashed line the absolute deviations in azimuth. Based on the trend lines it can be clearly seen that the error of measurement of the azimuth (blue solid line) notably increases towards the higher rigs, during which the error in the calculated elevation (green solid line) remains relatively constant. Note that

for loudspeaker number 1 in the MUMUTH no azimuth angle is required and specified, because this loudspeaker is perpendicular to the microphone and thus an angle in azimuth is irrelevant.

To find out whether there are significant differences in the absolute deviations according to the height of the loudspeakers in different rigs, an ANOVA was applied for both rooms. The results of the ANOVA for the MUMUTH are shown in Figure 4.31.

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	2.1307	2	1.06533	0.53	0.5923
Error	49.7993	25	1.99197		
Total	51.93	27			

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	32.339	2	16.1697	4.78	0.0175
Error	84.565	25	3.3826		
Total	116.904	27			

Figure 4.31: ANOVA of the Absolute Deviations according to the Loudspeaker Rigs, MUMUTH. (left: ANOVA table elevation, right: ANOVA table azimuth)

The ANOVA for the absolute differences in elevation (left) results in a P-value greater than 5%, which means that there are no significant differences in comparison of the various rigs. Whereas the ANOVA for the absolute deviations in azimuth (right) results in a P-value of 1,75%, which is less than 5%. Thus, the differences between rig means are significant.

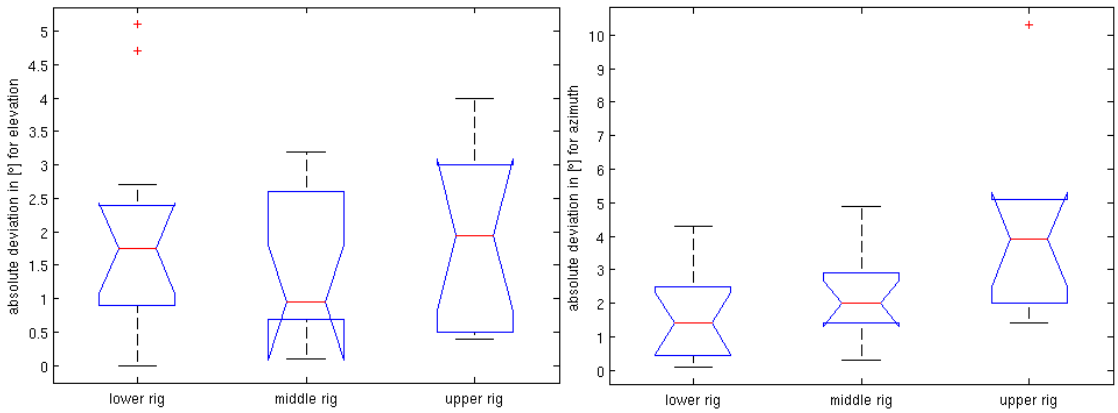


Figure 4.32: Boxplots from the Absolute Deviations of Measured Elevation (left) and Azimuth (right) in Comparison to the Different Loudspeaker Rigs, MUMUTH.

In Figure 4.32 the boxplots for the absolute deviations in the elevation in comparison to the different rigs of the MUMUTH are shown on the left and absolute deviations in the azimuth are shown on the right. As the result from the ANOVA for the elevation pre-

viously illustrated, no significant differences in comparison of the various rigs are shown (left illustration), whereas the absolute deviations in the upper rig for the azimuth value (right illustration) is already significant compared to the lower rig.

In Figure 4.33 the results of the ANOVA for the CUBE are shown.

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	28.1704	2	14.0852	5.5	0.012
Error	53.7492	21	2.5595		
Total	81.9196	23			

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	297.505	2	148.753	12.1	0.0003
Error	258.093	21	12.29		
Total	555.598	23			

Figure 4.33: ANOVA of the Absolute Deviations according to the Loudspeaker Rigs, CUBE. (left: ANOVA table elevation, right: ANOVA table azimuth)

The ANOVA for the absolute differences in elevation (left) results in a P-value of 1,2% and for the absolute differences in azimuth (right) in a P-value of of 0,3%. In both cases, there is significance between the rig means.

The boxplots of the absolute deviations for the CUBE are presented in Figure 4.34. On the left side the absolute deviations in the elevation and on the right in the azimuth.

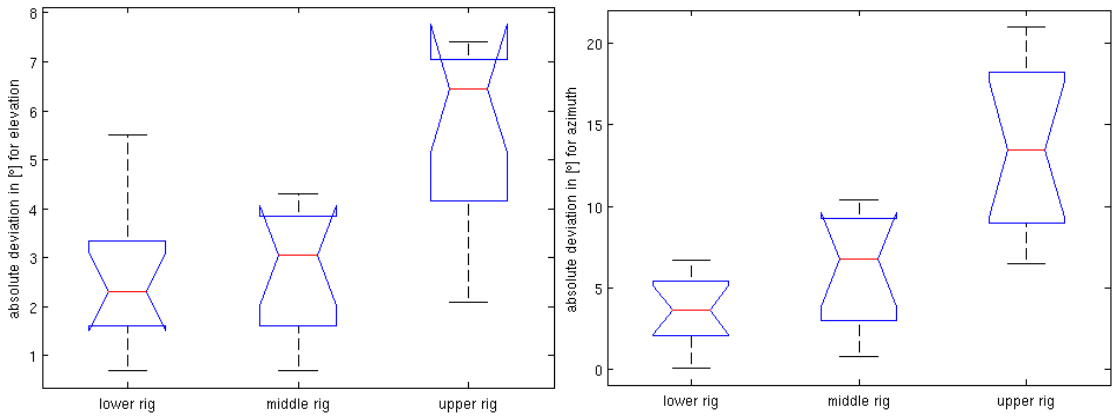


Figure 4.34: Boxplots from the Absolute Deviations of Measured Elevation (left) and Azimuth (right) in Comparison to the Different Loudspeaker Rigs, CUBE.

Immediately noticeable is the significant difference in elevation for the upper rig compared to the other rigs. Looking at the graph in Figure 4.30 and the elevation line in the upper rig, 3 loudspeakers step out of line concerning their deviations. The reason for

this could be that the microphone was not placed exactly right in the sweet spot during the measurement (cf. Figure 4.28). In the case of the azimuth, there are similar trends as previously shown in the results in the MUMUTH. The absolute deviations in the upper rig are notably larger compared to the other rigs. In the case of elevation (left illustration) the difference from the upper rig to the middle and lower rigs is highly significant and in the case of azimuth (right illustration) the upper rig is highly significant to the lower rig. In general, the variations in the measurements in the CUBE are larger than in the MUMUTH.

Subsequently, the errors in the direction detection in dependence of the frequency from the sound field are considered. The non-idealities of the directional patterns of real microphones can lead to errors in the directional analysis. The recording method, with the SoundField microphone, used in this thesis is described in section 2.2. Because the microphone capsules are not located exactly in one position, there is an inaccuracy in measuring the direction of incident sound especially at higher frequencies, where the wavelengths are close to the proportions of the microphone. The impact of higher frequencies on the measurement error is shown in Figure 4.35.

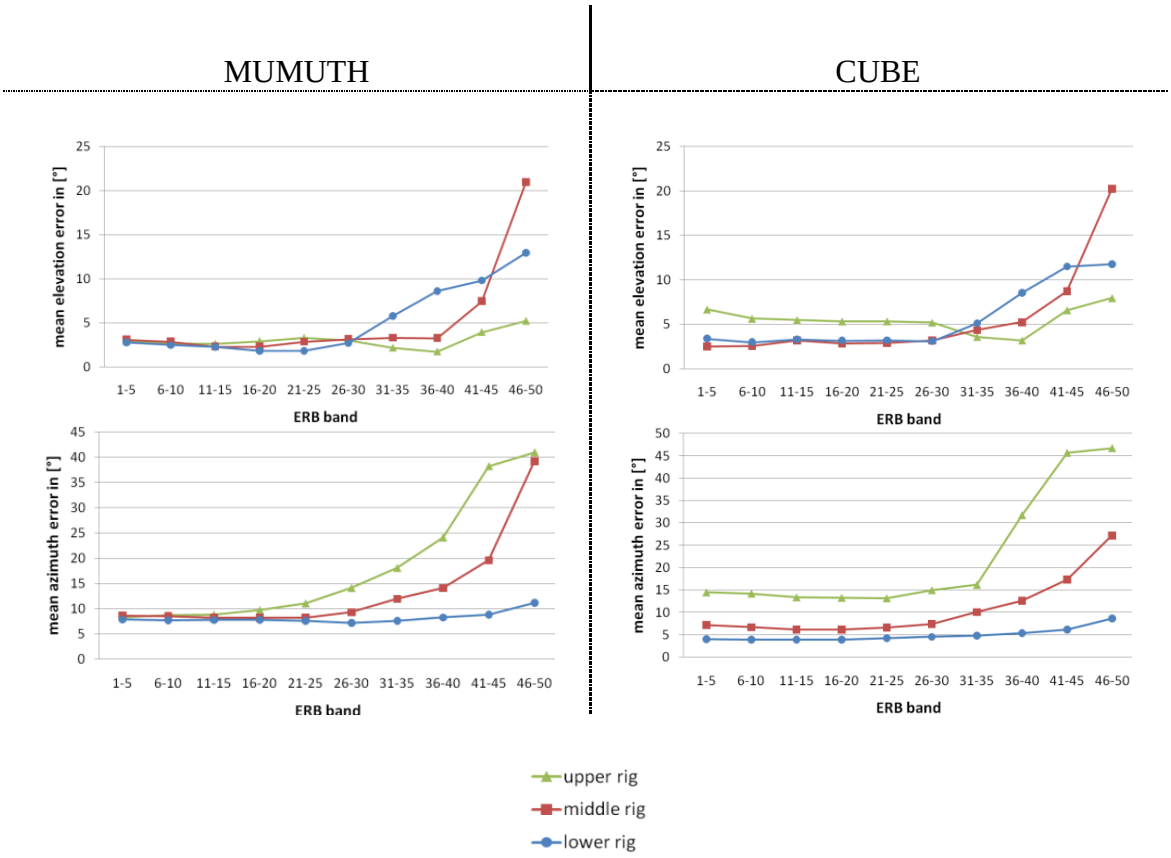


Figure 4.35: Mean Elevation and Azimuth Error from the Direction Measurements according to the Different Analyzed ERB Bands in the MUMUTH and the CUBE.

As mentioned above, the inaccuracy of direction reference increases towards higher frequencies. Figure 4.35 shows the measurement error of the SoundField microphone according to the different rigs and the different frequency bands analyzed. The green line shows the results for the upper rig, the red line for the middle rig and the blue line for the lower rig. On the left side the mean elevation error (top) and the mean azimuth error (bottom) are displayed for the MUMUTH and the same is done on the right side for the CUBE. Averaging is first done band-wise over all loudspeakers in a rig and then averaged in groups of 5 ERB bands each. In general the azimuth error is greater than the elevation error and with regard to the two elevation charts and the two azimuth charts, there is a great similarity between both rooms. Looking at the lower rigs the error of the azimuth remains relatively constant while the error in the elevation increases rapidly towards higher frequencies. The exact opposite occurs in the upper rigs. Here the error in the elevation is relatively constant whereas the error in the azimuth increases quickly with increasing frequency. In the case of the azimuth, the error is always the biggest in the upper rig at high frequencies in both rooms and the error in the elevation is always

at a maximum in the middle rigs at high frequencies. A possible explanation is provided by Batke [57], where he revises B-Format microphones in terms of their accuracy with respect to higher frequencies. The directivity patterns of the B-Format microphone become disturbed if the physical size of the microphone gets similar to half of the wavelength of the incoming sound field [13]. Thus, Batke found out that at higher frequencies the error in the omni-directional W-component is especially large in regions where the capsules of the microphone array have a greater distance from each other and furthermore for the figure-of-eight patterns (X-, Y-, Z-components) the highest error occurs at 45° of incidence [57].

Evaluation of the detected direction in direct sound and reflections and comparison with a room simulation

As a further evaluation and verification of the correctness of the onset detection and the detected direction of the reflections, simulations were performed on the rooms with the acoustics simulation program CATT-Acoustic[8]. A simple shoe box model with corresponding loudspeaker arrangement and the selected measurement position of the microphone has been implemented (cf. Figure 3.4 and Figure 3.6). As an exemplary example the simulation of the impulse response of the loudspeaker number 1 in the CUBE is shown and contrasted with the results from the real impulse response measurement.

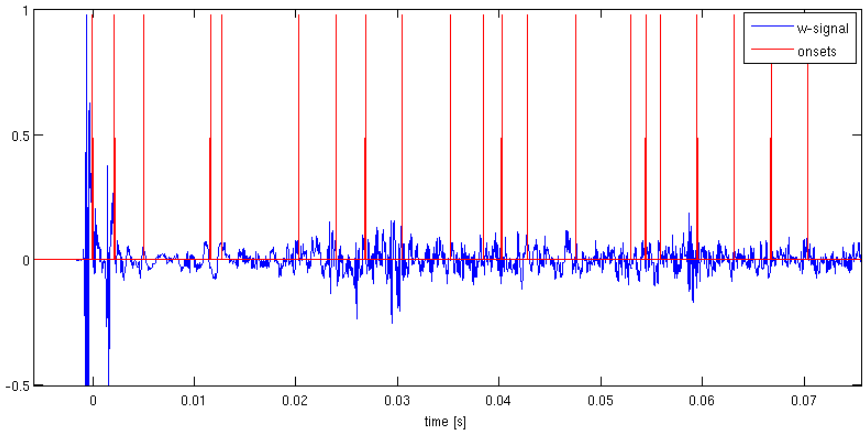


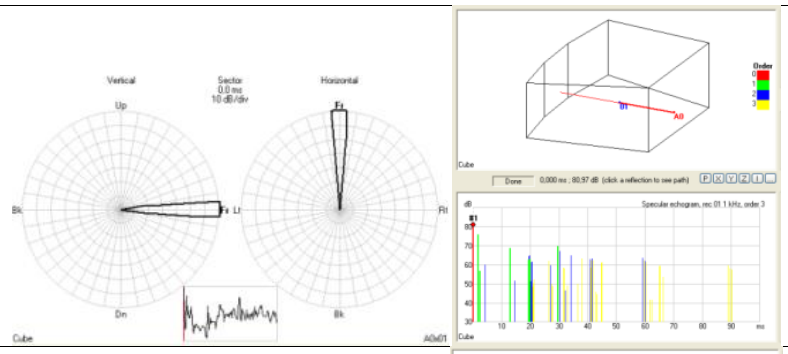
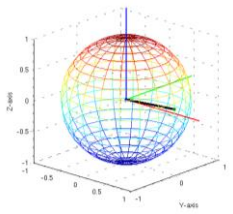
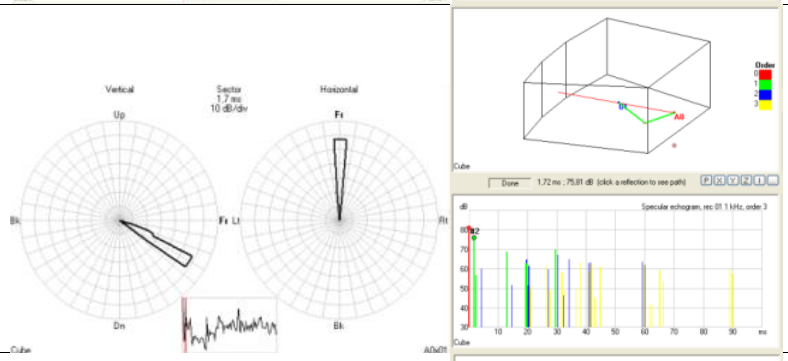
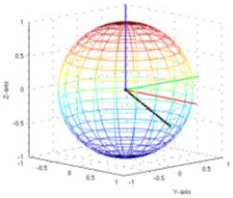
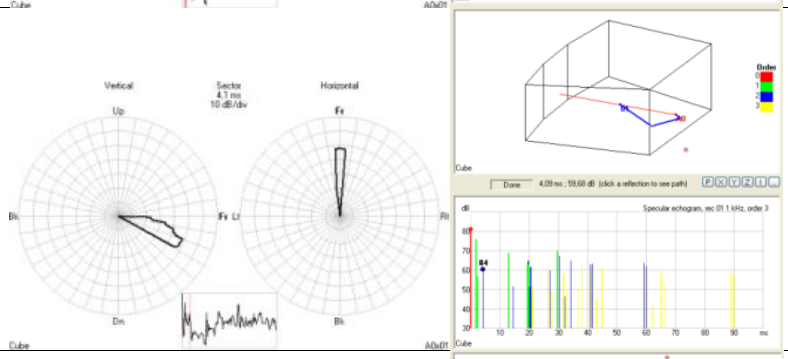
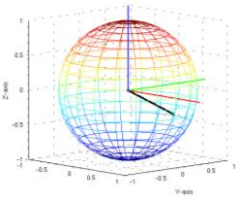
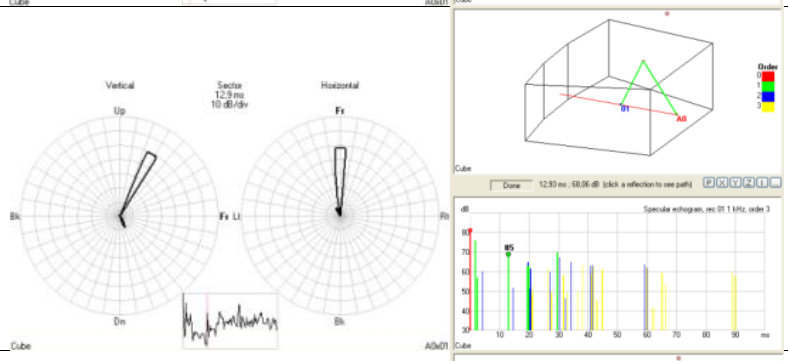
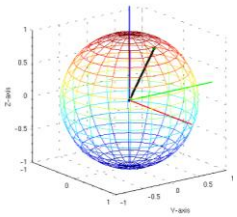
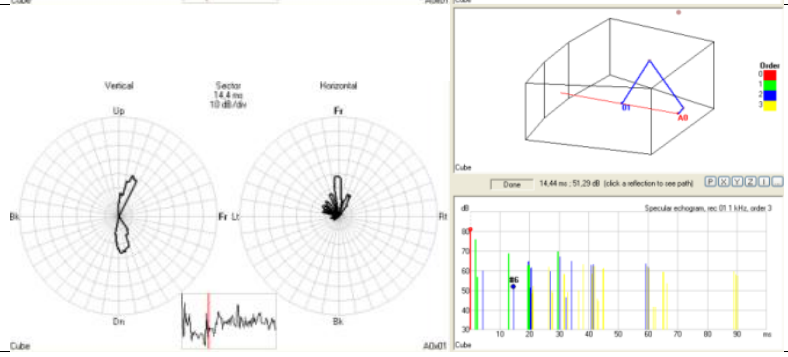
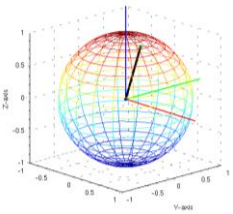
Figure 4.36: W-Channel Impulse Response (blue) from Loudspeaker Number 1 in the CUBE with the First Detected Onsets (red) up to 70 ms.

[8] www.catt.se

Figure 4.36 shows the W-channel of the measured impulse response, including the detected onsets marked in red. In Figure 4.37 a graphical comparison of the simulation and the real measurement from the direct sound until the 9^{th} reflection is shown. The first column shows the ranking of the detected sound events from the real measurement. The second and third column shows the times in milliseconds when the sound events occur, for the simulation and measurement respectively. The first illustration in the fourth column represents the directions from the simulation presented as a polar plot and the second illustration shows the direction of sound in the simulated room plus the simulated impulse response until 100 ms. In the fifth and last column, the direction results of the real measurement are shown. The red, green and blue axes define the three dimensions and the black line is the measured direction of the incident sound at any given time.

Looking at the results until the fourth reflection, it can be stated that both the time when the sound event occurs and the direction incident correspond well. In the special case of the fifth reflection at approximately 20 ms several reflections arrive almost simultaneously at the measuring point. This fact can be a result from the nearly symmetric spatial geometry in relation to the loudspeaker position. In such cases it is not possible for the onset detection algorithm to select all incident directions. In the simulation illustration, the incident direction of the reflection with the highest intensity was selected. For the measured direction and the algorithm this means that the detected sound direction tends to be "pulled" to the incident direction with the greatest intensity, but is also influenced by the simultaneously occurring sound events.

Any differences in time of sound event localization compared to the simulation are caused by the simple design of the room simulation with the shoe box model. Nevertheless, it can be said that aside from the limitation described above, the implemented algorithm performs correctly in both the sound event detection as well as direction detection.

	Sim.	Meas.	Simulation	Measure
DS	0	0		
1. Refl	1,7	2,1		
2. Refl	4,1	5		
3. Refl	12,9	11,5		
4. Refl	14,4	12,6		

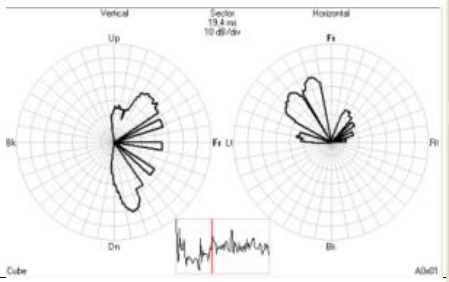
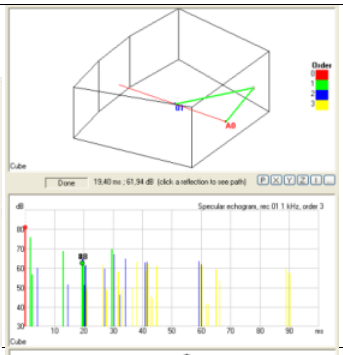
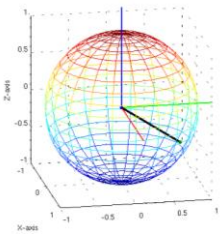
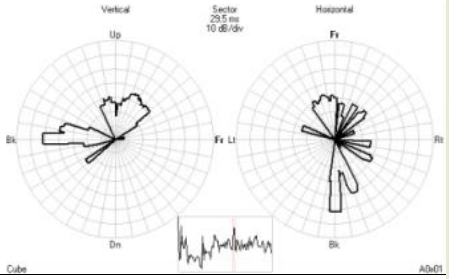
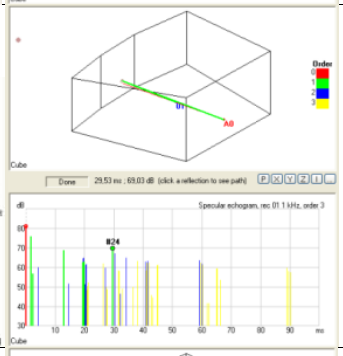
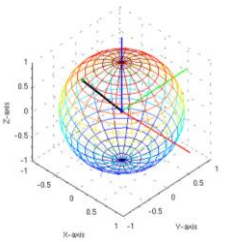
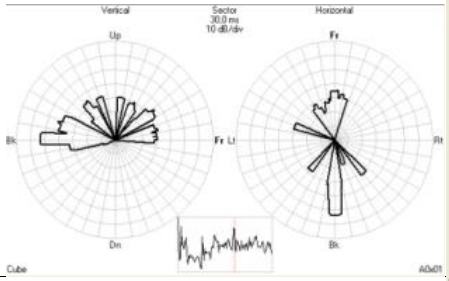
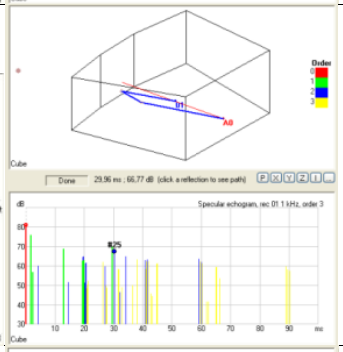
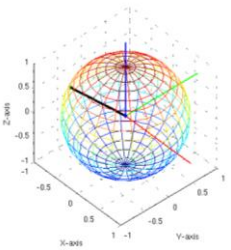
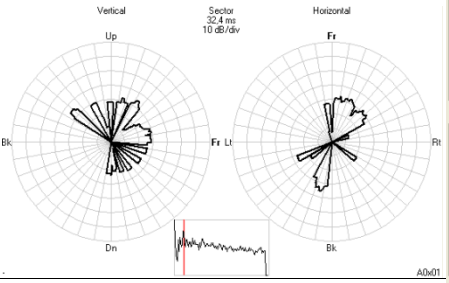
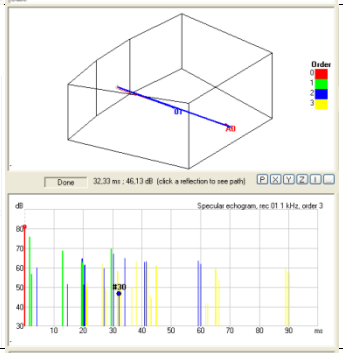
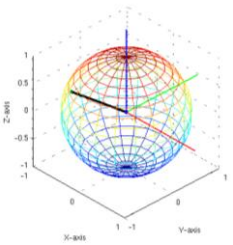
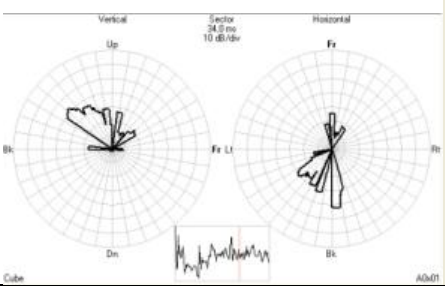
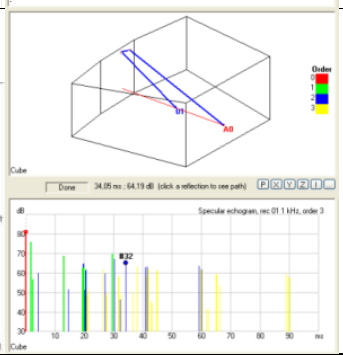
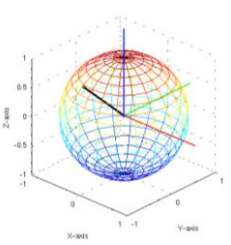
5. Refl	19,4	20,2			
6. Refl	29,5	23,9			
7. Refl	30	26,8			
8. Refl	32,3	30,4			
9. Refl	34	35,1			

Figure 4.37: Graphical Comparison of the Simulation and the Real Measurement of the Detected Directions from the Direct Sound up to the 9th Detected Reflection of the Impulse Response of Loudspeaker Number 1 in the CUBE. (1st column: detected sound events from the measurement (cf. Figure 4.36), 2nd and 3rd column: arise time in [ms] of the sound events, 4th column: results of the simulation, 5th column: results from the measurement)

4.4 Diffuseness Analysis

The transfer of energy in a sound field can be described by intensity vectors as mentioned in section 4.3.3, but also relative to the diffuseness of the sound field. The definition of diffuseness states that the energy density at all points within a perfectly diffuse field is constant. It is also defined that a propagation of sound in all directions is equally probable. Looking at these statements together no net transport of energy is thus implied [35].

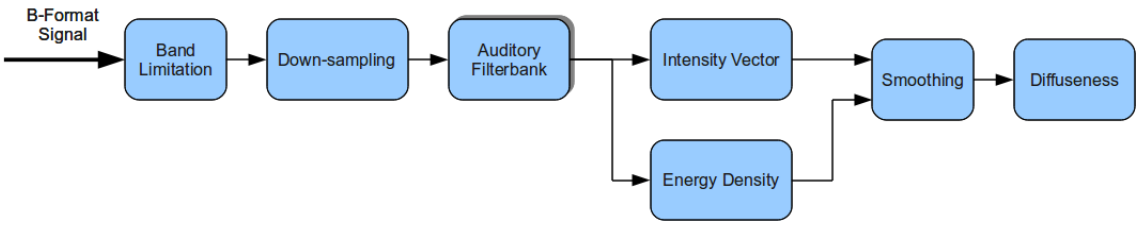


Figure 4.38: Diffuseness Analysis Procedure

The flow chart of the procedure for obtaining the diffuseness is shown in Figure 4.38. Before calculating the diffuseness, the B-Format signals must be pre-processing once again. As with the pre-processing for the directional analysis, the signals are band limited and down-sampled before they are divided into frequency bands by the auditory filterbank (see section 4.3.2). In addition to the calculated and documented intensity vectors (see section 4.3.3), energy density is required to describe the diffuseness at the measuring point.

4.4.1 Instantaneous Energy Density

As mentioned above, the energy density in the measured point of the sound field is needed in addition to the sound intensity in order to calculate the diffuseness. The instantaneous energy density of a general sound field can be computed as

$$E(n) = \frac{1}{2}\rho_0 \left(\frac{p^2(n)}{Z_0^2} + \boldsymbol{u}^2(n) \right) \tag{4.31}$$

The square of a vector denotes the square of the Euclidean norm of the vector. With the shown proportions in eq. (4.19) and eq. (4.20) for B-Format signals of the sound pressure and particle velocity, the estimate for the energy density in the frequency bands is then

$$E(n, f_c) = \frac{1}{2}\rho_0 Z_0^{-2}\left(W^2(n, f_c) + \frac{X^2(n, f_c) + Y^2(n, f_c) + Z^2(n, f_c)}{2}\right) \quad (4.32)$$

The diffuseness ψ is subsequently the ratio between sound intensity and energy density and assumes values between 0 and 1. From eq. (4.18) and eq. (4.31) ψ can be expressed as

$$\psi(n, f_c) = 1 - \frac{\|\langle \boldsymbol{I}(n, f_c)\rangle\|}{c\langle E(n, f_c)\rangle} = 1 - \frac{\sqrt{2}\sqrt{\langle I_x(n, f_c)\rangle^2 + \langle I_y(n, f_c)\rangle^2 + \langle I_z(n, f_c)\rangle^2}}{\langle W^2(n, f_c) + \frac{X^2(n, f_c) + Y^2(n, f_c) + Z^2(n, f_c)}{2}\rangle} \quad (4.33)$$

with the definition for the acoustic impedance Z_0 in eq. (4.21) and where $\|\cdot\|$ denotes the norm of the vector and $\langle\cdot\rangle$ denotes the temporal smoothing operator. The same 1-pole low-pass filter as mentioned in eq. (4.24) is used in this case. The diffuse parameter compares the active intensity of the direct sound in the numerator and the energy density in the denominator.

4.4.2 Diffuseness Results

Diffuseness is analyzed and defined in three different ways. In the global diffuseness from the broadband signal, the global diffuseness in frequency bands and the local diffuseness in frequency bands. The terms global and local are in relation to time.

- *global diffuseness from the broadband signal*

From each measured impulse responses, the global diffuseness is calculated. It is among other things, a criterion for the correction of incorrectly selected onsets (cf. section 4.2.4). Figure 4.39 shows an example of global diffuseness using the W-channel impulse response from loudspeaker number 1 in the MUMUTH. The signal is marked in

blue and the green line represents the global diffuseness. It can be seen clearly that in areas of sound events, the factor of the diffuseness decreases in contrast to regions with less spatial resolution.

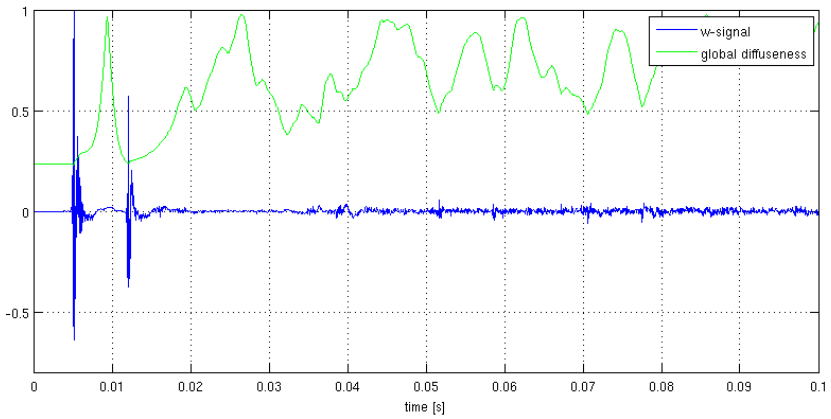


Figure 4.39: Global Diffuseness (green solid line) vs. Evolving Impulse Response (blue) in the Omni-Directed W-Channel Representation of Loudspeaker Number 1, MUMUTH.

- *global diffuseness in frequency bands*

The diffuseness is also identified globally in the individual frequency bands of impulse responses. These are then subsequently used as metadata for the synthesis phase. In the synthesis, they serve as the basis for the allocation of the W-signal in diffuse and non-diffuse streams (cf. section 5.1), as well as for defining the energy in the synthesized Ambisonics channels in the Higher Order Reproduction (cf. section 5.2.2).

- *local diffuseness in frequency bands*

It has been shown, in section 4.3.3, that local diffuseness plays an important role in the analysis of the direct sound direction. Frequency bands are discarded if the level of diffuseness is to high in relation to the direct component of sound when compared to other frequency bands. Such bands are not taken into account in the directional analysis, since they contain less direct information, and thus would worsen the outcome. An illustration of this can be seen in Figure 4.21, where the local diffuseness distribution in the region of the direct sound from the impulse response of two loudspeakers is shown.

An evaluation of the diffuseness in the regions where the direct sound in the signals occurs is shown in Figure 4.40.

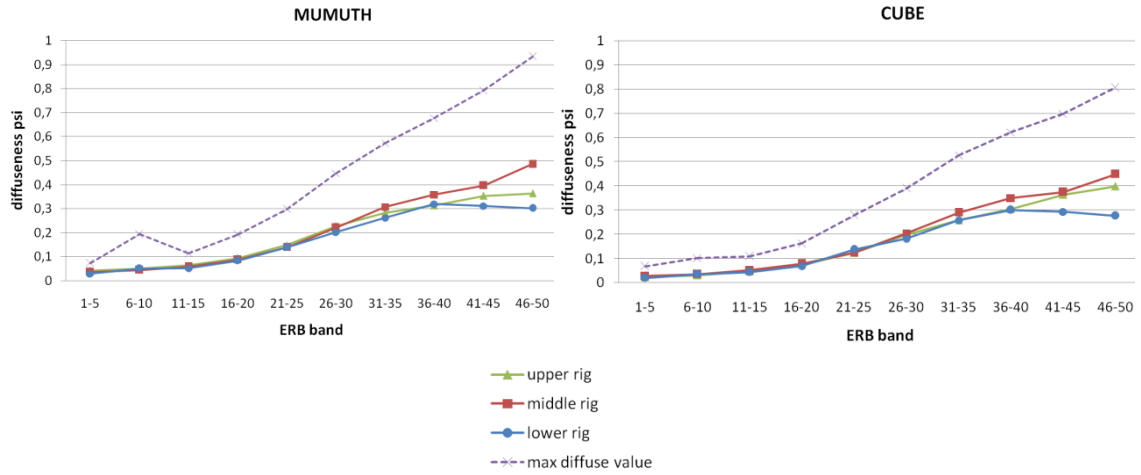


Figure 4.40: Mean and Maximum Diffuseness in ERB Bands from the Region of the Direct Sounds of the Measured Impulse Responses in the MUMUTH (left) and the CUBE (right).

The mean local diffuseness according to the different rigs and the different analyzed frequency bands is displayed for the MUMUTH on the left and for the CUBE on the right. The green solid lines show the results for the upper rig, the red solid lines for the middle rig and the blue solid lines for the lower rig. Additionally, the magenta dashed lines show the maximum diffuse value that occurs in a frequency band in the analyzed region of the direct sound. Averaging is first done band-wise over all loudspeakers in a rig and then averaged in groups of 5 ERB bands respectively. Again, it is distinguishable that there is a great similarity between both rooms. All rigs behave fairly similar and in comparison to the increase in direction errors at higher frequencies, analyzed in Figure 4.35 at section 4.3.4, an increase of diffuseness towards higher frequencies takes place. Especially, the highest diffuseness was measured in the higher ERB bands of the middle rigs. The higher the frequency the greater the influence of the diffuseness.

5 Synthesis Phase

This chapter deals with the synthesis phase in which further processing of the extracted metadata from the analysis phase takes place. In this decoding stage the goal is to improve the spatial resolution of the measured impulse responses performed for the MUMUTH and the CUBE in order to reproduce and/or transfer the natural room acoustics. The main part of the synthesis generally consists of processing the diffuse part and the non-diffuse part of the recorded impulse responses. With the aid of the analyzed parameters azimuth θ and elevation φ at the related instants from the onset detection and the given undirected W-signal, any arbitrary spherical harmonics order signal representation can be determined. A general overview of the executed steps in this synthesis phase is shown in the flow chart in Figure 5.1.

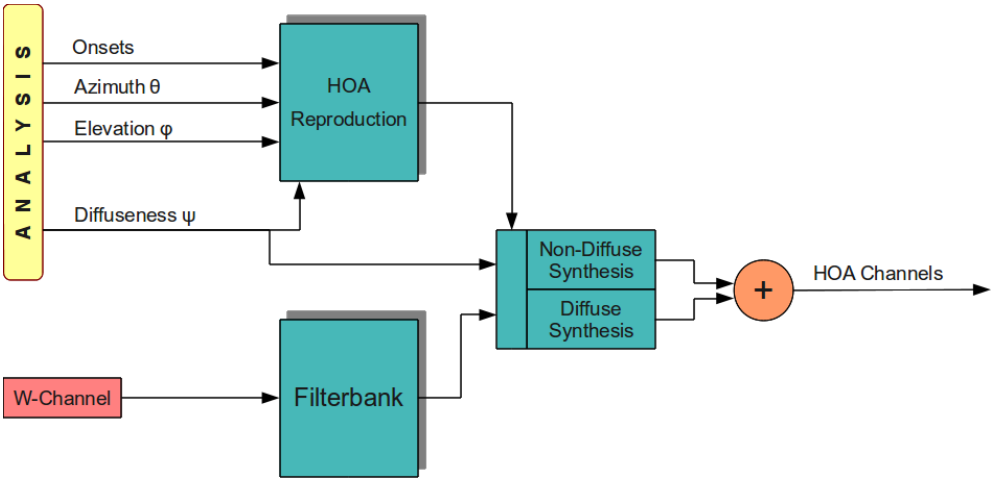


Figure 5.1: Flow Chart Synthesis Phase

Before a separate processing of diffuse and non-diffuse parts can be performed, a few adjustments to the metadata derived from the analysis phase are necessary. These concern especially the processing of the recorded omni-directional W-signal, the length of the impulse response[9] and thus the adaptation of the diffuseness data on the length of

[9] In the analysis phase, only the early parts of the impulse responses, where the relevant information concerning incoming sound events contain, are considered. In the synthesis phase, the existing portion of reverberation must be considered and parameterized.

the synthesized signals. These specifications are achieved in a pre-processing step in section 5.1. Thereafter, with the aid of the diffuseness data from the analysis, the measured impulse responses are divided into the diffuse streams and the non-diffuse streams. The synthesis of the non-diffuse streams is discussed in section 5.2 with the aim to obtain an improved spatial representation of the impulse responses by using Higher Order Ambisonics decoding. Section 5.3 deals with the synthesis of the diffuse streams and focuses on the decorrelation technique applied. The decorrelated diffuse streams have to be encoded to the desired spherical harmonics order and added to the non-diffuse streams to obtain the (Higher Order) Ambisonics channels. Finally the approach of the room acoustics transmission from the larger room the MUMUTH to the smaller room the CUBE is explained in section 5.3.2.

5.1 Pre-Processing

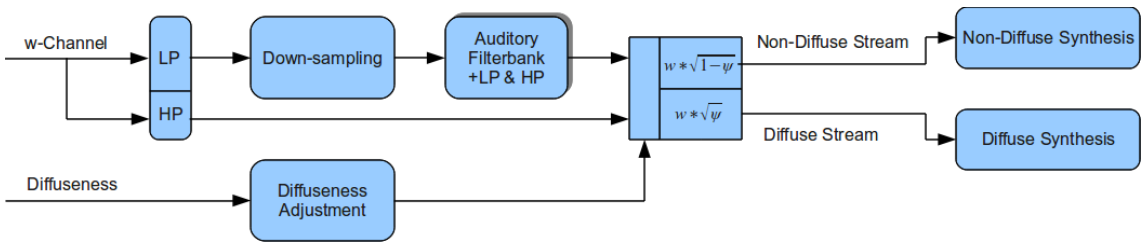


Figure 5.2: Pre-Processing for Synthesis Phase

Figure 5.2 shows the pre-processing stage for the synthesis phase. At the beginning the relevant W-channels of the measured impulse responses are divided by frequency. For this the signals are, in each case, filtered with a cut-off frequency of 4 kHz low-pass and high-pass once again. The low-pass filtered signals are down-sampled from 44100 Hz sampling rate to 11025 Hz sampling rate and divided with the auditory filter bank, as presented in section 4.3.2, into 50 frequency bands. Since this filterbank only consists of band-pass filters from 100 Hz to 4 kHz, the signal is additionally filtered with a low-pass filter with a cut-off frequency of 100 Hz and with a high-pass filter with the cut-off frequency of 4 kHz. This is because in the combination of the frequency bands those portions of the signal shall not be lost at a further stage. Thus, the result with the additional high-pass channel, which is directed parallel and contains information for all

spectral components up to 4 kHz. All in all, next to the 50 band-pass channels, 3 additional channels are carried and processed.

As described in the introduction to chapter 5, the W-signals for the synthesis are selected to be longer to consider the entire portion of the reverberation. A signal length of 2 seconds was chosen for this purpose. Thus, it is ensured that the entire reverberation time is taken into account. Due to this fact the calculated global diffuseness from the analysis must be adjusted, because global diffuseness information of only the first 200 ms were calculated. For this, the diffuseness channels are aligned to the length of the W-signals and the extended part is evaluated beginning from 10 ms after the last detected onset with the value 1, a maximum proportion of diffuseness. Additionally, diffuseness information is needed for the extra channels from the filtering described above. There are two possibilities to get the diffuseness information, either it will be calculated separately as in the analysis in section 4.4, or the diffuseness information of the first band-pass is used as an approximation for the low-pass channels and the diffuseness information of the last band-pass is used as an approximation for the high-pass channels. In this thesis the additional diffuseness information was calculated separately.

At least in this stage the W-channels are divided into a diffuse part by multiplying the signals by $\sqrt{\psi}$ and into a non-diffuse part by $\sqrt{1-\psi}$ where ψ is the diffuseness information from the analysis [39]. Figure 5.3 illustrate the two weighting functions for extracting the non-diffuse and diffuse parts from the signals.

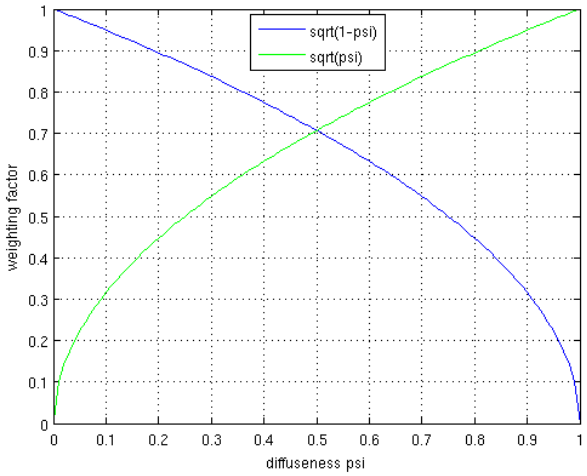


Figure 5.3: Non-Diffuse (blue) and Diffuse (green) Weighting Functions

To obtain the total energy, cross-fading is applied. At those periods where no transient occurs, a certain higher degree of diffuseness appears (cf. Figure 4.39) and the spatial resolution decreases. In case more diffuseness is present, more energy is feeding towards the diffuse stream and less energy is feeding towards the non-diffuse stream and vice versa. For further processing these two parts are synthesized in different ways.

5.2 Non-Diffuse Synthesis

This section gives an insight into the synthesis of the non-diffuse part of the spatial impulse response reproduction. Figure 5.4 shows the flow chart of the single steps implemented in this phase.

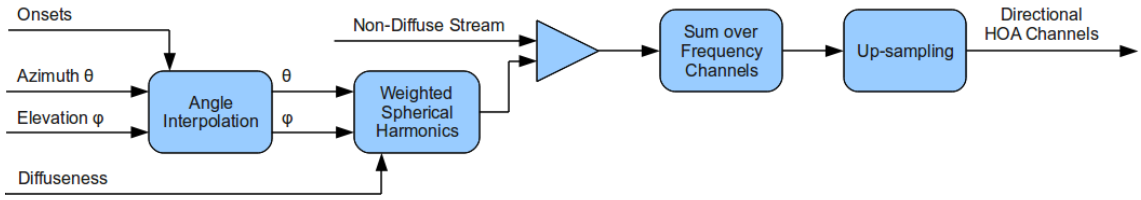


Figure 5.4: Non-Diffuse Synthesis

From the analysis of the impulse responses, the directional information of incident sound events to the respective onset times were calculated. Using this information, the non-diffuse signals should be encoded in higher order Ambisonics. Smooth transitions between the different sound events concerning the changes of directions should be guaranteed. Thus, an interpolation in time, of the changes in elevation angle and azimuth angle is carried out. The procedure for the angle interpolation is described in section 5.2.1. The higher spatial order decoding and the calculation of the spherical harmonics is presented in section 5.2.2. Furthermore, an approach is established in which the various spherical harmonics orders channels are additionally weighted depending on the obtained diffuseness information. Therefore, an improvement in the spatial resolution is targeted. Finally, to get the directional Ambisonics channels, the non-diffuse streams must be weighted with the calculated spherical harmonics channels.

5.2.1 Angle Interpolation

The direction information for each detected direct sound and reflection in the impulse responses has been elicited in the analysis. The information on the elevation and azimuth for an incoming sound event is defined for the area until the next incoming sound event occurs, and so on. The changes in direction of the incident sounds must not take place abruptly but must be interpolated to allow for a smooth transition to be achieved. For this purpose the approach of the shortest path between two points on a spherical surface is chosen. The shortest distance between two points on a spherical surface is described by a large circle (Great Circle Distance) [58]. Figure 5.5 illustrates this approach.

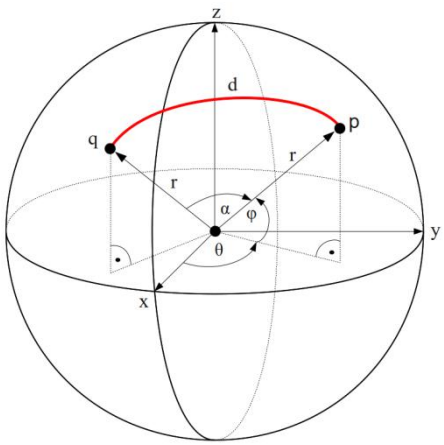


Figure 5.5: Interpolation along the Shortest Distance (Great Circle) between Two Direction Points (*q*, *p*) on a Spherical Surface. (cf. [58])

In this example, point q is the direction from which the sound entry event arrives and point p is the incidence direction of the subsequent sound event. The interpolation from point q to point p now occurs along the intermediate portions of the great circle (Orthodrome) which is, as already mentioned, the shortest distance d between the two points (red line).

The great circle distance is the portion of the circumference described by the included apex angle α hence

$$d = r\alpha. \tag{5.1}$$

From the Law of Cosines

$$\cos\alpha = \frac{\boldsymbol{q}\cdot\boldsymbol{p}}{\|\boldsymbol{q}\|\|\boldsymbol{p}\|} \tag{5.2}$$

the apex angle α (eq. (5.4)) can be calculated with

$$\cos\alpha = \frac{(r\cos\varphi_q\cos\theta_q)(r\cos\varphi_p\cos\theta_p) + (r\cos\varphi_q\sin\theta_q)(r\cos\varphi_p\sin\theta_p) + (r\sin\varphi_q)(r\sin\varphi_p)}{r^2} \tag{5.3}$$

$$\alpha = \arccos[(\cos\varphi_q\cos\theta_q)(\cos\varphi_p\cos\theta_p) + (\cos\varphi_q\sin\theta_q)(\cos\varphi_p\sin\theta_p) + (\sin\varphi_q)(\sin\varphi_p)] \tag{5.4}$$

To calculate the interpolation on the Orthodrome the given spherical coordinates are first converted to Cartesian coordinates according to eq. (2.1). The representation of the points in Cartesian coordinates is given in eq. (5.5).

$$\begin{aligned}\boldsymbol{q} &= [x_q, y_q, z_q]^T \\ \boldsymbol{p} &= [x_p, y_p, z_p]^T\end{aligned} \tag{5.5}$$

By calculating the vector cross product of the two points

$$\boldsymbol{n} = \boldsymbol{q}\times\boldsymbol{p} = \begin{pmatrix} q_y p_z - q_z p_y \\ q_z p_x - q_x p_z \\ q_x p_y - q_y p_x \end{pmatrix} \tag{5.6}$$

and after clamping of the orthogonal basis with

$$\begin{aligned}\boldsymbol{u} &= \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} \\ \boldsymbol{v} &= \frac{\boldsymbol{n}\times\boldsymbol{q}}{\|\boldsymbol{n}\times\boldsymbol{q}\|}\end{aligned} \tag{5.7}$$

where $\|\cdot\|$ denotes the norm of the vector, by further insertion into the circle equation in eq. (5.8)

$$\boldsymbol{X}_m = \boldsymbol{M} + r\boldsymbol{u}\cos m + r\boldsymbol{v}\sin m\,, \qquad 0 < m \leq d \tag{5.8}$$

the Cartesian coordinates of the interpolation steps on the Orthodrome are obtained. $\boldsymbol{M}$ represents the center of the sphere with $\boldsymbol{M} = [0\ 0\ 0]^T$, m the interpolation steps between 0 and the great circle distance d. $\boldsymbol{X}_m$ are the resulting Cartesian coordinates for the interpolation steps.

A reverse transformation into spherical coordinates is carried out according to eq. (2.2). Thus, a smooth transition of the angle azimuth and elevation before every onset and each new change of direction of the incident sound is achieved.

5.2.2 Higher Spatial Order Reproduction

The goal is indeed to decode the existing recordings of directional room impulse responses into higher order Ambisonics. From the channels for the azimuth and elevation angles the nth order normalized spherical harmonics Y_{nm}^{σ} (cf. Table 2.3) are now evaluated for each impulse response and each frequency band according to eq. (2.7). Similar to the case of diffuse adjustment (see section 5.1), the obtained spherical harmonics channels (except for the W-channel with omni-directional representation) are set to zero 10 ms after the last onset. Thus, there is no direction information from the spherical harmonics channels of higher order and only the diffuse component is present. Also the spherical harmonics information of the first band-pass channel for additional low-pass channel and the last band-pass filter for additional high-pass channels are taken here.

As an additional criterion a weighting of the spherical harmonics channels in different orders is applied according to the diffuseness parameters in each channels. The idea is to feed more energy into the spherical harmonics of low order channels representative towards pure spatial resolution in periods where no transients occur and thus the diffuseness percentage is higher. Therefore, higher orders of the spherical harmonics representation fade out earlier, whereby, their energy is recovered towards the lower order channel to retain the overall energy.

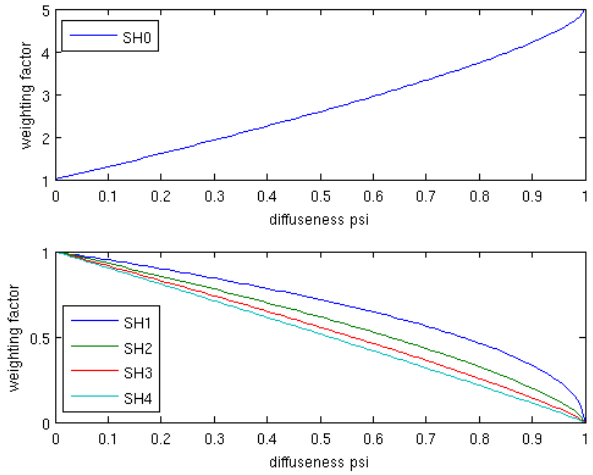


Figure 5.6: Weighting Functions for Spherical Harmonics (SH) up to Order $n = 4$ according to the Diffuseness (top: weighting function for spherical harmonics channels with the order $n = 0$, the omni-directional W-channels, bottom: weighting functions for spherical harmonics channels with order $n = 1$ to $n = 4$)

Figure 5.6 shows the implemented weighting functions for the spherical harmonics of order $n = 4$. The illustration below shows the weighting functions for the channels of order $n = 1$ to $n = 4$. The greater the diffuseness, the lower the spatial resolution and the smaller the weighting factor for these channels is. But therefore, more energy is plugged in the omni-directional W-channels, as the top illustration shows. The total energy remains the same and is only distributed according to the diffuseness influence.

The mathematical principles behind these weighting functions are defined in the following equations. For the weighting functions of the spherical harmonics of the order $n = 1$ to $n = 4$

$$SH_J = (1 - \psi)^{\log\left(\frac{1}{k_{SH_J}}\right)} \tag{5.9}$$

applies, where ψ are the diffuseness parameters, J stands for the chosen spherical harmonics orders n with $1 \leq J \leq n$ and the factor k_{SH_J} is defined in eq. (5.10)

$$k_{SH_J} = \frac{1}{2} \int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} {P_{J0}}^2 \cos\varphi \, d\varphi \tag{5.10}$$

with the associated Legendre functions P_{nm} (see Table 2.2) with order $n = J$ and mode $m = 0$. For the zero-order spherical harmonics channels, the omni-directional W-channels, the weighting function consists of the sum of the difference of the spherical harmonics of higher order from eq. (5.9) to the maximum weighting factor of 1.

$$SH_0 = 1 + \sum_{J=1}^{n} \left(1 - SH_J\right) \tag{5.11}$$

In order to obtain the spherical harmonics representation of the impulse responses, these weighted spherical harmonics functions are multiplied with each of the extracted non-diffuse streams (cf. eq. (2.11)).

Finally, the individual frequency groups of Ambisonics channels are added up and combined. This is followed by an up-sampling to come back from 11025 Hz sampling rate to the original sampling rate of 44100 Hz. Last but not least, the high-pass channel (cf. Figure 5.2) is added and thereby information on the entire frequency spectrum is gained.

5.3 Diffuse Synthesis

The synthesis of the diffuse streams is independent of the synthesis of the non-diffuse streams and is discussed in this section. The diffuse part of a sound has no stable and clear direction. It is supposed to surround the listener when it is synthesized and reproduced from multiple loudspeakers.

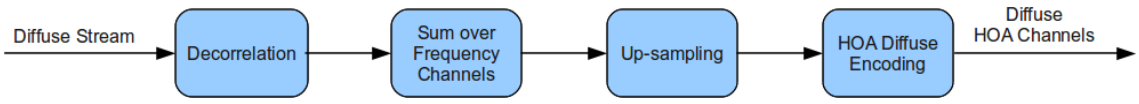


Figure 5.7: Diffuse Synthesis

Figure 5.7 shows the process and the steps of the diffuse synthesis which mainly deals with the topic of decorrelation (see section 5.3.1) hence by deriving decorrelated versions of the omni-directional W-signal. After decorrelation, the encoding of the diffuse streams to the desired higher order Ambisonics is mentioned in section 5.3.2.

5.3.1 Decorrelation

One way to decorrelate signals is to convolve the signals with a temporally diffuse impulse (TDI) which is generated by a noise burst [59]. It is important that the noise should have a magnitude response of unity but a random phase response. In the time domain, the noise burst can have constant amplitude or it may decay exponentially. The length of the noise bursts may not be too long, because it makes impulse-like sounds perceptually longer, but must not be too short, otherwise low frequencies are not sufficiently decorrelated. Decorrelated signals have no influence on the directional perception of non-diffuse sound because diffuse sound should come from all directions and should envelop the listener.

The decorrelation method applied in this thesis is based on frequency dependent exponentially decaying noise bursts [59]. Depending on the frequency band that shall be decorrelated the length of the noise burst is varied. Due to the larger number of oscillations in higher frequency bands, the persistence of the noise bursts can be shorter than at lower frequency bands. For the lowest frequency band with a center frequency of 100 Hz, a noise burst length of 50 ms is chosen, while for the highest frequency band with a center frequency of 4 kHz, a noise burst length of 5 ms is considered adequate. For the frequency bands between, the noise burst lengths are adjusted individually with a matching curve (see Figure 5.8). For each frequency band of all impulse responses that have to be decorrelated, a separate noise burst is generated to avoid accidental correlation.

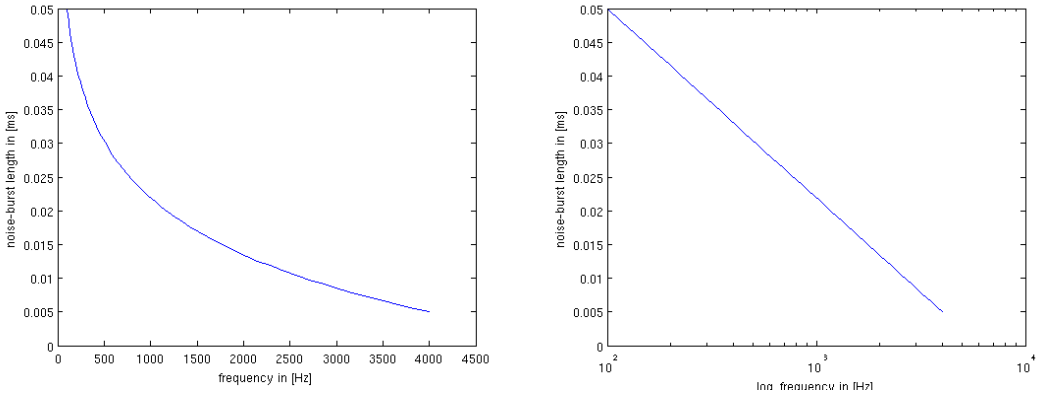


Figure 5.8: Matching Curve to find the Adequate Noise-Burst Lengths for the Different Frequency Bands. (left: linear representation, right: logarithmic representation)

For creating the TDI's each noise burst is modulated with a frequency band dependent exponentially decaying window to thereby obtain exponentially decaying noise bursts. The complex spectrum of the noise burst sequence is normalized by the minimum phase equalization with the help of the minimum-phase Fourier transformation of the burst. The technique of minimum phase equalization represents an all-pass response where the resulting phase response is the excess phase response of the burst. In MATLAB the minimum-phase impulse response can be derived from the impulse response of a burst sequence with the aid of the Hilbert transform of the log magnitude response [60] (see eq. (7.12)).

$$Burst_{min} = e^{conj\left(hilbert\left(\log\left(abs(fft(burst))\right)\right)\right)} \tag{5.12}$$

Thereafter, the TDI is calculated from the inverse Fourier transform of the Fourier transform of the burst, normalized by eq. (5.12).

$$TDI = ifft\left(\frac{fft(burst)}{Burst_{min}}\right) \tag{5.13}$$

Figure 5.9 shows two examples of a TDI assigned for a low frequency band (at the top on the left side) and for a high frequency band (at the bottom on the left side). In the illustrations on the right side of the figure it is shown that the desired magnitude res-

ponses remain unchanged over the frequency range and only the phase responses randomize.

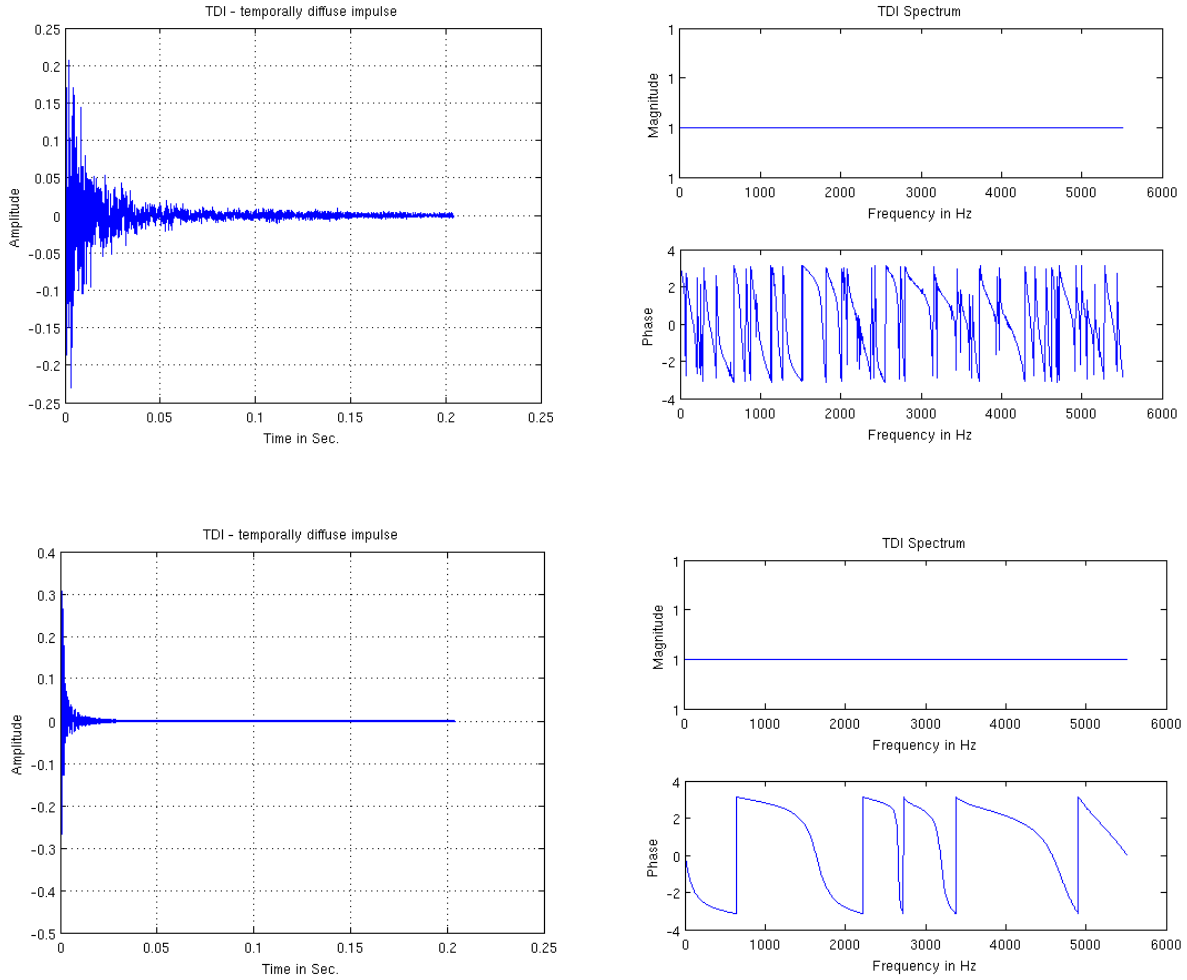


Figure 5.9: Temporally Diffuse Impulse (top left: TDI with an exponential decaying noise burst from 50 ms and minimum-phase equalization for decorrelation of the lowest frequency band, top right: magnitude and phase response from the TDI at the top left, bottom left: TDI with an exponential decaying noise burst from 5 ms and minimum-phase equalization for decorrelation of the highest frequency band, bottom right: magnitude and phase response from the TDI at the bottom left)

Thereafter, the signal decorrelation can be achieved by subsequent convolving of the obtained TDI filters with the above calculated diffuse-streams.

As in the non-diffuse synthesis the individual groups of the decorrelated frequency channels are summed up, followed by an up-sampling to come back from 11025 Hz sampling rate to 44100 Hz sampling rate.

5.3.2 Diffuse Higher Spatial Order Encoding

In a final phase of the diffuse synthesis the decorrelated W-channels will be encoded to the desired spherical harmonics order (see section 5.2.2). Because the diffuse-streams are weighted according to the diffuseness, it is possible to encode the whole W-signal into the spherical harmonics for the respective known loudspeaker positions (cf. Table 3.1 and Table 3.2) with eq. (2.7).

To obtain the final Ambisonics channels, the synthesized diffuse channels are added to the existing non-diffuse channels. As an example Figure 5.10 shows the first 200 ms of the resulting synthesized Ambisonics channels from loudspeaker number 1 in the MU-MUTH, up to order $n = 2$.

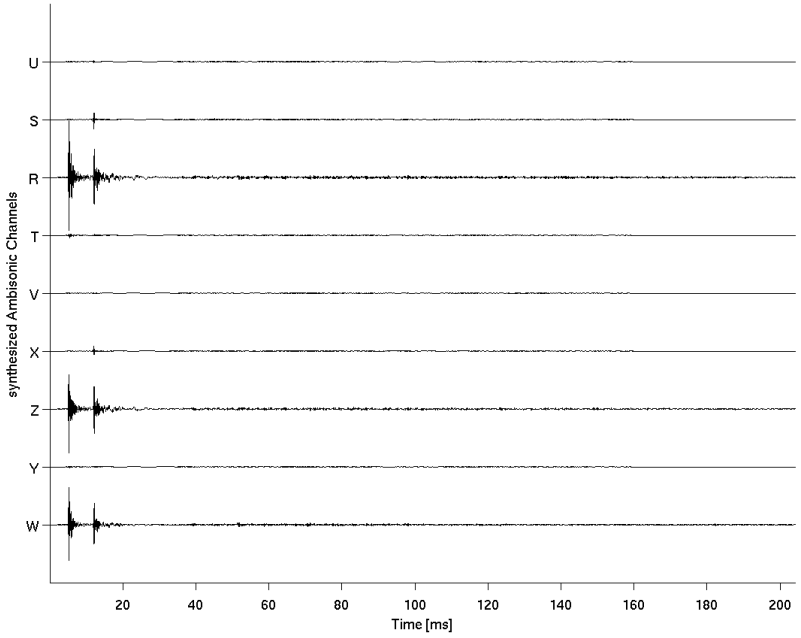


Figure 5.10: Synthesized Ambisonic Channels up to Order $n = 2$ from Loudspeaker Number 1, MUMUTH.

With regard to the loudspeaker position above the microphone during the measurement (cf. section 3.4), it can be expected that the direct sound arrives from above and therefore the first reflection arrives from the bottom. According to that, most energy is found in the Ambisonic channels with mode m = 0 (cf. Figure 2.4), with their representation in the Z direction. These are the omni-directional W-channel, the Z-channel, the R-channel and the corresponding channels in the higher orders.

5.4 Room Acoustics Transfer Approach

This section introduces a possible approach for the transfer of the acoustics of the larger room (MUMUTH) to the smaller room (CUBE), based on the author's publication [25], with the aim to design a decoder which has filter functions instead of scalar weights.

The following equations show the mathematical way to obtain a corresponding decoder. All representatives with the subscript 0 refer to the room where the acoustic properties are captured and the representatives with the subscript 1 refer to the room where these properties are transferred to. The representatives used in the equations are frequency dependent variables, whereby the argument is omitted in order to simplify the description. Furthermore, the vectors used are time-varying but the matrices remain invariant.

$$\mathbf{b}_0 = \mathbf{H}_0 \cdot \mathbf{x} \tag{5.14}$$

On the left side of eq. (5.14) the resulting spherical harmonics description of the overall room response at the microphone position is stated. Matrix $\mathbf{H}_0$ is defined as

$$\mathbf{H}_0 = \begin{bmatrix} H_1^{11} & \dots & H_l^{11} & \dots & H_L^{11} \\ \vdots & \ddots & & & \vdots \\ H_1^{nm} & & H_l^{nm} & & \\ \vdots & & & \ddots & \\ H_1^{NN} & \dots & & & H_L^{NN} \end{bmatrix} \tag{5.15}$$

and covers all the transfer functions from each loudspeaker l across the room to the spherical harmonics, ordered by the super index nm. The microphone array properties are already converted into higher order spherical harmonics by the synthesis as described in the sections above. The vector $\mathbf{x}$ represents the sound pressure excitation distribution over the loudspeaker arrangement. This distribution is obtained by a targeted virtual sound field representation $\mathbf{b}$ in spherical harmonics via the decoding process (decoder matrix $\mathbf{D}_0$) expressed in eq. (5.16).

$$\mathbf{x} = \mathbf{D}_0 \cdot \mathbf{b} \tag{5.16}$$

Combining eq. (5.14) and eq. (5.16), the transformed representation of the target sound field in spherical harmonics is obtained, which is caused primarily by the room transfer functions.

$$\mathbf{b}_0 = \mathbf{H}_0 \cdot \mathbf{D}_0 \cdot \mathbf{b} \tag{5.17}$$

Similar to eq. (5.17), the spherical harmonics representation obtained in the target room can be calculated by applying the targeted sound field representation $\mathbf{b}$.

$$\mathbf{b}_1 = \mathbf{H}_1 \cdot \mathbf{D}_1 \cdot \mathbf{b} \tag{5.18}$$

In order to transfer the acoustic properties from one room to the other, the same spherical harmonics representation at the origin of the rendering system must be calculated as stated in eq. (5.19).

$$\mathbf{b}_1 \equiv \mathbf{b}_0 \tag{5.19}$$

To obtain the expected matching of the two realizations, the modification matrix $\mathbf{M}$ have to be introduced in eq. (5.18) to get the following expression.

$$\mathbf{b}_1 = \mathbf{H}_1 \cdot \mathbf{D}_1 \cdot \mathbf{M} \cdot \mathbf{b} \tag{5.20}$$

By simply rewriting eq. (5.19) using eq. (5.20) and eq. (5.17) and solving for $\mathbf{M}$, the following expression is deduced.

$$\mathbf{M} = (\mathbf{H}_1 \cdot \mathbf{D}_1)^{-1} \cdot \mathbf{H}_0 \cdot \mathbf{D}_0 \tag{5.21}$$

For the acoustic transfer the Matrix $\mathbf{M}$ has to be applied either directly to the sound field representation $\mathbf{b}$ or a frequency dependent decoder has to be obtained. To handle such a mathematical operation as mentioned in eq. (5.21) with the calculation of the inversion of the matrix product, is a crucial and a nontrivial task and should be considered carefully. Furthermore, it should be noted that only controllable reflections can be modified. Upper hemisphere loudspeaker arrangements cannot be used to alter ground reflections [25]. To design an adequate decoder the author refers to [61] where a novel idea of a virtual spherical t-design decoding is presented.

6 Summary, Conclusion and Outlook

Chapter 2 gives an overview of the Ambisonics approach. In order to describe, record and reproduce a sound field in the three-dimensional space at least four channels are needed. For the so called first-order Ambisonics format a minimum of four microphones for recording and a minimum of four loudspeakers for the reproduction of a sound field are required. These channels are the W-channel with its omni-directional signal and the X-, Y-, and Z-channels with a figure-of-eight characteristic to capture direct information of the sound filed according to the three spatial axes. To obtain a better and more accurate resolution of the sound field, an encoding in higher order Ambisonics is possible by decomposing the sound field into a series of spherical harmonics functions. One advantage of Ambisonics is that the encoding process and decoding process are independent. Higher order Ambisonics encoding and decoding is discussed in this chapter.

Chapter 3 deals with the measuring of the directional room impulse responses in the two concert halls which were analyzed, namely the MUMUTH and the CUBE. The chosen impulse response measurement technique with the Exponential Sine Sweep (ESS) method is introduced. The two concert halls with their geometric properties and their installed Ambisonics reproduction system are presented. The room impulse response is measured from each loudspeaker of the Ambisonics system, whereas the loudspeakers are aligned towards the origin of the room where the microphone stands. Additionally coincident measurement to capture the sound field in first-order Ambisonics format is discussed.

Chapter 4 covers the analysis process of the recorded impulse responses. The impulse responses are subjected to a time-frequency analysis and an energy analysis. At first, different onset approaches are realized and compared to capture the incoming sound events, regarding both the direct sound and reflections. Thereafter, the signals are divided into frequency bands and subjected to a frequency-dependent energy analysis. Information about the directions of the sound events and the diffuseness are obtained. Furthermore, an evaluation of the information and results from the analysis is presented in

this chapter. Via code testing procedures and an additional comparison with an implemented room simulation shows that the onset detection provides good results and matches well with the applied simulation. The onset detection algorithm reaches its limits when reflections occur simultaneously or in a temporally very close succession. These onsets cannot be analyzed separately from each other anymore. Furthermore, it is shown that the microphone which is used to record the room impulse responses exhibit inaccuracies in the mapping of directions (cf. [57]). The extent of these inaccuracies depends on both the angle of incidence of sound events and the frequency range defined by the geometric arrangement of the microphone capsules. It is also shown that the diffuseness in higher frequencies has a greater influence than at lower frequencies.

Chapter 5 deals with the further processing and the synthesis procedure of the captured impulse responses with regard to the extracted data from the analysis phase. With the aid of this data and the omni-directional measurement captured, the encoding of the room impulse responses into a higher order Ambisonics representation in order to obtain a better and higher spatial resolution is carried out. The non-diffuse and diffuse parts of the signals are each synthesized differently. Weighting functions are introduced to weight the spherical harmonics of diverse orders in the non-diffuse parts differently. This weighting depends on the existing diffuseness occurrences at any given time, whereas diffuse parts are prepared separately and applied to the Ambisonics channels in a decorrelated form. Finally an approach for the transfer of the captured room acoustic properties from one room to another room is presented.

This thesis showed the analysis of directional room impulse responses, captured with a single B-Format microphone and the spatial re-synthesis into higher order spherical harmonics for a better resolution. Also an approach for the transfer of natural room acoustics is given. Possible future work may include the following tasks:

- Further tests with the implemented onset algorithm would bring further insight to improve the detection and analysis when sound events occur simultaneously or within a short time. If these sound events came from different directions, a frequency-dependent separation should be possible. In [37] an approach to determine positions of sound sources based on DirAC parameters is presented.

- Due to the changing recording quality of the SoundField microphone used, with regard to the sound incidence direction, comparisons to recordings of the room impulse responses with other coincident microphone arrays would be instructive. Furthermore, other recordings of room impulse responses of other concert halls should be made to get a larger repertoire of acoustic parameters of different rooms.

- Further studies in improving an adequate decoder for the transfer of room acoustics should be performed. To obtain an adequate mapping of the room's acoustic properties from one room in another, the reproduction room should be highly damped acoustically. Otherwise its equalization has to be considered, as stated in the approach in section 5.4. Formal listening tests should be arranged.

- More research and testing can be done on to the weighting functions for the different weightings of the non-diffuse streams developed in this thesis (see section 5.2.2). Also further decorrelation methods [59] [62] for the synthesis of the diffuse streams could be analyzed in order to ensure an adequate reproduction of the diffuseness (see section 5.3.1).

- An adaptation of the algorithm for analysis and synthesis of real spatially encoded recordings would be a further step. In contrast to room impulse response recordings, real recordings do not only consist of short and transient events, longer and continuous events occur often. An encoding of real recordings from Ambisonics B-Format into higher order Ambisonics would result in a better and more differentiated sound source representation.

7 References

[1] F. Zotter, H. Pomberger, and M. Frank, "An Alternative Ambisonics Formulation: Modal Source Strength Matching and the Effect of Statial Aliasing," in *AES 126th Convention*, Munich, 2009.

[2] F. Zotter, "Analysis and Synthesis of Sound-Radiation with Spherical Arrays," Ph.D. Thesis, University of Music and Performing Arts Graz, 2009.

[3] V. Pulkki and C. Faller, "Directional Audio Coding: Filterbank and STFT-based Design," in *AES 120th Convention*, Paris, 2006.

[4] M. Kallinger, et al., "Analysis and Adjustment of Planar Microphone Arrays for Application in Directional Audio Coding," in *AES 124th Convention*, Amsterdam, 2008.

[5] J. Vilkamo, T. Lokki, and V. Pulkki, "Directional Audio Coding: Virtual Microphone-Based Synthesis and Subjective Evaluation," *Journal of the AES*, vol. 57, no. 9, pp. 709-723, 2009.

[6] G. Del Galdo, O. Thiergart, F. Kuech, M. Taseska, and D. Sishtla, "Optimized Parameter Estimation in Directional Audio Coding Using Nested Microphone Arrays," in *AES 127th Convention*, New York, 2009.

[7] J. P. Bello, et al., "A Tutorial on Onset Detection in Music Signals," in *IEEE Transactions on Speech and Audio Processing*, 2005.

[8] A. Klapuri, "Sound Onset Detection by Applying Psychoacoustic," in *Proceedings of the IEEE International Conference on Acoustics*, Phoenix, 1999.

[9] P. Masri and A. Bateman, "Improved Modeling of Attack Transients in Music Analysis-Resynthesis," in *Proc. Int. Computer Music Conf. (ICMC)*, 1996, pp. 100-103.

[10] M. Neukom, "Ambisonic Panning," in *AES 123rd Convention*, New York, 2007.

[11] J. Daniel, "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. Thesis, Université Paris, 2000.

[12] M. A. Gerzon, "Periphony: With-Height Sound Reproduction," in *2nd Convention of the Central Europe Section of the Audio Engineering Society*, Munich, 1972.

[13] M. A. Gerzon, "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound," in *50th Audio Engineering Society Conference*, 1975.

[14] P. G. Craven and M. A. Gerzon, "Coincident Microphone Simulation Covering Three Dimensional Space and Yielding Various Directional Outputs ," U.S. Patent 4042779, Aug. 16, 1977.

[15] F. Adriaensen, "ATetrahedral Microphone Processor for Ambisonic Recording," in *LAC2007 - 5th International Linux Audio Conference*, Berlin, 2007.

[16] B. Wiggins, "An Investigation into the Real-time Manipulation and Control of Three-Dimensional Sound Fields," Ph.D. Thesis, University of Derby, 2004.

[17] A. Farina. (2006, Oct.) A-format to B-format Conversion. [Online]. http://pcfarina.eng.unipr.it/Public/B-format/A2B-conversion/A2B.htm

[18] D. G. Malham, "Space in Music - Music in Space," Ph.D. Thesis, University of York, 2003.

[19] M. A. Poletti, "The Design of Encoding Functions for Stereophonic and Polyphonic Sound Systems," *Journal of the AES*, vol. 44, no. 11, pp. 948-963, 1996.

[20] J. M. Zmölnig, "Entwurf und Implementierung einer Mehrkanal-Beschallungsanlage," Master Thesis, University of Music and Performing Arts Graz, 2002.

[21] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. London/San Diego: Academic Press, 1999.

[22] F. Hollerweger, "Periphonic Sound Spatialization in Multi-User Virtual Environments," Master Thesis, University of Music and Performing Arts Graz, 2006.

[23] P. Majdak and M. Noisternig, "Implementation kopfpositionsbezogener Binauraltechnik," Diploma Thesis, University of Music and Performing Arts, Graz, 2002.

[24] A. Sontacchi, "Dreidimensionale Schallfeldreproduktion für Lautsprecher- und Kopfhöreranwendungen," PhD Thesis, University of Technology, Graz, 2003.

[25] K. Hostniker and A. Sontacchi, "Transferable Acoustics Based on Spatial Analysis and Re-Composition," in *9. ITG-Fachtagung*, Bochum, 2010.

[26] M. Barron and A. H. Marshall, "Spatial Impression due to Early Lateral Reflections in Concert Halls: The Derivation of a Physical Measure," *Journal of Sound and Vibration*, vol. 77, no. 2, pp. 211-232, 1981.

[27] L. L. Beranek, "Concert Hall Acoustics," *J. Acoustic Soc. Am.*, vol. 56, no. 7/8, pp. 1-39, 1992.

[28] D. Griesinger, "General Overview of Spatial Impression, Envelopment, Localization, and Externalization," in *Proceedings of the Audio Eng. Soc. 15th Int. AES Conference*, Copenhagen, 1998.

[29] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross, "Objective Measures of Listener Envelopment in Multichannel Surround Systems," *J. Audio Eng. Soc.*, vol. 51, no. 9, pp. 826-840, 2003.

[30] D. Griesinger, "The Psychoacoustics of Listening Area, Depth, and Envelopment in Surround Recordings, and theirRelationship to Microphone Technique," in *Proceedings of the AES 19th International Conference*, Elmau, 2001.

[31] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of Different Impulse Response Measurement Techniques," *J. Audio Eng. Soc.*, vol. 50, no. 4, pp. 249-262, Apr. 2002.

[32] S. Müller and P. Masarani, "Transfer-Function Measurement with Sweeps," *J. Audio Eng. Soc.*, vol. 49, no. 6, Jun. 2001.

[33] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," in *110th AES Convention*, Paris, 2000.

[34] A. Farina, "Advancements in Impulse Response Measurements by Sine Sweeps," in *122nd AES Convention*, Vienna, 2007.

[35] J. Merimaa, "Analysis, Synthesis and Perception of Spatial Sound-Binaural Localization Modeling and Multichannel Loudspeaker Reproduction," Ph.D. Thesis, University of Technology Helsinki, 2006.

[36] R. Schultz-Amling, et al., "Planar Microphone Array Processing for the Analysis and Reproduction of Spatial Audio using Directional Audio Coding," in *AES 124th Convention*, Amsterdam, 2008.

[37] O. Thiergart, R. Schultz-Amling, G. Del Galdo, D. Mahne, and F. Kuech, "Localization of Sound Sources in Reverberant Environments Based on Directional Audio Coding Parameters," in *Proceedings of the 127th AES Convention*, New York, 2009.

[38] J. Zmölnig, W. Ritsch, and A. Sontacchi, "The IEM-CUBE," in *Proceedings of the 2003 International Conference on Auditory Display*, Boston, 2003.

[39] J. Merimaa and V. Pulkki, "Spatial Impulse Response Rendering I: Analysis and Synthesis," *Journal of the AES*, vol. 53, no. 12, pp. 1115-1127, 2005.

[40] V. Pulkki and J. Merimaa, "Spatial Impulse Response Rendering II: Reproduction of Diffuse Sound and Listening Tests," *Journal of the AES*, vol. 54, no. 1/2, pp. 3-20, 2006.

[41] V. Pulkki, "Spatial Sound Reproduction with Directional Audio Coding," *Journal of the AES*, vol. 55, no. 6, pp. 503-516, 2007.

[42] J. Blauert, *Communication Acoustics*, 1st ed. Berlin Heidelberg, Germany: Springer-Verlag, 2005.

[43] P. Masri, "Computer Modelling of Sound for Transformation and Synthesis of Musical Signals," Ph.D. Thesis, University of Bristol, 1996.

[44] J. P. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain," in *IEEE Signal Processing Letters*, vol. 11, London, 2004.

[45] C. Duxbury, J. P. Bello, M. Sandler, and M. Davies, "A Comparison Between Fixed And Multiresolution Analysis For Onset Detection In Musical Signalsc," in *Proc. of the 7th Int. Conference on Digital Audio Effects (DAFx-04)*, Naples, 2004.

[46] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "Complex Domain Onset Detection for Musical Signals," in *Proc. of the 6th Int. Conference on Digital Audio Effects (DAFx-03)*, London, 2003.

[47] I. Kauppinen, "Methods for detecting impulsive noise in speech and audio signals," in *Proceedings of the 14th International Conference on Digital Signal Processing (DSP2002)*, Santorini, 2002.

[48] F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51-83, 1978.

[49] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," Technical Report 35, 1993.

[50] B. R. Glasberg and B. C. J. Moore, "Derivation of Auditory Filter Shapes from Notched-Noise Data," *Hearing Research*, vol. 47, no. 1-2, pp. 103-138, 1990.

[51] R. D. Patterson, et al., "Complex Sounds and Auditory Images," in *Auditory Physiology and Perception, Proc. 9th International Symposium on Hearing*, Oxford, 1992, pp. 429-446.

[52] F. J. Fahy, *Sound Intensity*. Essex: Elsevier Science Publishers Ltd., 1989.

[53] P. Berens, "CircStat: A Matlab Toolbox for Circular Statistics," *Journal of Statistic Software*, vol. 31, no. 10, pp. 1-21, 2009.

[54] S. G. McGovern. (2003) A Model for Room Acoustics. [Online]. http://www.2pi.us/rir.html

[55] Wikipedia. Solid angle. [Online]. http://en.wikipedia.org/wiki/Solid_angle

[56] M. Frank, "Perzeptiver Vergleich von Schallfeldreproduktionsverfahren unterschiedlicher räumlicher Bandbreite," Diploma Thesis, University of Technology, Graz, 2009.

[57] J.-M. Batke, "The B-Format Microphone Revised ," in *Ambisonics Symposium*, Graz, 2009.

[58] D. Bernstein. (2004) Great Circle Distances. [Online]. https://users.cs.jmu.edu/bernstdh/web/common/lectures/summary_great-circle-distance_spherical.php

[59] M. O. J. Hawksford and N. Harris, "Diffuse Signal Processing and Acoustic Source Characterization for Applications in Synthetic Loudspeaker Arrays," in *AES 112th Convention*, Munich, 2002.

[60] M. O. Hawksford, "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design," *J. Audio Eng. Soc.*, vol. 45, no. 1/2, pp. 37-62, Jan. 1997.

[61] F. Zotter, M. Frank, and A. Sontacchi, "The Virtual T-Design Ambisonics-Rig Using VBAP," in *Proceedings of the 1st EAA-EuroRegio*, Ljubljana, 2010.

[62] M. Bouéri and C. Kyirakakis, "Audio Signal Decorrelation Based on a Critical Band Approach," in *AES 117th Convention*, San Francisco, 2004.