A Binaural 3D Sound System Applied to Moving Sources

Master Thesis

Performed and handed in by
Peter Lessing

For:	Institute of Electronic Music and Acoustics (IEM). University of Music and Dramatic Arts of Graz
Board of management:	o.Univ.-Prof.Mag.DI Dr. Robert Höldrich
In cooperation with:	Escuela Universitaria De Ingeniería Técnica (EUITT) de Telecomunicación. Universidad De Las Palmas De Gran Canaria
Surveyor:	o.Univ.-Prof.Mag.DI Dr. Robert Höldrich
Research Instructor:	o.Univ.-Prof.Mag.DI Dr. Robert Höldrich Juan Luis Navarro Mesa (EUITT).

Las Palmas De Gran Canaria, 25.08.2004

Abstract

This master thesis deals with the problem of generating binaural signals for moving sources in closed or in open environments. For this purpose a simulation program composed of three main parts is developed: propagation, reverberation and the effect of the head, torso and pinna. Before the development of each part, all of these effects are taken under consideration and are studied.

For propagation the effect of attenuation due to distance and molecular air-absorption is considered. Particularly in open environments the interaction of sound with the ground and some atmospheric conditions (e.g. fog) are simulated.

Related to the interaction of sounds with the environment, especially in closed environments is reverberation. For these environments an algorithm for estimating the early reflections up to fourth order, which utilize the method of mirror sources, is developed. Late reverberation is implemented by designing digital filters constructed by Schroeder and Jot.

The effects of the head, torso and pinna on signals that arrive at the listener are also objectives of the consideration. The set of Head-Related Transfer Functions (HRTF) that have been used, originates from the KEMAR database. Special attention has been given to the modelling and interpolation of HRTFs for the generation of new transfer functions, which are not at disposal in the database.

Some other aspects like definition of trajectories, definition of closed environment, etc. will also be considered for their inclusion in the program to achieve realistic binaural renderings. The evaluation is implemented in MATLAB.

Zusammenfassung

Diese Diplomarbeit behandelt die Verarbeitung binauraler Signale für Quellen, die sich im geschlossenen und offenen Raum bewegen. Hierzu wird ein Simulationsprogramm bestehend aus drei Teilen entwickelt: Ausbreitung, Hall und dem Effekt des Kopfes, des Torso und der Pinna.

Für die Ausbreitung wird der Effekt der Dämpfung auf Grund von Distanz und molekularer Luftabsorption, implementiert. Speziell im offenen Raum wird die Interaktion des Schalls mit dem Boden und einigen atmospherischen Bedingungen (z.B Nebel) simuliert.

Durch Interaktion von Schall mit der Umgebung, speziell in geschlossenen Räumen, entsteht Hall. In diesem Teil der Arbeit wird ein Algorithmus entworfen, der die ersten Reflexionen bis zur 4ten Ordnung über das Spiegelquellenverfahren berechnet. Der Nachhall wird mit digitalen Filtern von Jot und Schröder implementiert.

Ebenfalls werden die Auswirkungen des Kopfes, des Torso und der Pinna auf die Signale, die beim Hörer eintreffen, behandelt. Verwendet wurden Außenohr-Übertragungsfunktionen (HRTFs) der KEMAR Datenbank. Speziell wird auf die Modellierung und Interpolation von HRTFs eingegangen, um Übertragungsfunktionen zu generieren, die nicht in der Datenbank vorhanden sind.

Andere Aspekte wie die Defintion von Trajektorien, die Definition von geschlossenen Räumen usw. werden in das Programm eingebunden um eine realistische, binaurale Wiedergabe zu gewährleisten. Die Ausarbeitung erfolgt in MATLAB.

To my parents

I wish to thank
Juan Luis Navarro Mesa and
o.Univ.- Prof.Mag.DI Dr. Robert Höldrich
for being in charge of this master thesis.

Table of Contents

Introduction 1

Chapter One
Basics

1.1 Overview of Spatial Hearing

1.1.1 Environmental Context 2
1.1.2 The Human Ear 4
1.1.3 A Model of Natural and Spatial Hearing 6

2.1 Perception of Azimuth and Elevation

1.2.1 The Duplex Theory (IIDs and IDTs) 8
1.2.2 Spectral Cues 11
1.2.3 Movements of Head and Source 14

3.1 Sound Distance and Reverberation

1.3.1 Distance 15
1.3.2 Reverberation 18

Chapter Two
Head Related Transfer Functions (HRTFs)

2.1 Head Related Transfer Functions (HRTF)

2.1.1 What Influences HRTF Curves 20
2.1.2 HRTF Magnitude Characteristics 21
2.1.3 HRTF Phase Characteristics 22

2.2 Localization with HRTF Cues

2.2.1 Spectral Cues 24
2.2.2 Spectral Band Sensitivity and Directional Bands 26

2.3 How to Measure HRTF Sets

2.3.1 Measurement 27
2.3.2 Equalisation 29

2.4 Existing HRTF Material

2.4.1 KEMAR 31
2.4.2 CIPIC 34

Chapter Three
The Acoustical Environment

3.1 Indoor and Outdoor Sound Propagation

3.1.1 Air Absorption 35
3.2.2 Wall Absorption 40
3.1.3 Ground Interaction 42
3.2.4 Early Reflections and the Source Image Model 45
3.2.5 Late Reverberation and it's Implementation 52

Chapter Four
Modelling and Interpolation of HRTFs

4.1 HRTF - Modelling

4.1.1 Common-Acoustical-Pole and Zero Modelling (CAPZ) 64

4.2 HRTF - Interpolation

4.2.1 Vector-Base-Amplitude Panning (VBAP) 77
4.2.2 Common-Acoustical-Pole and Residue Interpolation (CAPR) 82

Chapter Five

The 3D System

5.1 The System

5.1.1 The Signal Routing 86
5.1.2 The Manual 91

References 105

Introduction

Many different sound system technologies are in use. The simplest, a monophonic sound system, is incapable of reproducing the spatial characteristics of sound.

Two canal stereo sound systems are by far superior, enabling the reproduction of sound image that are spatially distributed between two loudspeakers. The capabilities of stereo sound systems can be augmented by adding additional speakers to the sides or rear of the listener. The resulting surround systems are generally able to reproduce sound images anywhere in the horizontal plane surrounding the listener [Gardner, 1998].

A spatial auditory display (also called virtual acoustic display or 3-D audio system) is a system capable of rendering sound images positioned arbitrarily around a listener. There are two approaches to build such a system. The first one is to completely surround the listener with a large number of loudspeakers and the second one is to reproduce only at the ears of the listener the acoustic signal that would occur in the natural listening situation to be simulated. This method is called binaural audio.

During the last decades binaural technology has become enabling technology with a significant impact on various fields, such as information technology, hearing aids and advanced sound measurement techniques.

There are three aspects, which comprise a binaural system: The physical, psychoacoustic and the psychological aspect. The physical aspect is concerned with the way of the source signals before they reach the inner ears. The Psychoacoustic Aspect deals mainly with the signal processing in the subcortical auditory system and the Psychological Aspect focuses on cognitive brain function.

It is clear that today's binaural technology rests mainly on our knowledge of the physical aspect, with an increasing use of psychoacoustics.

This master thesis only deals with the Aspects of physics and psychoacoustics as the subcortical part is not that important for the implementation of the 3-D system [Begault,1994].

Chapter 1
Basics

1.1 Overview of Spatial Hearing

1.1.1 Environmental Context

Direct and Indirect Sound

The acoustic pressure wave expands radially outward, reaching walls and other surfaces where energy is reflected, diffracted and absorbed. Technically speaking, all this reflected energy is reverberation and gives us a spatial impression [3].

Figure 1.1 shows a simple model of how a listener (L) in an enclosure will hear. The direct path, which the listener first hears is called the *direct sound (DS)*, followed by the *early reflections (ER)*, the ones of nearby surfaces. After a few hundred milliseconds, the number of reflected waves becomes very large and because there is equal energy propagation in all directions, these reflections are called *diffuse reverberation* or *late reflections* (LR).

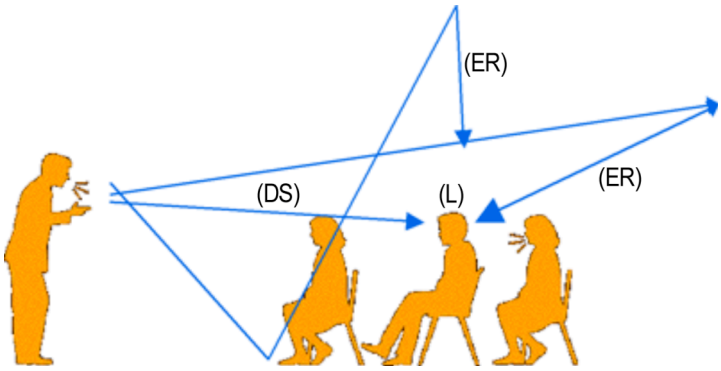


Figure 1.1 Room with reflections [Internet]

These two reflections, the early reflections and the late reflections, define the *indirect sound* and yield information for perception about the sound source and the environmental context.

At the head of the listener, sound pressure waves are, depending on the wavelength of the sound in comparison with the dimension of the head, also diffracted (>1.5 Hz) and reflected (<1.5 Hz). These two phenomena yield information about the source's position in three dimensions. Therefore, in order to specify the location of the source relative to the listener, a coordinate system has to be defined.

Coordinate System

To specify the location of sound source relative to the listener, we need a coordinate system. One natural choice is *the head-centred rectangular – coordinate system* shown in Figure 1.2. Here the 'x' axis goes (approximately) through the right ear, the 'y' axis points straight ahead and the 'z' axis is vertical.

This defines three standard planes: The xy or horizontal plane, the xz or frontal plane and the yz or median plane. Clearly, the horizontal plane defines up-down separation, the frontal plane front-back separation and the median plane right-left separation.

However, due to the fact that the human head is roughly spherical, a spherical coordinate system is used. Unfortunately, there is more than one way to describe these coordinates and different people define them in different ways.

The vertical-polar system shown in Figure 1.3 is the most popular. This system contains two angles and one scalar shown in Figure 1.4.

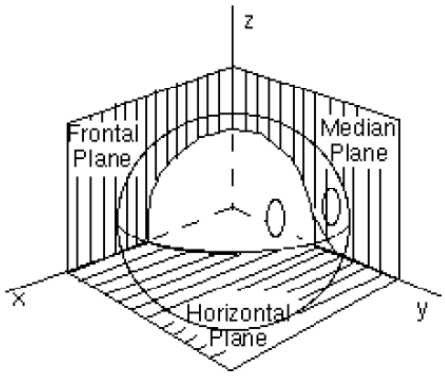


Figure 1.2 Head-centered rectangular-coordinate system [20]

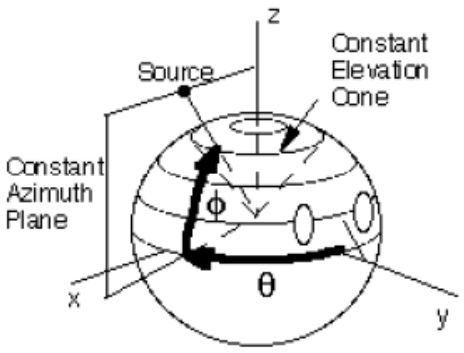


Figure 1.3 Vertical polar coordinate system [20]

The angle in the horizontal plane and goes from 0˚ (exact in front of the listener) to 360˚ is called azimuth (θ) and is measured from the median plane to a vertical plane containing the source and the z axis. The second angle is called elevation (ϕ) and is measured from the horizontal plane to the distance vector. It goes from 0˚ to 90˚ (above the listener) and 0˚ to - 90˚ (below the listener). The scalar defines the distance of the source.

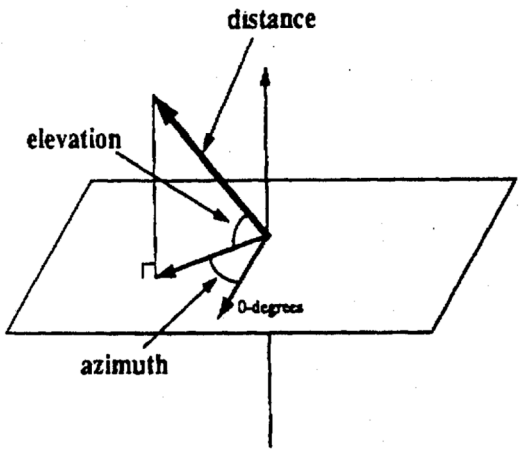


Figure 1.4 Azimuth, elevation and distance in the vertical polar system [1]

1.1.2 The Human Ear

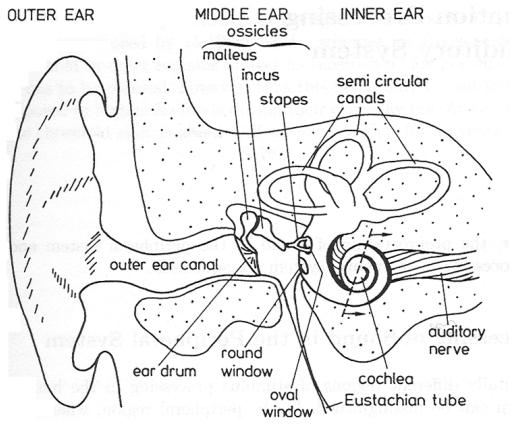


Figure 1.5 The human ear (plus arrows to the different parts) [10]

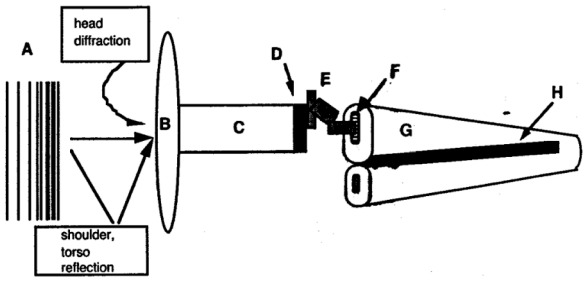


Figure 1.6 Schematic overview of the auditory system [1]

The human ear consists of 3 main sections: The *outer ear*, the *middle ear* and *the inner ear* [1]. The sound is first transformed by the *pinnae*, the visible portion of the outer ear (**B**), and proximate parts of the body like the shoulders and head. Followed by the effects of the *ear channel* (or *meatus* - **C**) that leads to the middle ear, which consists of the *eardrum* (**D**) and the *ossicles*, that are small little bones called *"hammer-anvil-stirrup"* (**E**). There sound is transformed from acoustical level (at the *ear drum*) to a mechanical one. At the *oval window* (**F**) the mechanical energy again is converted into fluid pressure within the *inner ear* (the *cochlea* - **G**). The fluid pressure causes frequency dependent vibration patterns of the *basilar membrane* (**H**) within the *inner ear*, which causes numerous fibers producing from auditory *hair cells*. These activate electrical action potentials within the neurons of the auditory system, which are combined and processed at higher levels with information from the opposite ear.

1.1.3 A Model of Natural & Virtual Spatial Hearing

Begault D. R. [1] mentioned that, from an engineering standpoint, a perceptually meaningful, segmented approach for effecting synthetic binaural hearing can result from understanding natural spatial hearing. The nonlinearities between the various source-medium-receiver (Figure 1.7) stages can be described statistically via psychoacoustic research results related to 3-D sound. Each element of the model, shown in the figure above, contains a number of physical, neurological or perceptual transmissions in the communication of the spatial location of the sound source [1].

In the case of natural spatial hearing (Figure 1.8), the sound is first influenced by the environmental context and then by the auditory system. Other sound sources within the environmental context might be considered undesirable. These are generally classified as *noise* sources. This diffuse field is processed at higher levels of the brain.

Considering the model of virtual spatial hearing, shown in Figure 1.9, the source is the electrical representation of sound. It is transformed in a 3-D audio system and afterwards processed in the same manner, at higher levels.

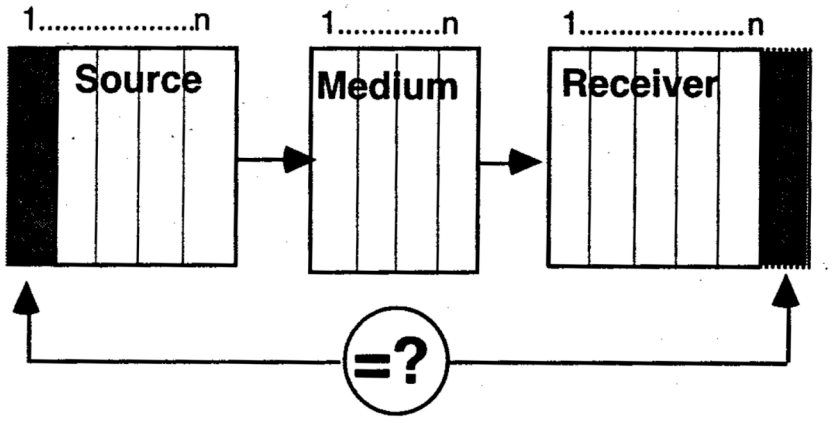


Figure 1.7 A source-medium-receiver model [1]

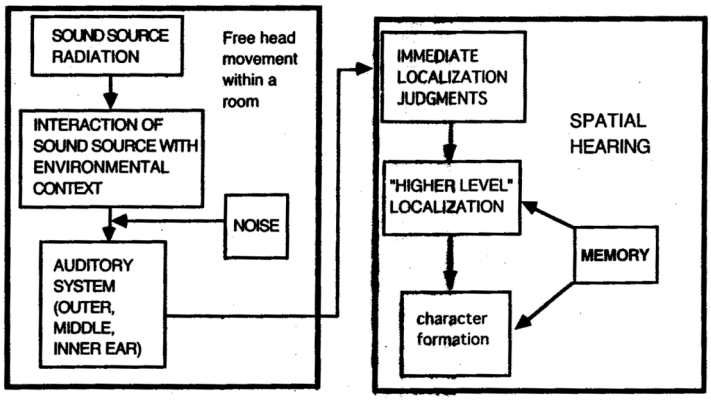


Figure 1.8 A model of natural spatial hearing [2]

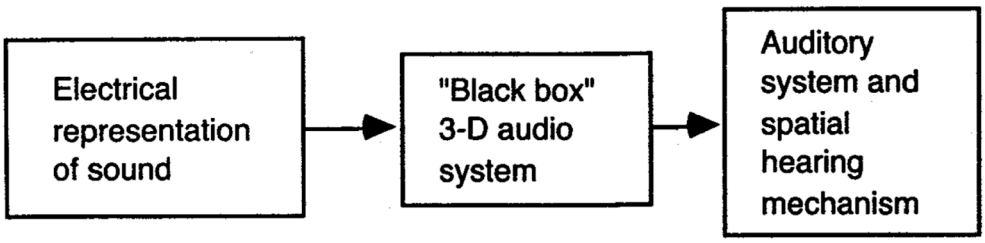


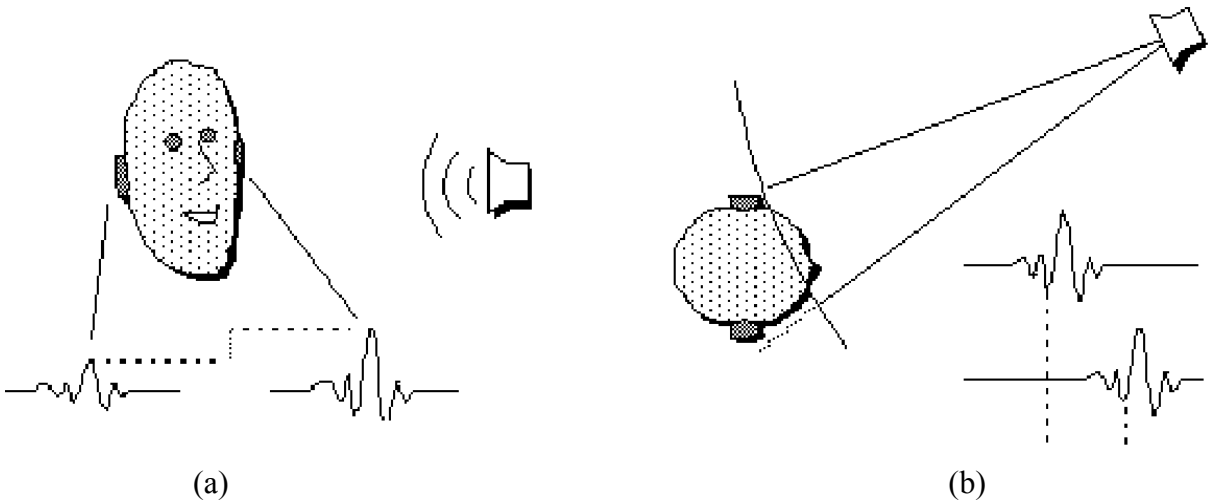
Figure 1.9 A model of virtual spatial hearing [3]

1.2 Perception of Azimuth & Elevation

1.2.1 The Duplex Theory

For 100 years it has been known that the principal cues for sound localisation, particularly the left-right localisation, are time and level differences at the ears of the listener (Rayleigh's "duplex theory", 1907). This theory states that low frequencies are localized using time (phase) cues, and high frequencies are localized using interaural level (intensity) cues, as pointed in Gardner W. G [4].

Interaural level differences (ILDs) or also called interaural intensity differences (IID) are defined as level differences generated between the right and the left ears by the sound. A simple method of deriving ILDs is depicted in Figure 1.10 (a). Rayleigh observed that the incident sound waves are diffracted by the head. As you might expect, ILDs are highly frequency dependent. At low frequencies (below approximately 1 kHz), where the wavelength is long relative to the head diameter, there is hardly any difference in sound pressure at the two ears. However for higher frequencies, where the wavelength is short relative to the head diameter (frequencies greater than about 1.5 kHz), the head acts like an obstacle, creating a "shadow" effect at the contra lateral side. This "shadow" effect increases with increasing frequency (i.e., decreasing size of the wavelength). For instance, a 3 kHz sine wave at 90 degrees will be attenuated by about 10 dB, a 6 kHz wave by about 25 dB and a 10 kHz wave by about 35 dB, [1.5].



(a) (b)

Figure 1.10 A simple method for deriving IID (a) and ITD (b) [Internet]

Interaural time differences (ITDs), shown in figure 1.10 (b), result from the difference in arrival times of sound's wavefront at the right and left ears.

Consider a sound wave from a distant source that strikes a spherical head of the radius r from a direction specified by the azimuth angle θ, shown in Figure 1.11. When the wavefront A-B arrives at the right ear, there is still a path of the length (a+b) to travel before it arrives at the left ear. As a result of the symmetry of the configuration the b section is given by $b = r\sin\theta$. A section represents a proportion of the circumference, subtended by θ. Therefore the corresponding path length (a+b) is

$$(a+b) = \left(\frac{\theta}{360}\right)2\pi r + r\sin\theta \tag{1.1}$$

Divided by the speed of sound c, we obtain the following formula for the interaural time difference

$$IDT = \frac{r}{c}(\theta + \sin\theta), \qquad -90^{\circ} \leq \theta \leq +90^{\circ} \tag{1.2.}$$

When the source is directly ahead, θ tends 0˚, the ITD is zero. When θ tends to 90˚, we have a maximum of the ITD. Suppose the head diameter is 17.5 cm, then the path length is about 22.5 cm. With a sound speed of $c = 330 m/s$, the ITD is about $656\,\mu s$. This represents a difference of arrival time of about 0.7 ms for a typical size human head.

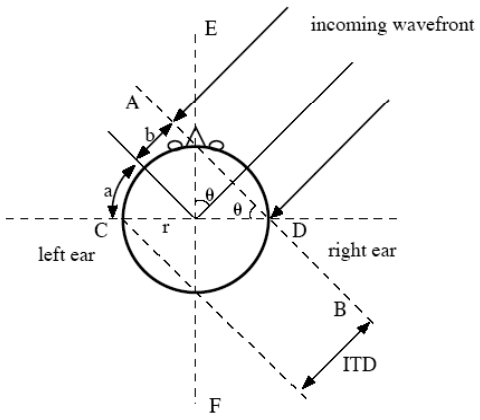


Figure 1.11 ITD [20]

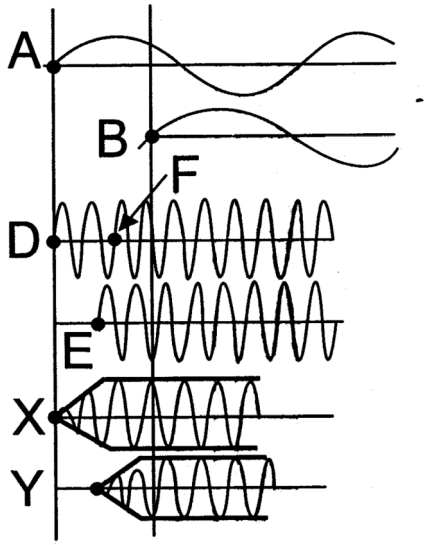


Figure 1.12 Phase locking [1]

Both ITD and IID are frequency dependent and important parameters for the perception of sound location in the azimuth plane (e.g., perception of sound in the left-right direction). In general, a sound is perceived to be closer at the *ipsilateral ear*. In other words, for pure sinusoids, perceived lateral displacement is proportional to the phase difference of the received sound at the two ears. However, at frequencies above a critical point of about 1.5 KHz, sine waves become smaller than the diameter of the head and the phase information in relationship to relative time arrival at the ears can no longer convey which the leading wavefront is. But research has revealed that timing information can be used for higher frequencies because the timing differences in amplitude envelopes are detected [1]. This so called ITD envelope cue is illustrated in figure 1.12. A and B show sine waves with 800 Hz, whose phase information can be detected because the half period of the waves is larger than the size of the head. Signal D and E are waves with a frequency over the critical point. Here you cannot say whether D leads E or E leads F. But if the sine waves are increased and decreased in amplitude (amplitude modulation) then an amplitude envelope is imposed on the sine wave [1]. Shown in figure 1.12 X and Y. The auditory system somehow extracts the overall amplitude envelope of higher frequency components at both ears and measures the difference in time of arrival of the two envelopes.

1.2.2 Spectral Cues

Let us think about an ideal model of the head, where the interaural differences are at a minimum. For a sound originating from any point of the median plane, ITDs and IIDs are zero. However, because listeners can differentiate sound originating from points in this plane, there has to be a monoaural hearing mechanism for the median plane. The spectral coloration of a sound produced by the external ear, or pinna, the torso and the head is a well known effect that provides the primary cues for elevation.

This spectral filtering of a sound source before it reaches the ear drum is called head-related-transfer function (HRTF). In the time domain this function is termed head-related impulse response (HRIR). The binaural HRTF (terminology for referring to both left and right ear HRTFs) can be thought of as a frequency dependent amplitude and time delay difference that result primary from the complex shaping of the pinna. More about HRTFs is mentioned in chapter 2.

Several studies have tested sound localization to measure the influence of the pinna. These studies show that pinna cues are significant for elevation and azimuth localization. Figure 1.13 shows measured frequency responses for two different directions of arrival.

In each case we see a direct path and a longer path following a reflection from the pinna. At low frequencies these two signals arrive in phase and the pinna collects additional sound energy. However at high frequencies the signal is out of phase with the reflected one and destructive interference occurs. The greatest destructive interference occurs when the path length d is a half wavelength:

$$f = \frac{c}{\lambda} \quad \rightarrow \quad f = \frac{c}{2d} \tag{1.3}$$

This frequency produces a “pinna notch“ around 10 kHz. With typical values for d, the notch frequency is usually in the 6 kHz to 16 kHz range.

As the pinna is a more effective reflector for sounds coming from the front of the listener than sounds from above, the resulting notch is more pronounced for sources from the front. In addition, the length of the part changes with the elevation angel and therefore the frequency of the notch moves with elevation.

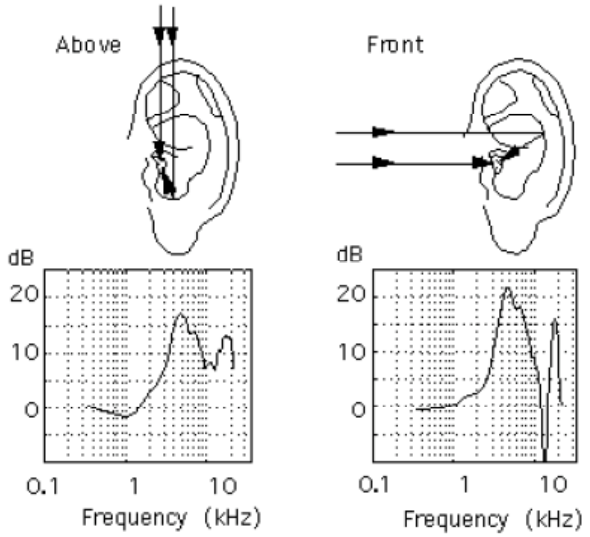


Figure 1.13 Frequency responses of the pinna for two directions of arrival [1]

The torso and the head have a measurable influence on the HRTF. The best known effect is the so called “shoulder bounce“ , the reflection from the upper torso that occurs when the source is above the subject. The reflected wave introduces a series of comb filter notches into the HRTF spectrum [6,7,and 8]. From overhead sources, the first notch appears at frequencies as low as 600 Hz, which is important for sources that do not certain have much high-frequency energy. In general, the longest time delay between the direct pulse and the reflected one occurs when the source is overhead. As the source elevation is reduced, the delay time shrinks, eventually becoming zero when the ray from the source to the ear becomes tangent to the torso. The set of tangent rays defines what we call the “torso-shadow cone“, shown in figure 1.14. There are two different cases depending on whether the source is located inside or outside the torso-shadow cone. When the source is outside of the cone, a direct path and a shoulder reflection is detected. If the source is inside of the torso-shadow, the pressure wave has to diffract around the body to reach the ear. It also has been detected that sources in the ipsilateral zone of the shadow are shadowed by just the torso, while sources in the contra lateral zone are shadowed by both the torso and the head. The model, which describes these effects is called the “Snowman Model“, shown in figure 1.14.

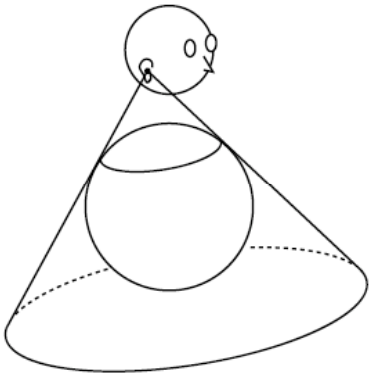


Figure 1.14 The Snowman model and The Torso-Shadow Cone [6]

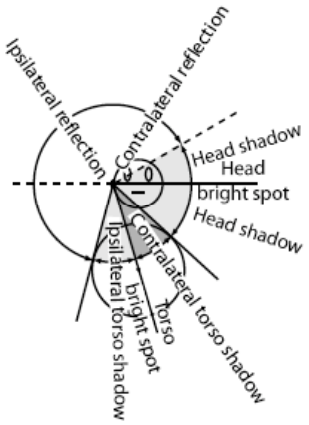


Figure 1.15 Zones of the right ear response [6]

The "snowman" model exhibits two bright spots: One due to the head around the elevation angle $\phi = 180$ and the other due to the torso around $\phi = 255$, shown in figure 1.15. The notches, shown in figure 1.16, that are symmetric about $\phi = 90$ due to specular reflections from the upper torso and the deeper notches around 210˚ to 250˚ are caused by torso shadow [6].

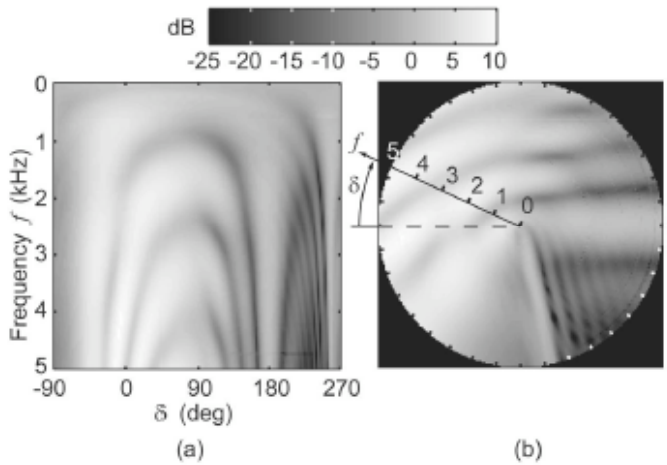


Figure 1.16 The computed HRTF for the physical snowman model [6]

1.2.3 Movements of Head and Source

Thinking about a spherical head and the pinnae disregarded, sources at many different locations produce essentially the same ITD and IID if they are located in the so called "cone of confusion". It comes without saying that the human head is not spherical and the statement pointed out above is only theoretical, but when IID and IID cues are maximally similar, such confusion exists. In this cone of confusion, shown in figure 1.17, the source cannot be localized clearly.

To improve localization we move our head in order to minimize the interaural differences, using our head as a kind of "pointer" [1]. Moving the head causes dynamic changes for a fixed source, a moving source will cause dynamic changes for a fixed head.

One of the main cues for moving sources is the Doppler shift. This effect represents the change in pitch if the source passes a stationary listener.

Another important parameter is the minimum audible movement angle (MAMA). MAMAs can range up to 3 degrees under optimal conditions (narrowband source, velocity: 2.8˚ to 360˚/sec), but can increase as a function of movement velocity, location of movement and source type [1].

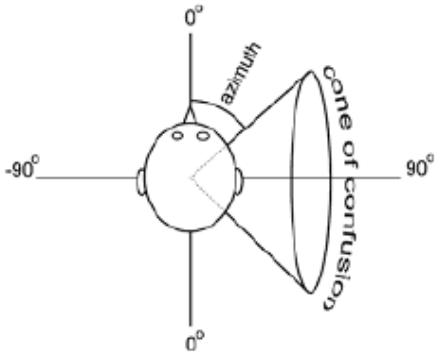


Figure 1.17 The Cone of Confusion [6]

1.3 Sound Distance and Reverberation

1.3.1 Distance

As mentioned in Begault D. R. [1], the intensity of a sound source (and its interpretation as loudness) is the primary distance cue. Auditory distance is learned from a visual-aural observation, correlating the physical displacement of sound source with corresponding increases and reductions in intensity.

For an omni directional sound source, intensity reduction is adjusted to be 6 dB for doubling of distance, shown in figure 1.18, with a dependence of $1/r^2$ (where r represents the distance to the source) or also called inverse square law. If the sound source is a line source, then the intensity will fall 6 dB for each doubling of distance from the source, with a dependence of $1/r$.

Intensity when measured in dB, expresses the ratio of a sound source's intensity to a reference level, but loudness is the perceived magnitude of intensity. Judgments of half or doubled loudness have been shown to be more closely related to the sone scale, than to the inverse square law. Loudness is frequency dependent, which can be shown with the so called "isophons" or equal loudness contours (figure 1.19).

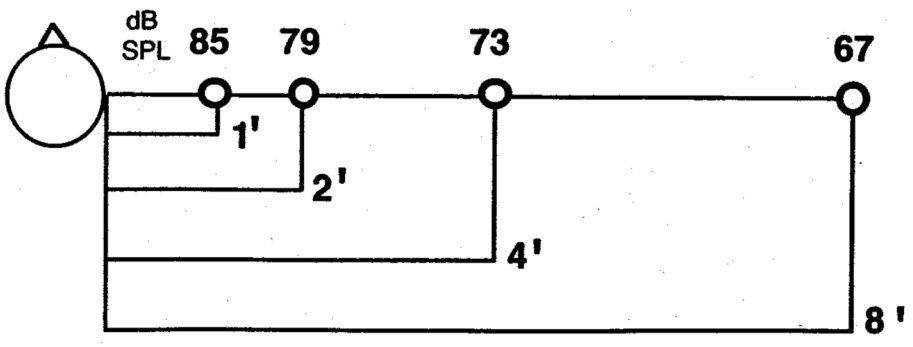


Figure 1.18 Reduction in intensity for doubling distances [1]

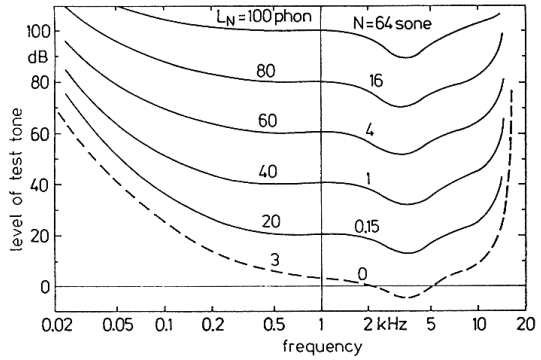


Figure 1.19 Isophons or equal loudness contours [10]

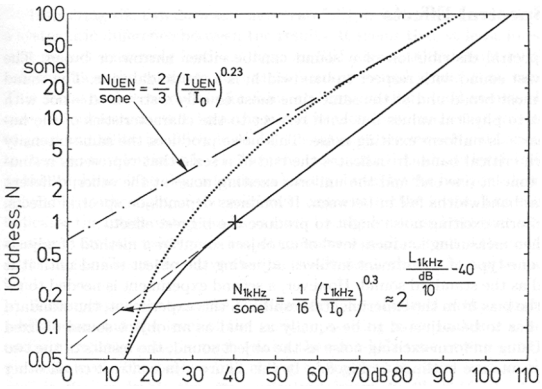


Figure 1.20 Sone Curve [10]

Isophones were derived from studies where subjects were asked to adjust sine waves to be twice as loud. Double the number of sones and you have doubled the loudness. Between 400-5000 Hz and 40-100 dB, this corresponds to an increase of 10 dB if the loudness is doubled (figure 1.20). The relation between intensity and perceived magnitude of intensity is shown in figure 1.21.

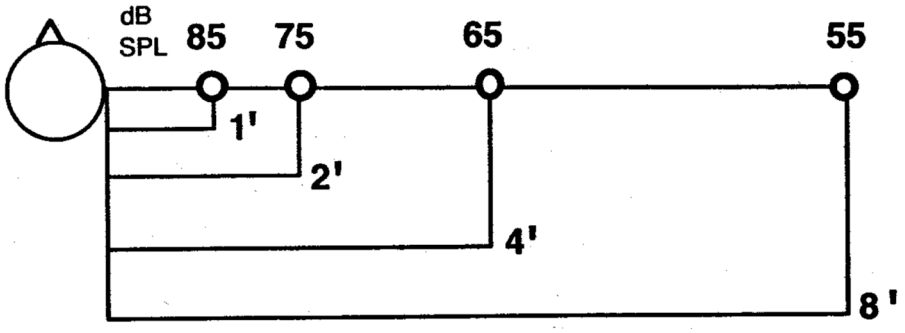


Figure 1.21 Reduction of intensity using loudness scales based on sones [1]

Spectral content of the sound source relative to a receiver position can also vary as a function of its distance. The effect includes the influence of atmospheric conditions, molecular absorption of the air and the curvature of the waveform [1]. Compared to loudness and reverberation cues, this is a relative weak cue. For high frequencies and large distances an air absorption coefficient can be calculated, shown in figure 1.22.

For nearby sources, the tone darking effect can be noticed. This effect is related to the equal loudness contours, which show that sensitivity to low frequencies increases with increasing sound pressure level.

A weak effect is the wind profile and ground cover effect, shown in Begault D. R. [1].

Using headphones in a 3D audio system can cause Inside-the-head localisation (IHL), especially without reverberation [1]. IHL can be avoided, as the stimulation approximates more closely a stimulation that is natural. The likely sources of these natural interaural attributes include the binaural HRTF, head movement and reverberation.

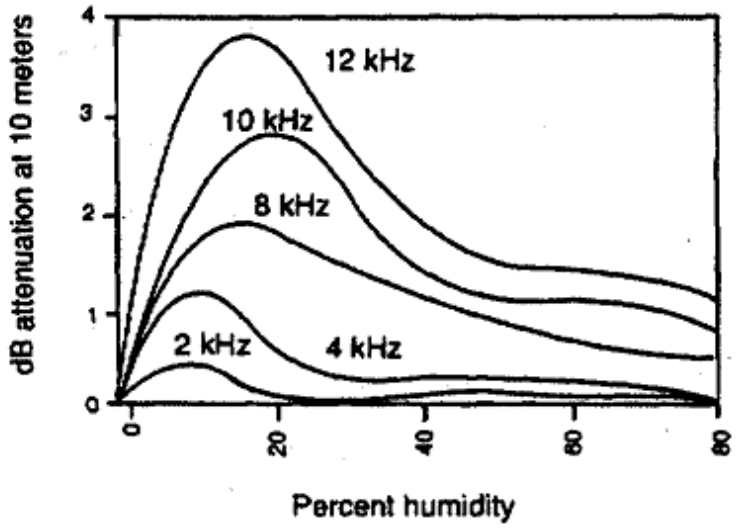


Figure 1.22 AIR Absorption [1]

1.3.2 Reverberation

As mentioned in 1.1.2 there are direct and indirect or reflected sound. Direct sound is defined as the wavefront that reaches the ear first by a linear path, without any reflections and reflected sound (reverberation) refers to the energy of a sound source that reaches the listener indirectly, by reflecting from surfaces within the surrounding space occupied by the sound source and the listener. Characteristics of a chamber can be described by an impulse response (figure 1.23) or a reflectogram (figure 1.24), both mapped in the time domain.

Reverberation, depending on the time of arrival to the listener can be classified into two sections: Early reflections (ER) and late reflections (LR). Early reflections are the first reflections on the floor and the roof and reach the receiver within a period around 1-80 msec. Followed by the late reflections or late reverberation, which contain less energy overall and result from many subsequent reflections from surface to surface of the environmental context. Psychoacoustically, the individual earlier reflections are less likely to be masked (inaudible) than later reflections. The psychoacoustic evaluation of reverberation places most of emphasis on the first 80 msec of the response, as pointed in [1].

Early and late reflections are often described by the following three physical parameters: The ratio of reverberant to direct sound, the reverberant time and several criteria related to the arrival time and spatial incidence of early reflections.

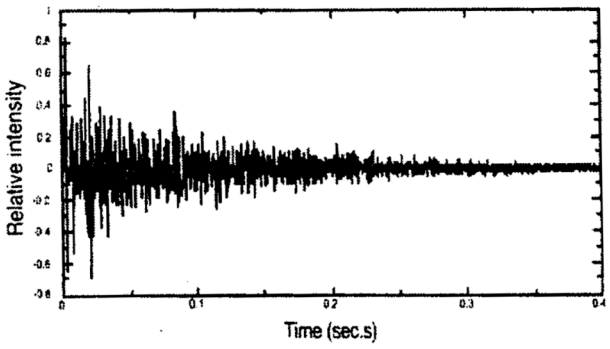


Figure 1.23 Impuls response [1]

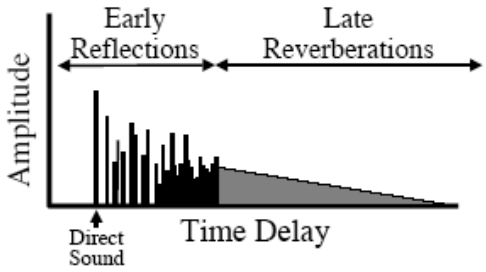


Figure 1.24 Reflectogram [Internet]

The ratio of reverberant to direct sound (R/D) has been cited in many studies as a cue to distance. Von Békésy (1960) observed that when he changed the R/D ratio, the loudness of the sound remained constant, but a sensation of changing distance occurred.

The reverberation time (t_{60}) is defined as the duration necessary fort he energy of reflected sound to drop 60 dB below the level of the direct sound.

The third parameter is related to temporal and spatial patterns of early reflections. Different spatial-temporal patterns can affect distance perception. By measuring the similarity of the reverberation over a specific time window at two ears, a value for interaural cross-correlation can be obtained (Blauert and Cobben, 1978; Ando, 1985).

The critical distance is the distance where the R/D ratio is 1.

Chapter 2

Head Related Transfer Functions (HRTFs)

2.1 Head Related Transfer Functions (HRTF)

2.1.1 What Influences HRTF Curves

Head-related transfer functions (HRTFs) are influenced by directional and nondirectional components (after Genuit, 1984; Gierlich, 1992). Directional components are the torso, shoulder reflection, head diffraction and reflection and pinnae, as described in chapter 1.2.2. Nondirectional components are the cavum conchae dominant resonance, the ear canal and eardrum impedance. A structural model of these components is shown in figure 2.1.

The influence of the upper body and shoulders become apparent in the range of 100 Hz- 2 kHz. The shoulder influence is about $\pm 5dB$ and the influence of the torso about $\pm 3dB$. The ear canal can be seen as an acoustic transmission line between the eardrum and the outer ear. With a length of 2.5 cm and a diameter of 7-8 mm, it has a significant resonance at 3-4 kHz. The level difference between the eardrum and the entrance of the ear is mapped in figure 2.2.

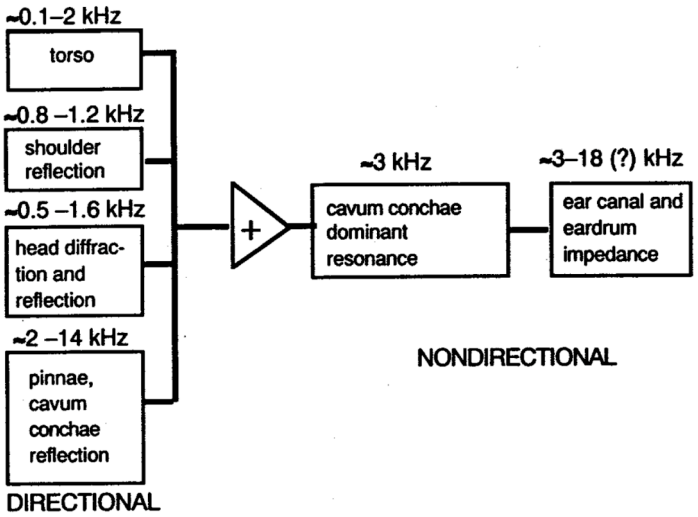


Figure 2.1 Directional and nondirectional components of the HRTF [1]

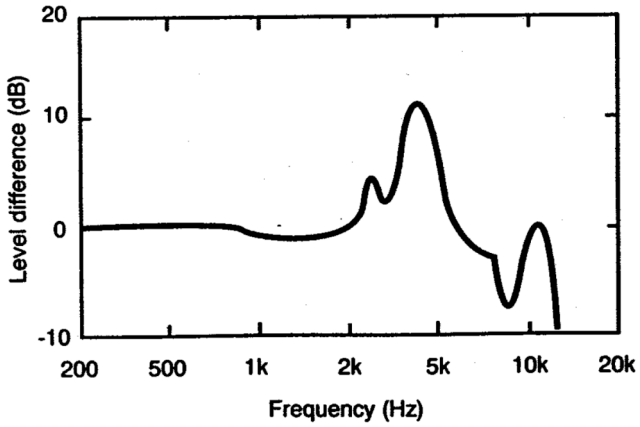


Figure 2.2 Ear canal resonance [1]

The largest resonant area of the pinnae is called cavum conchae and is located at the opening of the ear canal. Blauert (1983) reports experiments, where he demonstrated the difference in the resonant effect of the cavum conchae on a stimuli heard from the front compared to the back (0˚-180˚). A difference of around 5 dB occurred at 10 kHz.

2.1.2 HRTF Magnitude Characteristics

Figures 2.3-2.5 show examples of individualized HRTFs from three persons measured in the same laboratory. The differences between the three subjects are caused by the fact that every

pinna has an individual shape. In most 3-D systems individualized HRTFs cannot be used. That is why, in most cases nonindividualized HRTFs are taken, which can be derived from individualized HRTF averages. But in averaging HRTFs problems could occur when significant peaks of the direction are diminished overall. Another option is measuring the HRTF set with a dummy head with average size of torso, shoulders and pinnae.

2.1.3 HRTF Phase Characteristics

If a wideband sound source containing all audible frequencies were played at the pinnae, some frequencies would arrive later at the eardrum [1]. This delay can be measured in degrees. Figure 2.6 shows, for a frequency range of 500-4000 Hz, the unwrapped phase difference for a single person at 30˚ increments of azimuth from 0˚-150˚, at 0˚ elevation. The bold lines show the conversion from phase delay to interaural time delays at 0.25 msec, 0.5 msec and 1 msec.

As mentioned in Begault D. R. [1], the phase response at a single pinna is less critical for localization than the interaural time difference.

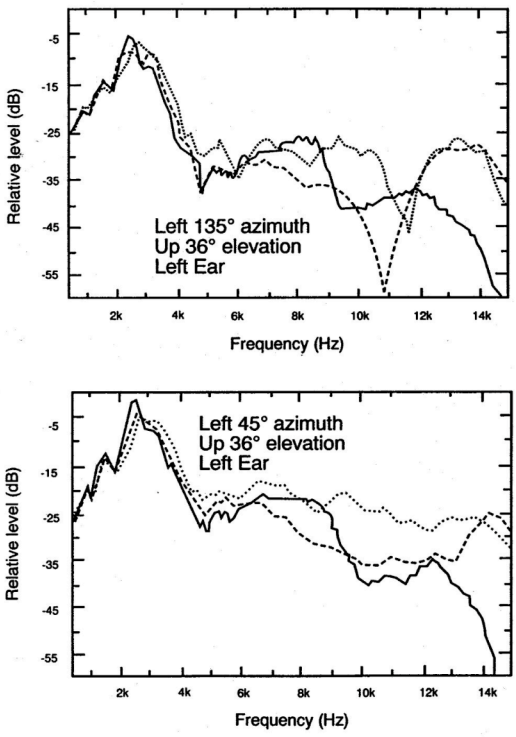


Figure 2.3 HRTFs of left and right ear, 45˚ azimuth and 36˚ elevation, measured for three people [1]

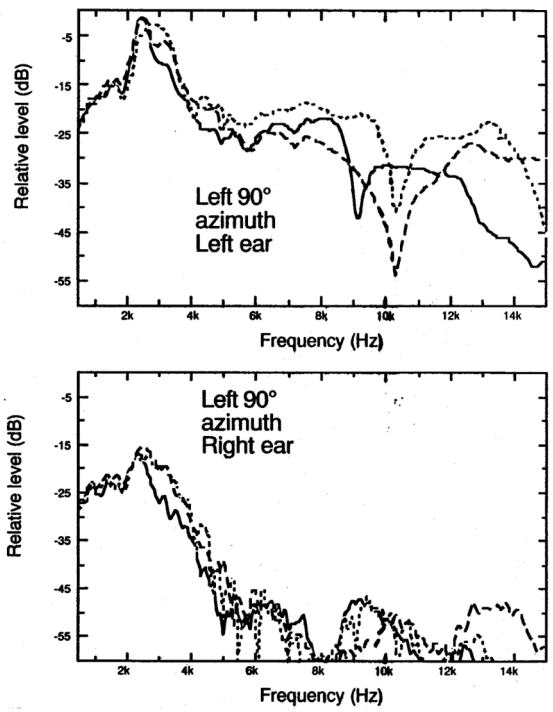


Figure 2.4 HRTFs of left and right ear, 90˚ azimuth and 0˚ elevation, measured for three people [1]

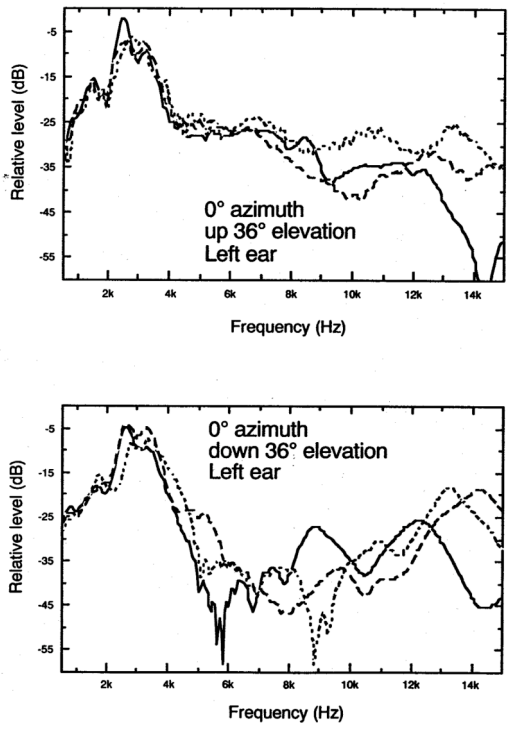


Figure 2.5 HRTFs of left and right ear, 0˚ azimuth and 36˚ elevation, measured for three people [1]

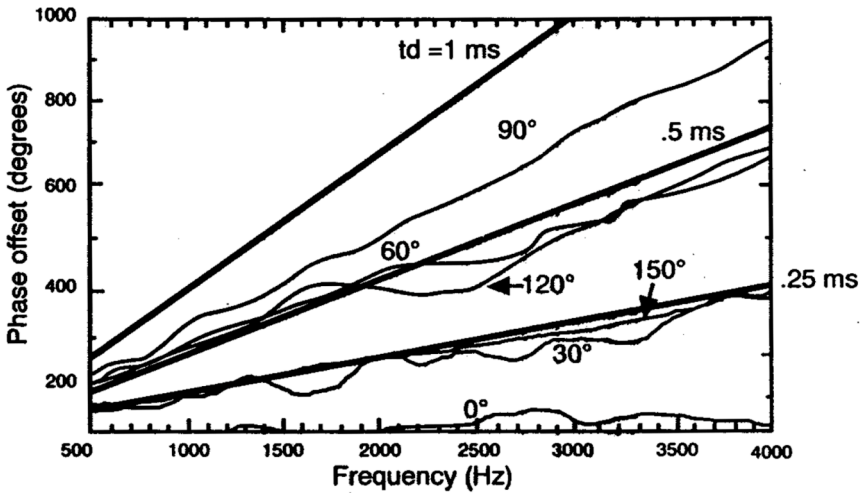


Figure 2.6 Unwrapped interaural phase difference [1]

2.2 Localization with HRTF Cues

2.2.1 Spectral Cues

The main role of HRTF cues from a psychoacoustic standpoint is thought to be the disambiguation of front to back for sources on the cone of confusion and as an elevation cue for disambiguating up from down.

Figure 2.7, for example shows the spectral difference in the HRTF at one ear of a single person, for two positions on the cone of confusion: 60 and 120 degrees azimuth (0 degrees elevation). The main difference occurs primarily in the upper frequencies, especially the broad peak at 5 kHz and the notch around 9 kHz. The relationship between spectral weighting and front-back confusions can be seen in evaluations of reversals as a function of single frequencies. In one study, confusions were most pronounced in the frequency region where ITD "takes over" for IID, i.e. around 1.5 kHz (Mills, 1972). Stevens and Newman (1936) showed that sine waves below 3 kHz were judged as front or back on the order of chance, independent of actual location. This evidence, combined with the cone of confusion model and the fact that the binaural HRTF shows minimal IID below 400 Hz, suggests the role of spectral shaping at higher frequencies for front-back spectral cueing.

For vertical localization, the directional effects of the pinnae are considered to be particularly important. Several studies have shown that without the pinnae's effect on a broadband source, vertical localization is diminished (Gardner and Gardner, 1973; Oldfield and Parker 1984). To examine the HRTF spectral cues for elevation, some studies have isolated the role of ITD and IID by examining the special case of sources along the median plane of the listener; i.e. the plane that intersects the left and the right sides of the head, at 0 and 180 degrees azimuth. Notice that this includes those positions where, beside IID and ITD, the spectral differences would be minimized between the HRTFs to the degree that the left and right pinnae are identical. The role of the pinnae is also evaluated by comparing judgements made under normal conditions to a condition where the pinnae are bypassed or occluded (Musicant and Butler, 1984, Oldfield and Parker 1984). These studies almost always show that the pinnae improve vertical localization. Other experiments where one ear was blocked support the fact that spectral cues also work monaurally; i.e. the spectral difference between the HRTFs is considered to be less important than the overall spectral modification at a particular ear (Middlebrooks and Green, 1991), although there is some evidence for binaural pinnae cues (Searle, 1975).

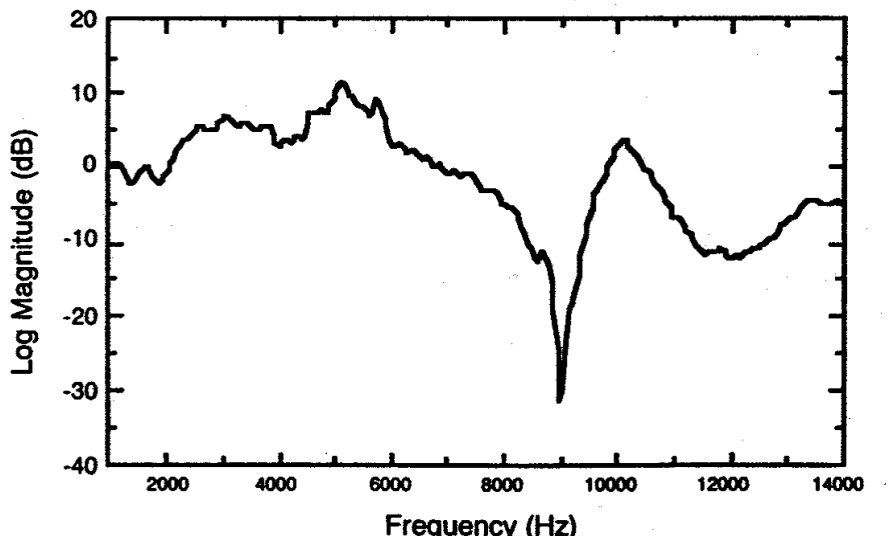


Figure 2.7 Difference in spectra between two front-back source locations on a cone of confusion: 60 and 180 degrees azimuth, 0 degrees elevation. [1]

2.2.2 Spectral Band Sensitivity and Directional Bands

As pointed out above, spectral cues can influence the perception of direction independent of the location of a sound source. There have been several researches on how peaks and troughs influence our perception of localization. One of the most important results is the "spectral band sensitivity". Blauert (1969) used third-octave-filtered bands of noise across frequency to determine the directional bias associated with the spectral band, in terms of three categories: Above, in front, or behind [1]. Blauert determined that "directional bands" existed, shown in figure 2.8.

The accuracy of localization is intersubjectively different. There are humans with good localization and others are bad localizers. Wenzel (1988) showed that a person who was a relatively good localizer in elevation judgments would degrade his or her ability when listening through the pinnae of a person who was inherently worse in free-field elevation judgments. While the converse question – could bad localizer improve listening with good localizer's HRTFs – has not been fully evaluated. A study (Wenzel, 1993) showed that neither deterioration nor improvement was detected, while listening with HRTFs of better localizers.

Perceived location	Center frequency kHz	Bandwidth kHz
overhead	8	4
forward (band #1)	0.4	0.2
forward (band #2)	4	3
rear (band #1)	1	1
rear (band #2)	12	4

Figure 2.8 Centre frequencies and frequency bandwidths for "directional bands"[1]

2.3 How to Measure HRTF Sets

2.3.1 Measurement

One of the most important properties of a transfer function – or in our case of a head related transfer function – is its impulse response. Equation 2.1 describes a LTI system, where x(n) is the input signal, y(n) the output signal and h(n) the impulse response:

$$y(n) = h(n) * x(n) \tag{2.1}$$

Converted to the frequency domain a convolution is a simple multiplication:

$$Y(z) = H(z)X(z) \tag{2.2}$$

Thus,

$$H(z) = \frac{Y(z)}{X(z)} \tag{2.3}$$

An impulse response can be measured with different methods. In general, the following demands on a measure method are made:

- The excitation signal has to be absolutely reproducible
- For an interferenceless processing, the signal to noise ratio has to be greater than 80 dB.
- High levels of the excitation signal can produce harmonic distortion of the speaker. These nonlinearities should be cancelled well by the measurement method.

The most common measurement methods are MLS (Maximum Length Sequence) and TDS (Time Delayed Spectrometry). An alternative method for cancelling harmonic distortion is called Swept Sine Technique [9].

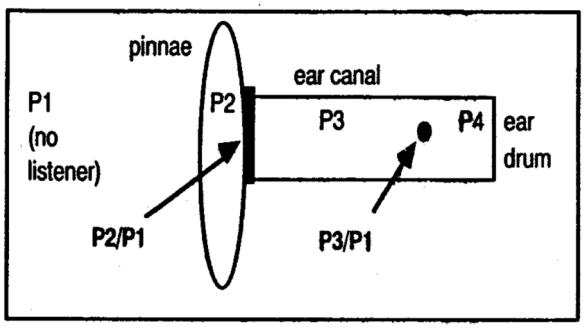


Figure 2.9 Measurement points for the HRTF [1]

In addition, HRTF measurements can be calculated either in terms of a point within the ear canal (figure 2.8 – P3/P1) or at the entrance of the blocked ear canal (figure 2.8 – P2/P1), as mentioned in Begault D. R. [1]. P3/P1 includes both directional and nondirectional aspects of the HRTF, whereas in P2/P1 only directional aspects are included. When this point is used, the nondirectional aspect, the spectral cue of the ear canal (P3) can be calculated once and then convolved with each measurement.

Figure 2.10 shows a possible HRTF measurement, used at the institute of electronic music (IEM) in Graz/Austria.

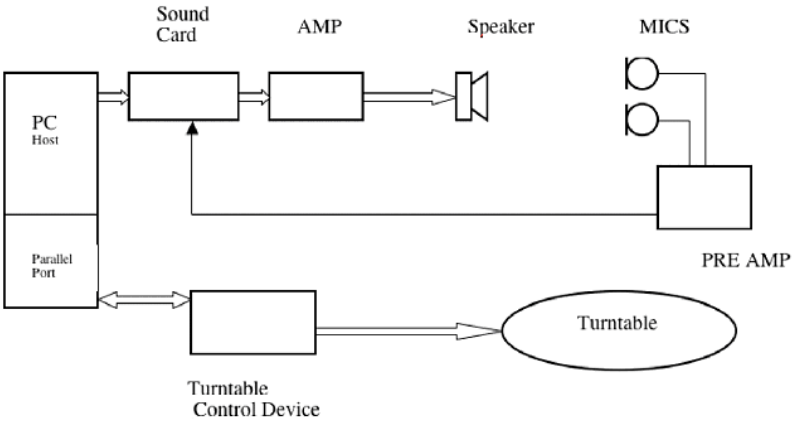


Figure 2.10 HRTF measurement

2.3.2 Equalisation

Raw impulse measurements must be modified in both time and frequency domain before using in a 3-D audio system. The time domain modifications are relatively easy, while the modification in the frequency domain, the compensations of the material, is more complicated.

The first time domain modification is to discard the blank portion at the beginning that results from the shortest flight time (the measurement position with the ear nearest to the speaker) of the sound to the microphone. This can be applied to all HRTFs.

The next step is to normalize the HRTF set by finding the loudest sample in a given sequence and then multiplying all samples in that sequence so that the loudest sample is at the maximum quantization value.

After that, the transfer function of the microphone, the speakers, the headphone and the ear canal (if is not part of the impulse response measurement) is compensated in the frequency domain.

The frequency domain transfer function for one ear can be represented as follows :

A(Z) = analytic signal in frequency domain
M(Z) = transfer function of the microphone
C(Z) = transfer function of the ear canal
L(Z) = transfer function of the loudspeaker
HP(Z) = transfer function of the headphone
H(Z) = naturally occurring HRTF in a free field
RAW(Z) = uncorrelated HRTF for virtual simulation
COR(Z) = correlated HRTF for virtual simulation
INV(Z) = inverse filter for virtual simulation
YE(Z) = the signal arriving at the eardrum
YM(Z) = the signal arriving at the microphone
X(Z) = an input signal to be spatialized

In natural spatial hearing, a sound source played by a loudspeaker can be described as:

$$YE(Z)_{natural} = X(Z)H(Z)C(Z)L(Z) \tag{2.4}$$

In a virtual spatial hearing simulation, we desire

$$YE(Z)_{virtual} = YE(Z)_{natural} \tag{2.5}$$

First, the uncorrelated HRTF is measured for a particular direction by playing the analytic signal through the loudspeaker:

$$RAW(Z) = YM(Z) = A(Z)M(Z)C(Z)H(Z)L(Z) \tag{2.6}$$

Second, the headphone and ear canal are measured by playing the analytic signal through the headphones:

$$YM(Z) = A(Z)M(Z)C(Z)HP(Z) \tag{2.7}$$

To obtain COR(Z), it is first necessary to find the inverse filter INV(Z):

$$INV(Z) = \frac{1}{YM(Z)} = \frac{1}{A(Z)M(Z)C(Z)HP(Z)} \tag{2.8}$$

Then the uncorrected HRTF is corrected as follows:

$$\begin{aligned} COR(Z) &= RAW(Z)INV(Z) \\ &= \frac{A(Z)M(Z)C(Z)H(Z)L(Z)}{A(Z)M(Z)C(Z)HP(Z)} \\ &= \frac{H(Z)L(Z)}{HP(Z)} \end{aligned} \tag{2.9}$$

To create a virtual sound source, the spectrum of the input is convolved with COR(Z), and then played through headphones via the ear canal of the listener:

$$\begin{aligned} YE(Z) &= \left[X(Z)COR(Z)\right]\left[HP(Z)C(Z)\right] \\ &= \left[X(Z)\frac{H(Z)L(Z)}{HP(Z)}\right]\left[HP(Z)C(Z)\right] \\ &= X(Z)H(Z)C(Z)L(Z) \end{aligned} \tag{2.10}$$

The final part of the previous equation shows that the natural spatial hearing and the virtual simulation are equivalent:

$$YE(Z)_{natural} = YE(Z)_{virtual} = X(Z)H(Z)C(Z)L(Z) \tag{2.11}$$

2.4 Existing HRTF Material

2.4.1 KEMAR Database

Figure 2.11 The KEMAR mannequin head and torso [11]

The following paragraph deals with the measurement technique, the measurement procedure, and the data of the KEMAR HRTFs and is totally taken over from Gardner B. and Martin K. [11].

Measurements were made using a Macintosh Quadra computer equipped with an Audiomedia-II DSP card, which has 16-bit stereo A/D and D/A converters that operate at a 44.1 kHz sampling rate. One of the audio output channels was sent to an amplifier which drove a Realistic Optimus Pro 7 loudspeaker. This is a small two way loudspeaker with a 4 inch woofer and 1 inch tweeter. The KEMAR mannequin (figure 2.11), Knowles Electronics model DB-4004, was equipped with model DB-061 left pinna, model DB-065 (large red) right pinna, Etymotic ER-11 microphones, and Etymotic ER-11 preamplifiers. The outputs of the microphone preamplifiers were connected to the stereo inputs of the Audiomedia card.

The impulse responses were obtained using MLS. The sequence length was N = 16383 samples, corresponding to a 14-bit generating register. Two copies of the sequence were concatenated to from 2*N sample sound which was played from Audiomedia card. Simultaneously, 2*N samples were recorded on both the left and right input channels. For each input channel, the following technique was used to recover the impulse response. The first N samples of the result were discarded, and the remaining N samples were duplicated to form a 2*N sample sequence. This was cross-correlated with the original N sample MLS using FFT based block convolution, forming a 3*N-1 sample result. The N sample impulse response was extracted starting at N-1 samples into this result.

The measurement was made in MIT's anechoic chamber. The KEMAR was mounted upright on a motorized turntable which could be rotated accurately to any azimuth under computer control. The speaker was mounted on a boom stand which enabled accurate positioning of the speaker to any elevation with respect to the KEMAR. Thus, the measurements were made one elevation at a time, by setting the speaker to proper elevation and then rotating the KEMAR to each azimuth. With the KEMAR facing forward toward the speaker (0 degrees azimuth), the speaker was positioned such that a normal ray projected from the centre of the face of the speaker bisected the interaural axis of the KEMAR at a distance of 1.4 meters. The speaker was always within 0.5 inch of the desired position, which corresponds to an angular error of ± 0.5 degrees.

The spherical space around the KEMAR was sampled at elevations from -40 degrees (40 degrees below the horizontal plane) to +90 degrees (directly overhead). At each elevation, a full 360 degrees of azimuth was sampled in equal size increments. The increment sizes were

chosen to maintain approximately 5 degree great-circle increments. Figure 2.12 shows the number of samples and azimuth increment at each elevation.

The 1.4 meter air travel corresponds to approximately 180 samples, and there is an additional delay of 50 samples inherent in the playback/recording system. In order to reduce the size of data set without eliminating anything of potential interest, the first 200 samples of each impulse response were discarded and the following 512 were saved. Each HRTF response is 512 samples long and stored as 16-bit signed integers, with the most significant byte (MSB) stored in the low address (Motorola 68000 format). The dynamic range of the 16-bit integers (96 dB) exceeds the signal to noise ratio of the measurements, which had been measured to be 65 dB. An inverse filter was also designed by zero-padding the measured impulse response and taking the DFT of the zero-padded sequence. The resulting complex spectrum was inverted by negating the phase and inverting the magnitude.

A data-reduced set of 128 point symmetrical HRTFs also exist.

Elevation	Numbers of Measurements	Azimuth Increment
-40	56	6.43
-30	60	6.00
-20	72	5.00
-10	72	5.00
0	72	5.00
10	72	5.00
20	72	5.00
30	60	6.00
40	56	6.43
50	45	8.00
60	36	10.00
70	24	15.00
80	12	30.00
90	1	x.xx

Figure 2.12 Number of samples and azimuth increment at each elevation

2.4.2 CIPIC

The following paragraph deals with the measurement technique of CIPIC and is totally taken over from the CIPIC Database [13]:
The CIPIC Interface Laboratory at U.C. Davis has measured HRTFs at high spatial resolution for 43 human subjects (Release 1.0), excluding the KEMAR mannequin with large and small pinnae. All HRTFs were measured with the subject seated at a centre of 1 meter radius hoop whose axis was aligned with the subject's interaural axis. The position of the subject was not constrained, but the subject could monitor his or her head position.
Both Acoustimass loudspeaker (5.8 cm cone diameter) were mounted at various positions along the hoop. A modified Snapshot system from Crystal River Engineering generated Golay-code signals. The subject's ear canals were blocked, and Etymotic Research ER-7C probe microphones were used to pick up the Golay-code signals. The microphone outputs were digitized at 44.1 kHz, 16-bit resolution and processed by Snapshot's "oneshot" function to yield a raw HRIR. A modified Hanning window was applied to the raw HRIR measurements to remove room reflections, and the results were free-field compensated to correct for the spectral characteristics of the transducers. The length of each HRIR is 200 samples, corresponding a duration of about 4.5 ms.
Sound source location was specified by the azimuth angle and elevation angle in interaural-polar coordinates. Elevations were uniformly sampled in 360/64 = 5.625˚ steps from -45˚ to +230.625˚. To obtain roughly uniform density on the sphere, azimuths were sampled at -80˚, -65˚, -55˚, from -45˚ to 45˚ in steps of 5˚, at 55˚, 65˚, and 80˚. This leads to spatial sampling at 1250 points.

Chapter 3
The Acoustical Environment

3.1 Indoor and Outdoor Sound Propagation

3.1.1 Air Absorption

Sound energy is dissipated in air by two major mechanisms:

• Viscous losses due to friction between air molecules, which result in heat generation, called "classical absorption".

• Relaxational processes: Sound energy is momentarily absorbed in the air molecules and causes the molecules to vibrate and rotate. These molecules can re-radiate sound at a later instant (like small echo chamber) which can partially interfere with the incoming sound.

These mechanisms have been extensively studied, empirically quantified, and codified into an international standard for calculation: ANSI Standard S1-26:1995, or ISO 9613-1:1996.
For a standard pressure of one atmosphere, the absorption coefficient α (in dB/100m) can be calculated as a function of frequency f (Hz), temperature T (degrees Kelvin) and molar concentration of water vapour h (%) by:

$$\alpha = 869 \times f^2 \left\{ 1.84 \times 10^{-11} \left(\frac{T}{T_0} \right)^{\frac{1}{2}} + \left(\frac{T}{T_0} \right)^{-\frac{5}{2}} \left[0.01275 \frac{e^{-2239.1/T}}{F_{r,O} + f^2 / F_{r,O}} + 0.1068 \frac{e^{-3352/T}}{F_{r,N} + f^2 / F_{r,N}} \right] \right\} \qquad (3.1)$$

With the Oxygen relaxation frequency [Hz]

$$F_{r,O} = 24 + 4.04 \times 10^4 \, h \frac{0.02 + h}{0.391 + h} \qquad (3.2)$$

and the Nitrogen relaxation frequency [Hz]

$$F_{r,N} = \left(\frac{T}{T_0} \right)^{-\frac{1}{2}} \left(9 + 280 \times h \times e^{\left\{ -4.17 \left(\frac{T}{T_0} \right)^{-\frac{1}{3}} - 1 \right\}} \right) \qquad (3.3)$$

where

$$T_0 = 293.15 K \quad (20\ ^\circ C)$$

A plot of the absorption coefficient for air at 20 ˚C and 70 % relative humidity is shown in figure 3.1 (ref. ANSI standard S1.26). The predominant mechanism of absorption (the classical and rotational relaxation) is proportional to the square of frequency. The vibration relaxation effect depends on the relaxation frequencies of the gas constituents (O and N) and is highly dependent on the relative humidity. Figure 3.2 shows the relation of humidity. It is interesting to note that absorption generally decreases with increasing humidity. The exception is totally dry air, which has the least absorption.

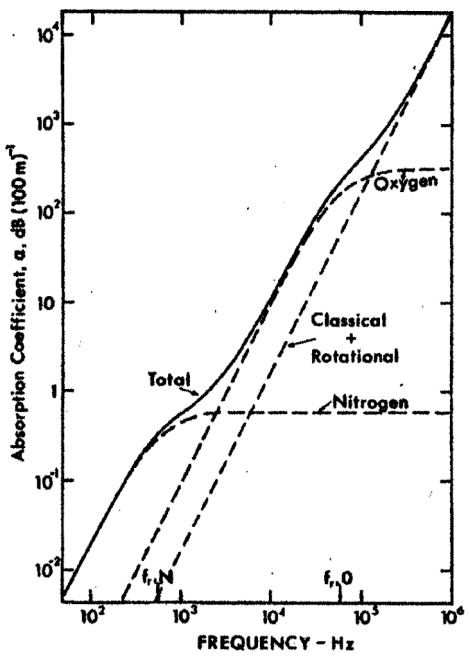


Figure 3.1 Predicted atmospheric absorption in dB/100m for a pressure of 1 atm, temperature of 20˚C and relative humidity of 70 % [15]

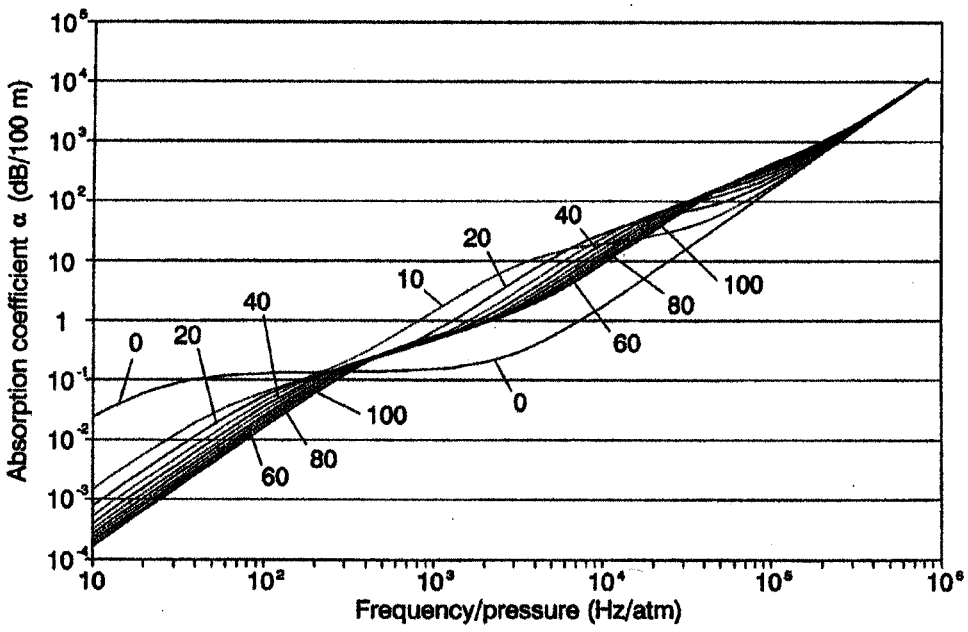


Figure 3.2 Sound absorption coefficient in air (dB/100 m) versus frequency/pressure ratio for various relative humidity at 20˚C. [15]

To implement the air absorption a FIR (Finite Impulse Response) and an IIR (Infinite Impulse Response) filter is designed and compared to the desired signal. The number of filter coefficients is chosen to be 20. Normally this quantity of coeffiecients is not necesarry but since this system is not a real time application we aspire to design our filters as well as possible.

For the FIR filter implementation a Parks-McClellan optimal equiripple FIR filter is chosen. This filter uses the Chebyshev approximation theory and an algorithm where the maximum error between the desired frequency response and the actual frequency response is minimized [19]. Filters designed this way exhibit an equiripple behavior in their frequency response and therefore they are sometimes called equiripple filters. Figure 3.3 shows the desired frequency response and the one designed with the Parks-McClellan optimal equiripple FIR filter. The distance is chosen to be 50 meters, the relative humidity is 70 % and the environmental temperature 20 ˚C. The number of filter coeffiecients is set to be 3. The error between these two signals is plotted below in figure 3.3. The mean error between the desired and the actual frequency response in the frequency range from 0Hz to 10 kHz is 0.96 dB.

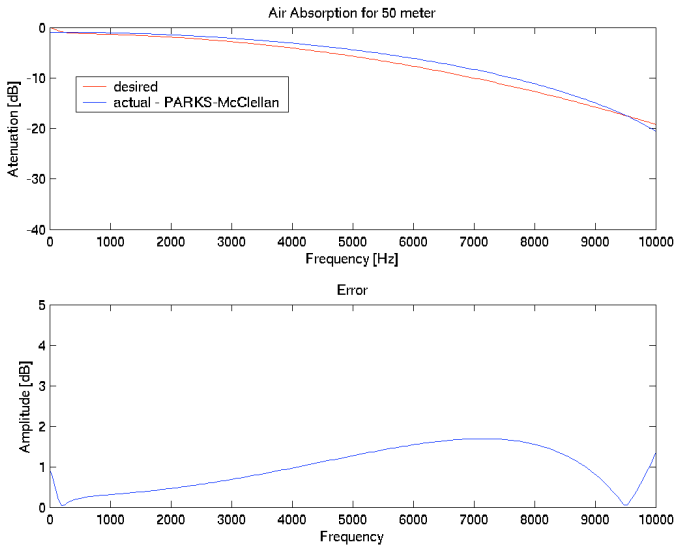


Figure 3.3 Desired frequency response and actual frequency response designed with the Parks-McClellan optimal equiripple FIR filter

The second filter is implemented with a recursive IIR filter using least-squares (called Yulewalker-algorithm [19]) to fit a specified frequency response. Figure 3.4 shows the desired frequency response and actual frequency response designed with the Yulewalker IIR filter. The distance is chosen to be 50 meters, the relative humidity is 70 % and the environmental temperature 20 ˚C. The number of filter coeffiecients is set to be 3. The error between these two signals is plotted below in figure 3.3. . The mean error between the desired and the actual frequency response in the frequency range from 0 Hz to 10 kHz is 0.66 dB.
Since there is no perceptual difference between these two filters, the IIR filter with better approximation is chosen.

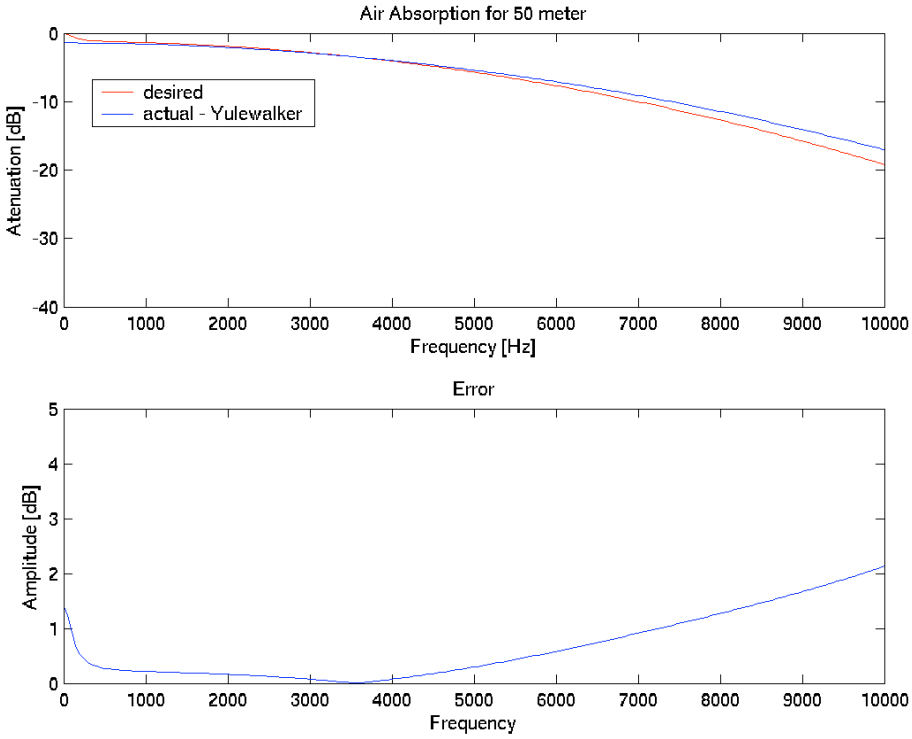


Figure 3.4 Desired frequency response and actual frequency response designed with a Yulewalker recursive IIR filter

3.1.2 Wall Absorption

Room acoustics is concerned with sound propagation in enclosures where the sound conducting medium is bounded on all sides by walls, ceiling and floor. These room boundaries usually reflect a certain fraction of the sound energy impinging on them. Another fraction of the energy is absorbed, i.e. it is extracted from the sound field inside the room, either by conversion into heat or by being transmitted to the outside by the walls.

The sound absorption properties of a material are quantified by its sound absorption coefficient. The sound absorption coefficient of a material can have a value between 0 and 1, with 0 representing no absorption and total reflection, and 1 representing total absorption of all the incident sound. The sound absorption coefficient varies with the frequency of sound.

After several experiments with different filter design methods the best transfer functions have been obtained with a recursive IIR filter using a least-squares called Yulewalker-algorithm. The number of filter coefficients is chosen to be 5. Figure 3.5 and 3.6 show absorption coefficients for different wall, roof and floor materials which are used in this work and in figure 3.7 absorption filters for Wood and Tapestry (0.35 Kg/m2, right) are pictured.

	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	16 kHz
Concrete	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.03
Brick	0.02	0.02	0.03	0.04	0.05	0.05	0.05	0.05
Wood	0.10	0.10	0.05	0.05	0.04	0.04	0.04	0.04
'Wood with air behind	0.30	0.25	0.25	0.17	0.15	0.1	0.1	0.1
Glas	0.4	0.4	0.3	0.3	0.2	0.2	0.2	0.2
'Tapestry (0.35 Kg/m2)	0.04	0.05	0.11	0.18	0.3	0.35	0.35	0.35
'Tapestry (0.60 Kg/m2)	0.14	0.35	0.55	0.75	0.7	0.6	0.6	0.6
Slate	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02
caoutchouc	0.04	0.03	0.04	0.04	0.03	0.02	0.02	0.02
'Wool - Carpet	0.2	0.25	0.35	0.40	0.50	0.75	0.75	0.75

Figure 3.5 Different absorption coefficients for different wall, roof and floor materials

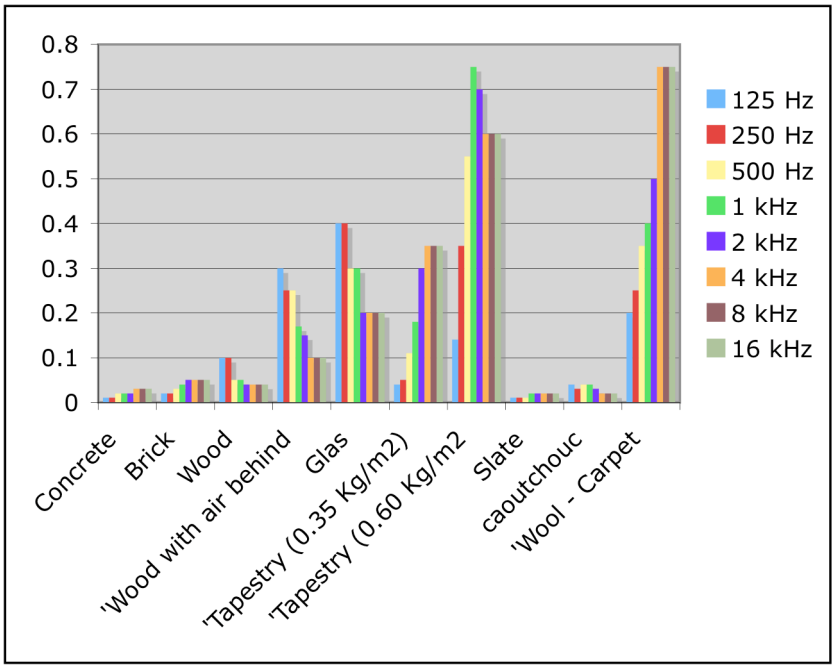


Figure 3.6 Table of different absorption coefficients for different wall, roof and floor materials

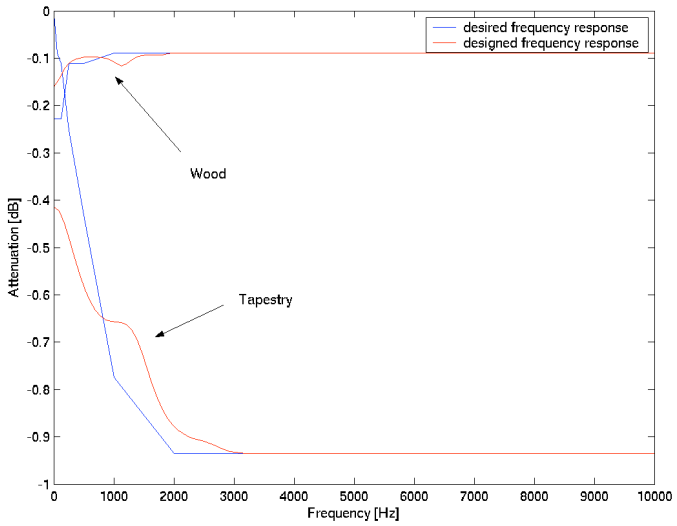


Figure 3.7 Frequency responses of different wall absorption filters

3.1.3 Ground Interaction

In chapter 3.1.2, sound propagation in enclosures, is treated. In contrast to this, outdoor sound propagation takes place in a medium unbounded in every direction, except to the floor.
The surface over which sound propagates can seldom be considered perfectly rigid or totally reflective (with the possible exceptions of open water, ice, or concrete). Typical soil surface with or without vegetation tends to absorb energy from incident acoustic waves. Accurate prediction of ground effects requires knowledge of the absorptive and reflective properties (the acoustic impedance) of the surface.
The ground surface itself also provides a significant path for transmission of acoustic energy, particularly at low grazing angles and low frequencies. Incident acoustic energy is transformed into vibration energy and is transmitted along the surface layer. This vibration disturbance can propagate for long distances, before dissipating or reradiating as sound. At long distances, the transmission of low frequency sound can be dominated by this surface wave mechanism. The geometry for the behaviour of sound propagation over ground of finite impedance is illustrated in figure 3.8.
When airborne sound is incident on a locally reacting fluid, the transmitted wave is refracted at right angles into the surface. The reflected portion of the wave leaves the surface at the angle of incidence, with its amplitude and phase modified by the impedance of the surface. The reflected wave propagates to the receiver, along with direct wave from the source. Depending of their relative phases and amplitudes, they may constructively add or destructively interfere. In the limit, for both source and receiver near the ground and perfect reflection and no atmospheric turbulence, the sound level at the receiver will be increased by 6 dB. Effectively, the receiver sees two sources, the actual source and a reflected or " image " source and the sound pressure is doubled.

In this work, recursive filters using the recursive least-squares method of Yule-Walker [19] with three coefficients have been designed to simulate different environments like foggy ones, meadows, forests. This method for designing filters is appropriate for our purposes because we only have some transfer function values at given frequencies. It matches the magnitude frequency response given in the tables of absorption coefficients. Figure 3.9 and 3.10 show tables of different absorption coefficients α [dB/100 m] of different environments. Figure 3.11 shows frequency responses of the Forest and Grass interaction filters for a source-listener-distance of 10 meters.

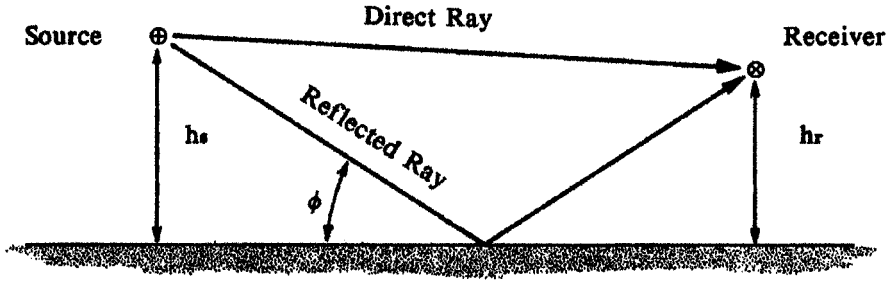


Figure 3.8 Geometry for reflection of sound from level ground of finite impedance.

	63 Hz	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz
Fog	0.8	1.0	1.3	1.6	2.0	2.5	3.0	4.0
Grass	0.7	1.0	1.4	2.0	2.8	4.0	5.6	8.0
Forest	5.0	7.0	10	14	20	28	40	56

Figure 3.9 Table of different absorption coefficients in dB/100 m of different environments

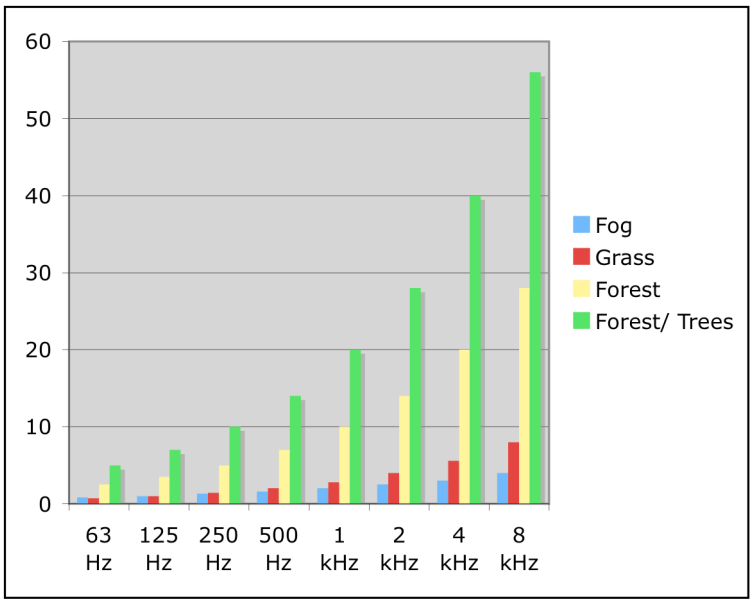


Figure 3.10 Diagram of different absorption coefficients of different environments

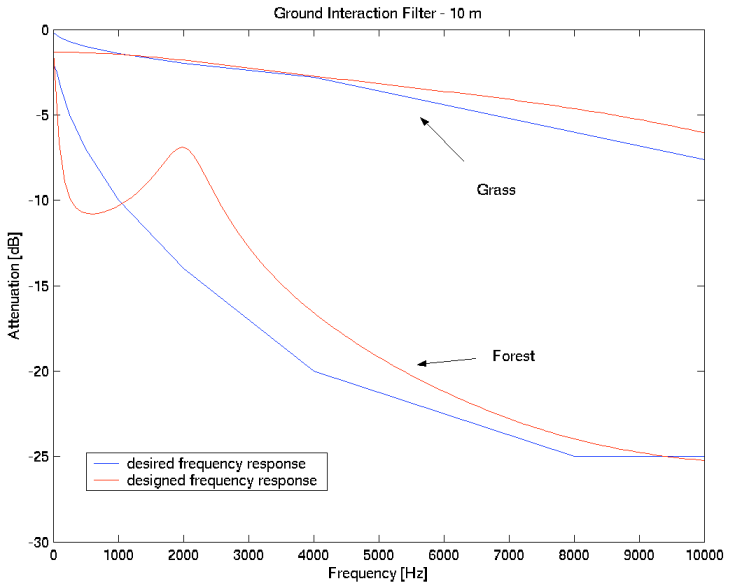


Figure 3.11 Frequency responses of Forest and Grass interaction filter

When we are enjoying a musical performance, speaking to colleagues in the office, walking outdoors on the city street, or even in the woods, the sounds we hear are invariably accompanied by delayed reflections from many different directions, as discussed in chapter 1.3.2. If the reflections occur soon after the initial sound, the result is not perceived as separate sound events. Instead, the reflections modify the perception of the sound, changing the loudness, timbre, and most importantly, the spatial characteristics of the sound. The importance of reverberation in recorded music has resulted in the creation of artificial digital reverberators. We will speak of a reverberation algorithm, or more simply, a reverberator, as a linear discrete-time system that simulates the input-output behaviour of a real or imagined room.

When the room to be simulated, like in our case, does not exist, we can attempt to predict its impulse response based on purely physical considerations. This requires detailed knowledge of the geometry of the room, properties of all surfaces in the room and the position and the directivities of the sources and receivers. Given this prior information, it is possible to apply the laws of acoustics regarding wave propagation and interaction with surfaces to predict how the sound will propagate in the space [13].

The fact that the early and the late reverberation have different physical and perceptual properties permits us to logically split the study of reverberation into early reflections and late reverberation.

3.1.4 Early reflections and the Source Image Method

Early reverberation is mostly easily studied by considering a simple geometrical model of the room. These models depend on the assumption that the dimensions of the reflected surfaces in the room are large compared to the wavelength of the sound. Consequently, the sound wave may be modelled as a ray that is normal to the surface of the wavefront and reflects specularly, like bouncing off a mirror, when the ray encounters a wall surface. Figure 3.12 shows a wall reflection using the ray model. The source is at point A, and we are interested in how sound will propagate to a listener at point B.

The reflected ray may also be constructed by considering the mirror image of the source as reflected across the plane of the wall. In figure 3.12, the image source thus constructed is denoted A'. This technique of reflecting sources across wall surfaces is called the source image method. The method allows a source with reflective boundaries to be modelled as multiple sources with no boundaries.

The image source A' is a first order source, corresponding to a sound path with a single reflection. Higher order sources corresponding to sound paths with multiple reflections are created by reflecting lower order sources across wall boundaries. To create the path of multiple reflections, shown in figure 3.13, we first construct the image source A'. This image source enables us to find the next section of the ray path. Then a second image A'' is constructed from A' associated with the wall at which the ray will arrive next. We continue in this way, obtaining more and more image sources as the total path length of the ray increases.

For a given enclosure and sound source position, the image source can be constructed without referring to a particular sound path. Suppose the enclosure is made up of N plane walls, each wall is associated with one image of the original sound source. Now each of these image sources of first order is mirrored by each wall, with leads to N(N-1) new images which are of second order. By repeating this procedure again and again a rapidly growing number of images is generated with increasing distance from the original source.

When the room is rectangular, as shown in figure 3.14, the pattern of image sources is regular and trivial to calculate.

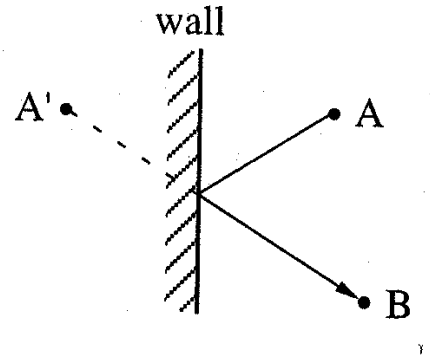


Figure 3.12 Single wall reflection and corresponding image source A'

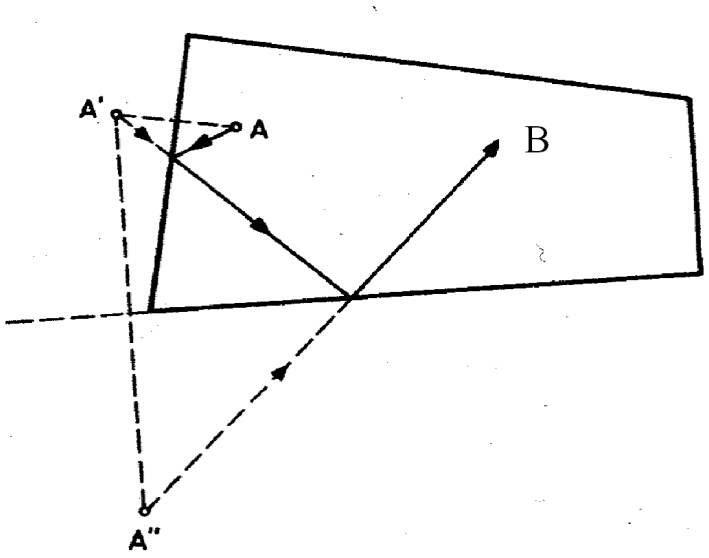


Figure 3.13 Image source of first and second order

The number of images of order i *is* $N(N-1)^{i-1}$ for $i \geq 1$. The total number of images is obtained by adding equation 3.4.

$$N(i) = N\frac{(N-1)^i - 1}{N-2} \tag{3.4}$$

The source image method is impractical for studying late reverberation because the number of sources and consequently the computation of the signal increase exponentially.

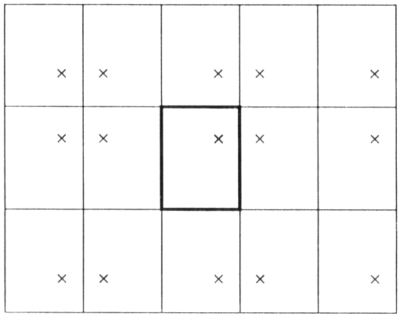


Figure 3.14 A regular pattern of image sources occurs in an ideal rectangular room

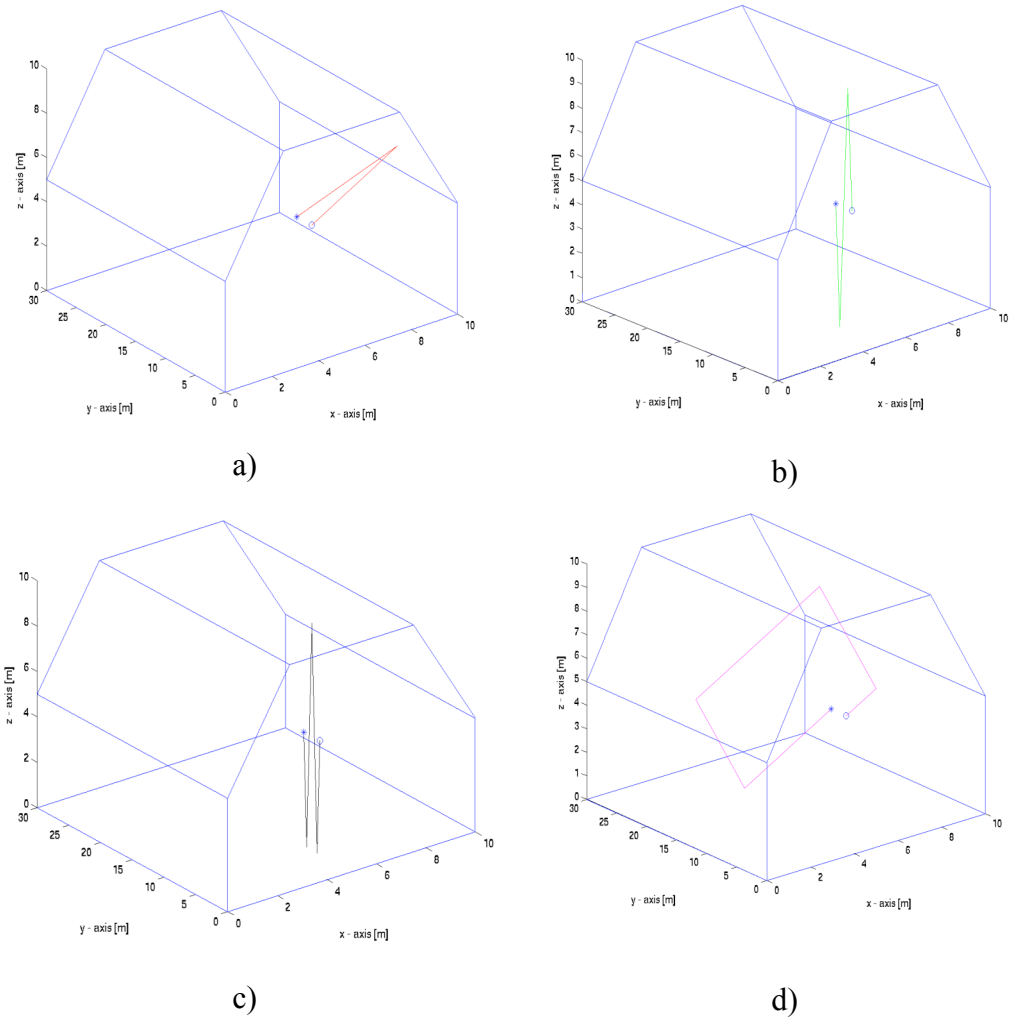


Figure 3.15 Examples of first a), second b), third c) and fourth 4) order reflections for a simple room.

Figure 3.13 shows some examples of first, second, third and fourth order reflections for a simple room. The position of the listener is $\begin{bmatrix}5 & 5 & 5\end{bmatrix}^T$ and marked with a circle; the position of the source is $\begin{bmatrix}5 & 7.5 & 5\end{bmatrix}^T$ and marked with a star. For every single reflection, the direction of arrival to the listener and distance from the source to the listener is calculated.

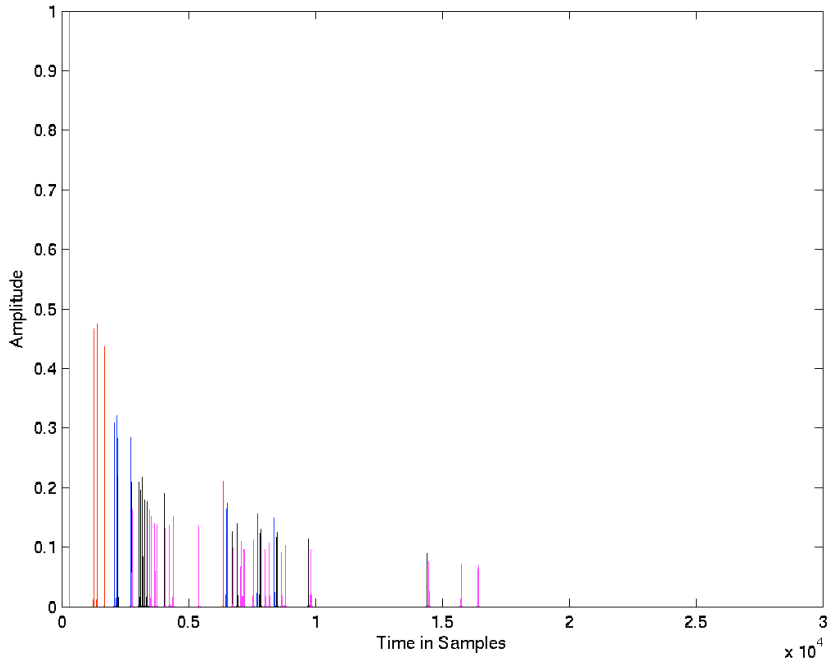


Figure 3.16 Reflectogram

Figure 3.16 shows the reflectogram of the room. The 'x' axis is pictured in samples with a sampling rate of 44100 Hz and the y axis is normalized to the direct sound.
The direct sound is pictured green, the first reflections red, the second ones blue, the third reflections black and the fourth magenta.

First the mirror sources S' of the source $S = \begin{bmatrix}x_S & y_S & z_S\end{bmatrix}^T$ are calculated using the plane equation E:

$$A \cdot x_E + B \cdot y_E + C \cdot z_E + D = 0 \tag{3.5}$$

and the plumb $\vec{x}$ and the intersection point L with the plane E:

$$\vec{x} = \begin{pmatrix} x_E \\ y_E \\ z_E \end{pmatrix} + t \cdot \begin{pmatrix} x_S \\ y_S \\ z_S \end{pmatrix} \tag{3.6}$$

$$L = \vec{x} \cap E \tag{3.7}$$

we obtain:

$$\vec{s}' = \vec{s} + 2(\vec{l} - \vec{s}) \tag{3.8}$$

Then a direct path between the mirror source and the listener position is drawn (eq. 3.9) and it is tested if there is a point of intersection of the line and the plane (eq. 3.10).

$$\vec{x}_{S-S'} = \begin{pmatrix} x_S \\ y_S \\ z_S \end{pmatrix} + t \cdot \left[\begin{pmatrix} x_{S'} \\ y_{S'} \\ z_{S'} \end{pmatrix} - \begin{pmatrix} x_S \\ y_S \\ z_S \end{pmatrix} \right] \tag{3.9}$$

$$L = \vec{x}_{S-S'} \cap E \tag{3.10}$$

If there is a point of intersection it is tested if this point is located inside the plane defined by the vertices of the wall. This is implemented with angle functions to ensure that every wall-shape can be processed.

To proof if the path is an allowed one it is tested if there are other points of intersection with the reflected ray. If there is no other point of intersection with other walls, the path is an allowed one.

Thus not just simple rooms can be calculated but even more complex environments like the cathedral shown in figure 3.17.

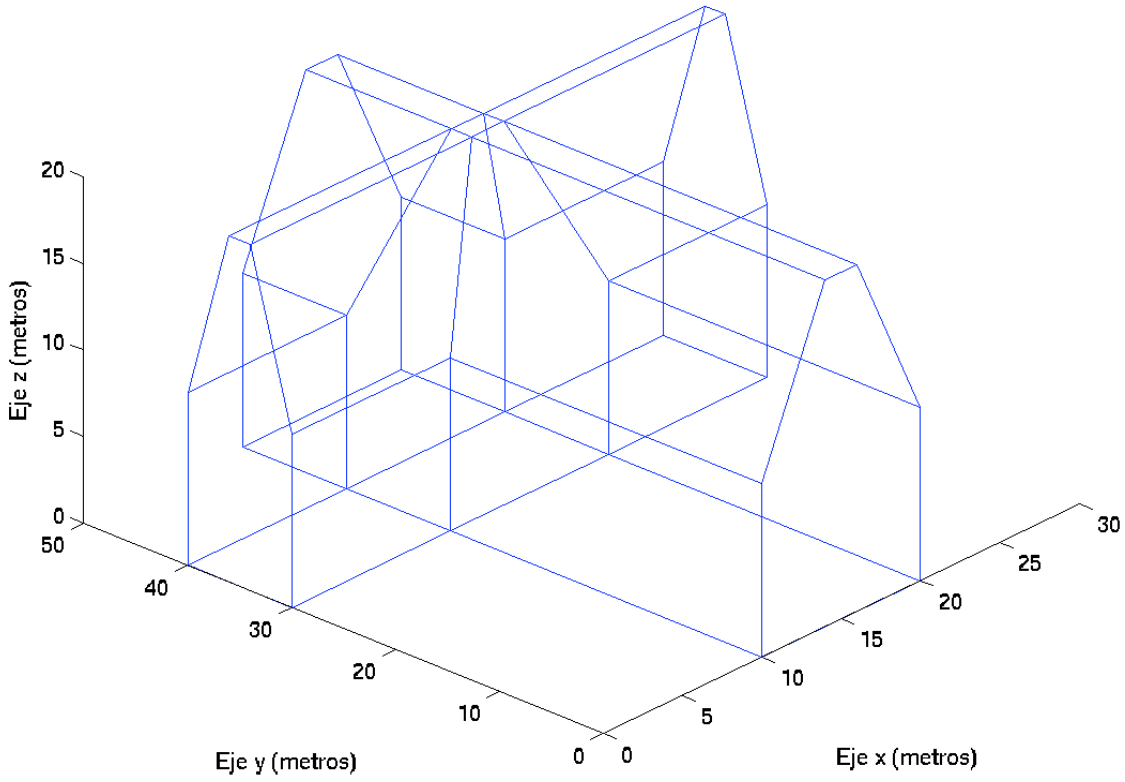


Figure 3.17 A more complex environment

The perceptual effects of early reflections can be studied by considering a simple soundfield consisting of the direct sound and a single delayed reflection. If the reflection delay is small (< 80 ms), the reflection and the direct sound fuse into one sound, but with a tonal coloration that causes the cancellation of a periodic set of frequencies. When the reflection comes from the lateral direction, the reflection can affect the spatial character of the sound. Small reflection delays (< 5 ms) can cause an apparent location shift, whereas larger delays can increase the size of the source, depending on the frequency content, or can create the sensation of being surrounded by sound.

The level of the direct sound relative to the reverberation changes as a function of source distance, and serves as an important distance cue.

Figure 3.18 shows a table of the first 14 reflections with their delay, direction, amplitude and the surface of the reflection. The listener position in meters is $L = [5 \quad 2 \quad 5]^T$ and the source position in meters is $S = [1 \quad 7 \quad 5]^T$. The roomsize is $RS = 10m \times 30m \times 10m$. The floor and roof material is wood and the wall material is tapestry (0.35 kg/m2).

Delay [ms]	Elevation [degrees]	Azimuth [degrees]	Amplitude	Surface
18.18	8	200	0.11	wall
19.16	0	198	0.089	wall
33.2	66	207	0.05	roof
33.2	-66	207	0.05	floor
35.9	-58	198	0.004	floor, wall
38.3	0	108	0.03	wall
38.8	34	349	0.04	roof
40.2	3	115	0.0031	wall, roof
40.7	0	117	0.0026	wall, wall
42.8	39	349	0.0034	roof, roof
42.9	0	352	0.029	wall
48.9	38	108	0.0024	wall, roof
48.9	-38	108	0.0024	wall, floor
52.5	-35	352	0.0021	wall, floor

Figure 3.18 Table of the first 14 reflections with their delay, direction, amplitude and the surface of the reflection

3.1.5 Late reverberation as a linear filter

As it is mentioned above, the source image method is computationally expensive and rather inflexible for studying late reverberation, because the number of sources increases exponentially with the number of order.

That is why late reverberation is realized by designing a digital filter. The first artificial reverberators based on discrete time signal processing were constructed by Schroeder in the early 1960's. Schroeder's original proposal was based on comb and allpass filters. The comb filter is shown in figure 3.19 and consists of a delay whose output is recirculated to the input. The z transform of the comb filter is given by:

$$H(z) = \frac{z^{-m}}{1 - gz^{-m}} \tag{3.11}$$

where *m* is the length of the delay in samples and g is the feedback in gain. The time response of this filter is an exponentially decaying sequence of impulses spaced *m* samples apart.

The system poles occur at the complex *m*th roots of g, and are thus harmonically spaced on a circle in the z plane. The frequency response is therefore shaped like a comb, with *m* periodic peaks that correspond to the pole frequencies.

Schroeder determined that the comb filter could be easily modified to provide a flat frequency response by mixing the input signal and the comb filter output as shown in figure 3.20. The resulting filter is called allpass filter because its frequency response has unit magnitude for all frequencies. The z transform of the allpass filter is given by:

$$H(z) = \frac{z^{-m} - g}{1 - gz^{-m}} \tag{3.12}$$

The poles of the allpass filter are thus the same as for the comb filter, but the allpass filter now has zeros at the conjugate reciprocal locations. The frequency response of the allpass filter can be written:

$$H(e^{j\Omega}) = e^{j\Omega m} \frac{1 - ge^{+j\Omega m}}{1 - ge^{-j\Omega m}} \tag{3.13}$$

In this form it is easy to see that the magnitude response is unity, because the first term in the product, e^{-jwm} has a unit magnitude, and the second term is a quotient of the complex conjugates, which also has unit gains. Thus

$$\left|H(e^{j\Omega})\right| = 1 \tag{3.14}$$

The Phase response of the allpass filter is a non linear function of frequency, leading to a smearing of the signal in the time domain.

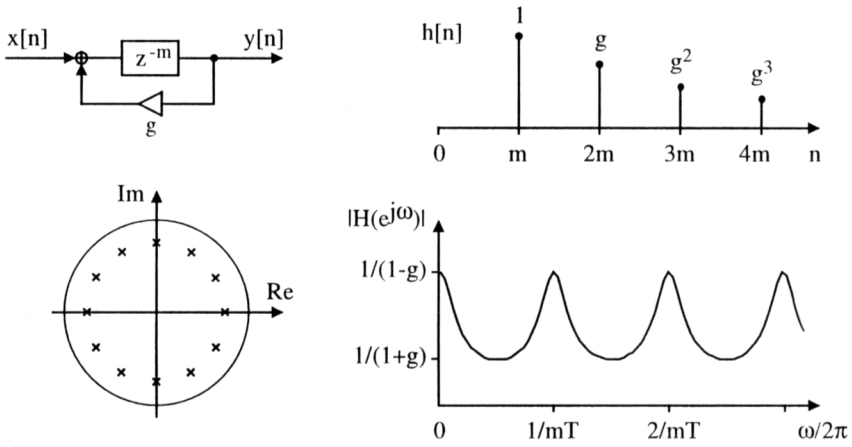


Figure 3.19 Comb filter: flow diagram, time response, frequency response, and pole diagram (clockwise from top-left) [3]

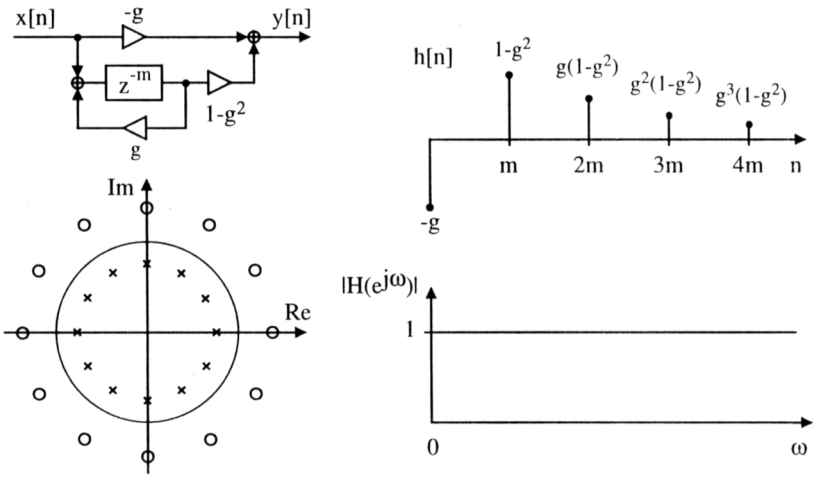


Figure 3.20 Allpass filter formed by modification of a comb filter: flow diagram, time response, frequency response, and pole-zero diagrams (clockwise from top-left) [3]

For short delay times, which yield rapidly occurring echoes, the frequency response is characterized by widely spaced frequency peaks. These peaks correspond to the frequencies that will be reverberated, whereas frequencies falling between the peaks will decay quickly. When the peaks are widely spaced, the comb filter has a noticeable and unpleasant characteristic timbre. We can increase the density of peaks by increasing the delay length, but this causes the echo density to decrease in the time domain. Consequently, the reverberation is heard as a discrete set of echoes, rather than a smooth diffuse decay.

An allpass filter has a flat magnitude response, and we might expect it to solve the problem of timbral coloration attributed to the comb filter. However, the response of an allpass filter sounds quite similar to the comb filter, tending to create a timbral coloration. This is because our ears perform a short time frequency analysis, whereas the mathematical property of the allpass filter is defined for infinite time integration.

By combining two elementary filters in series, we can dramatically increase the echo density, because every echo generated by the first filter will create a set of echoes in the second. Comb filters are not good candidates for series connection, because the only frequencies that will pass are those that correspond to peaks in both comb filter responses. However, any numbers of allpass filters can be connected in series and the combined response will still be allpass. Consequently, series allpass filters are useful for increasing echo density without affecting the magnitude response of the system.

A parallel combination of comb filters with incommensurate delays is also a useful structure, because the resulting frequency response contains peaks contributed by all of the individual comb filters. Moreover, the combined echo density is the sum of the individual densities. Thus, we can theoretically obtain arbitrary density of frequency peaks and time echoes by combining a sufficient number of comb filters in parallel [3].

Schroeder proposed a reverberator consisting of parallel comb filters and series allpass filters, shown in figure 3.21. The delays of the comb filters are chosen such that the ratio of the largest to the smallest is about 1.5 (Schroeder suggested a range of 30 to 45 msec).

The allpass delays are t_5 and t_6 are much shorter than the comb delays, perhaps 5 and 1.7 msec, with both allpass gains set to around 0.7. Consequently, the comb filters produce the long reverberant decay, and the allpass filters multiply the number of echoes output by the comb filters. Figure 3.22 shows the impulse response of Schroeder's reverberator.

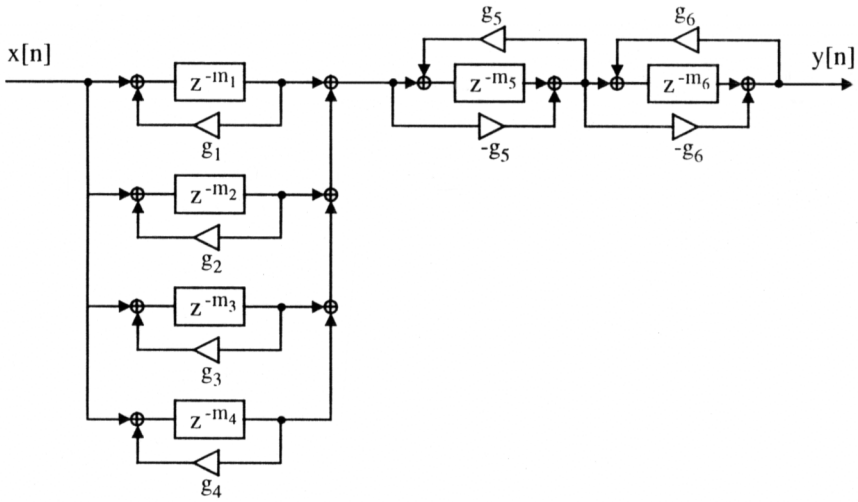


Figure 3.21 Schroeder's reverberator consisting of a parallel comb filter and a series allpass filters [3]

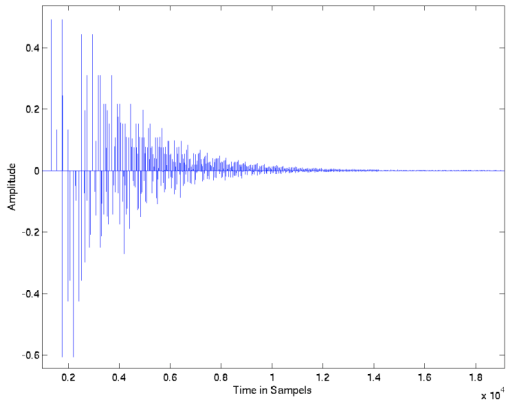


Figure 3.22 Impulse response of the Late Reverberation Filter of Schroeder

Another reverberator was proposed by Stautner and Puckette in 1982, shown in figure 3.23. They designed a four channel reverberator consisting four delay lines with a feedback matrix. The feedback matrix allows the output of each delay to be recirculated to each delay input, with the matrix coefficients controlling the weights of these feedback paths. The structure can be seen as generalization of Schroeder's parallel comb filters, which would arise using a diagonal feedback matrix. This structure is capable of much higher echo densities than parallel comb filter, given a sufficient number of nonzero feedback coefficients and incommmensurate delay lengths [3]. The delays were chosen in accordance with Schroeder's suggestions.

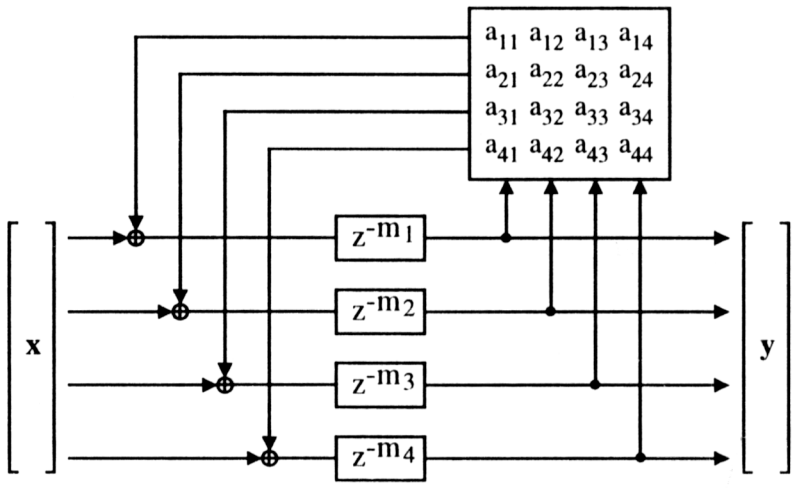


Figure 3.23 Stautner and Puckette's four-channel feedback delay network [3]

Stautner and Puckette make a number of important points regarding this system:

• Stability is guaranteed if the feedback matrix A is chosen to be the product of a unitary matrix and a gain coefficient g, where $|g| < 1$. They suggested the matrix

$$A = g\frac{1}{\sqrt{2}}\begin{bmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & -1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 1 \end{bmatrix} \tag{3.15}$$

where g controls the reverberation time. If $|g| = 1$, A is unitary.

• The outputs will be mutually incoherent, and thus can be used in a four-channel loudspeaker system to render a diffuse soundfield.

• Absorptive losses can be simulated by placing a lowpass filter in series with each delay line.

• The early reverberant response can be customized by injecting the input signal appropriately into the interior of the delay lines.

The authors, Kahrs M. and Brandenburg [3] mentioned that fluttering and tonal coloration is present in the late decay of this reverberator. To reduce the tonal coloration, they suggest randomly varying the length of the delays.

We now discuss the recent and important work by Jot, who has proposed a reverberator structure with two important properties.

• A reverberator can be designed with arbitrary time and frequency density while simultaneously guaranteeing absence of tonal coloration in the late decay.

• The resulting reverberator can be specified in terms of the desired reverberation time $T_r(w)$ and frequency response envelope $G(w)$.

This is accomplished by starting with an energy conserving system whose impulse response is perceptually equivalent to stationary white noise. Jot calls this a reference filter, but we will also use the term lossless prototype. Jot chooses prototypes from the class of unitary feedback system. In order to effect a frequency dependent reverberation time, absorptive filters are associated with each delay of the system. This is done in a way that eliminates coloration in the late response, by guaranteeing the local uniformity of pole modulus.
Jot generalizes the notion of a monophonic reverberator using the feedback delay network (FDN) structure shown in figure 3.24. The structure is a completely general specification of a linear system containing N delays [3].
Using vector notation and the z transform the equations for the output of the system $y(z)$ and the delay lines $s_i(z)$ are:

$$y(z) = c^T s(z) + dx(z) \tag{3.16}$$

$$s(z) = D(z)\left[As(z) + bx(z)\right] \tag{3.17}$$

where:

$$s(z) = \begin{bmatrix} s_1(z) \\ \vdots \\ s_N(z) \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix} \quad c = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \tag{3.18}$$

$$D(z) = \begin{bmatrix} z^{-m_1} & & 0 \\ & \ddots & \\ 0 & & z^{-m_N} \end{bmatrix} \quad A = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix} \tag{3.19}$$

The FDN can be extended to multiple inputs and outputs by replacing the vectors b and c with appropriate matrices. The system transfer function is obtained by eliminating s(z) from preceding equations:

$$H(z) = \frac{y(z)}{x(z)} = c^T \left[D(z^{-1}) - A \right]^{-1} b + d \tag{3.20}$$

The system zeros are given by

$$\det\left[A - \frac{bc^T}{d} - D\left(z^{-1}\right) \right] = 0 \tag{3.21}$$

The system poles are given by those values of z that nullify the denominator of equation 3.11, in other words the solutions to the characteristic equation:

$$\det\left[A - D(z^{-1}) \right] = 0 \tag{3.22}$$

Assuming A is a real matrix, the solutions to the characteristic equation 3.20 will either be real or complex conjugate pole pairs. Equation 3.20 is not easy to solve in general case, but for specific choices of A the solution is straightforward. For instance, when A is diagonal, the system represents Schroeder's parallel comb filter, and the poles are given by:

$$\prod_{i=1}^{N} \left(a_i - z^{m_i} \right) = 0 \tag{3.23}$$

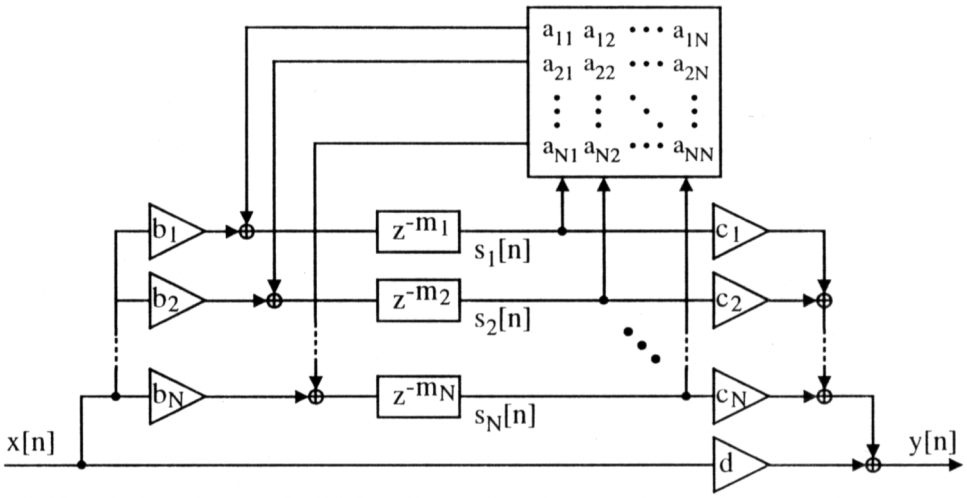


Figure 3.24 Feedback delay network as a general specification of a reverberator containing N delays [3]

The FDN (Feedback Delay Network) shown in figure 3.24 is used to simulate a monoaural Reverb. The disadvantage of a monoaural Reverb is the in head localization. Therefore two different FDNs, one for the right ear and one for the left ear have to be designed. This is implemented with a permutation matrix M that mutates the Feedback matrix and therefore the impulse response of the system. In this system M is chosen to be:

$$M = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{3.24}$$

The permutation matrix M is multiplied with the feedback matrix A. Therefore, using the permutation matrix M (eq. 3.24), the first feedback channel is permuted with the fourth and the second channel with the third like shown in figure 3.25. Figure 3.26 shows the two different impulse responses of the permuted FDN and the original one for the left and the right ear.

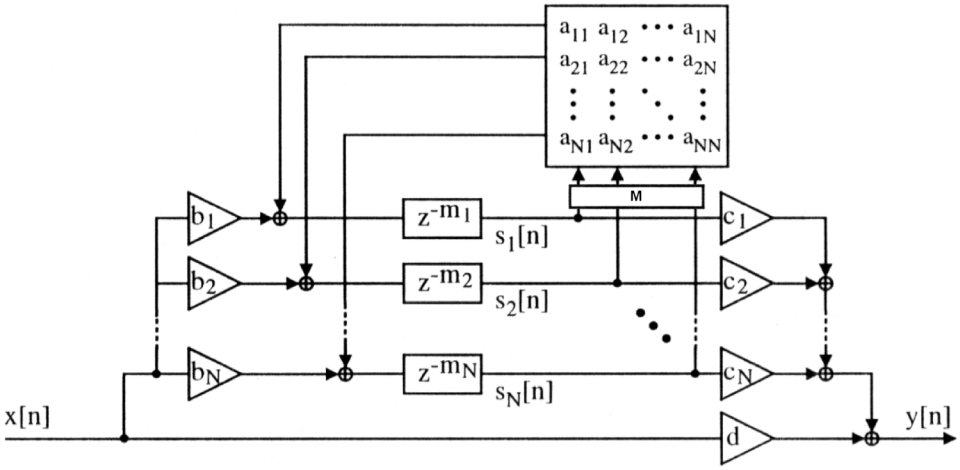


Figure 3.25 Feedback delay network as a general specification of a reverberator containing N delays and a permutation matrix M

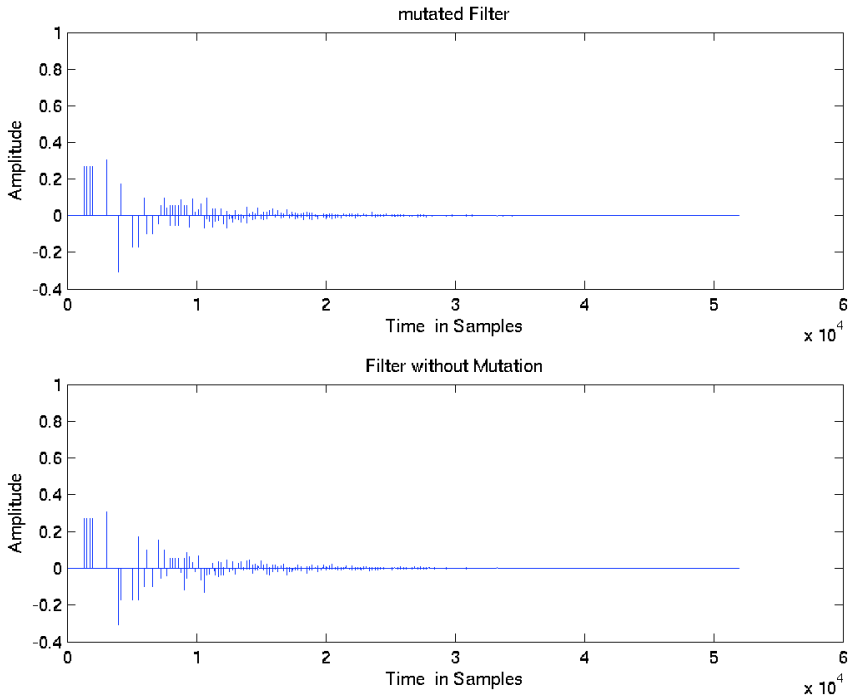


Figure 3.26 Impulse Response of the permuted FDN and the original FDN of Jot

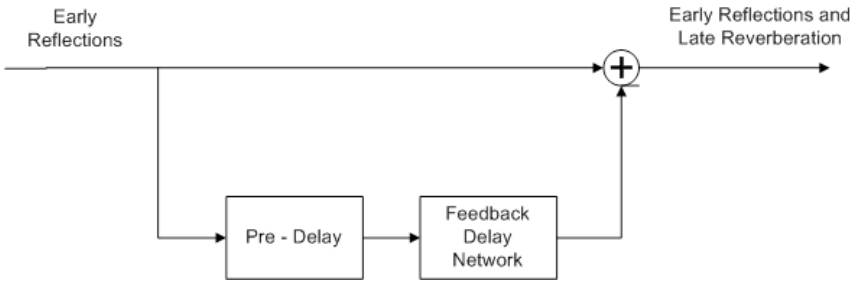


Figure 3.27 Simplified signal flow of the system

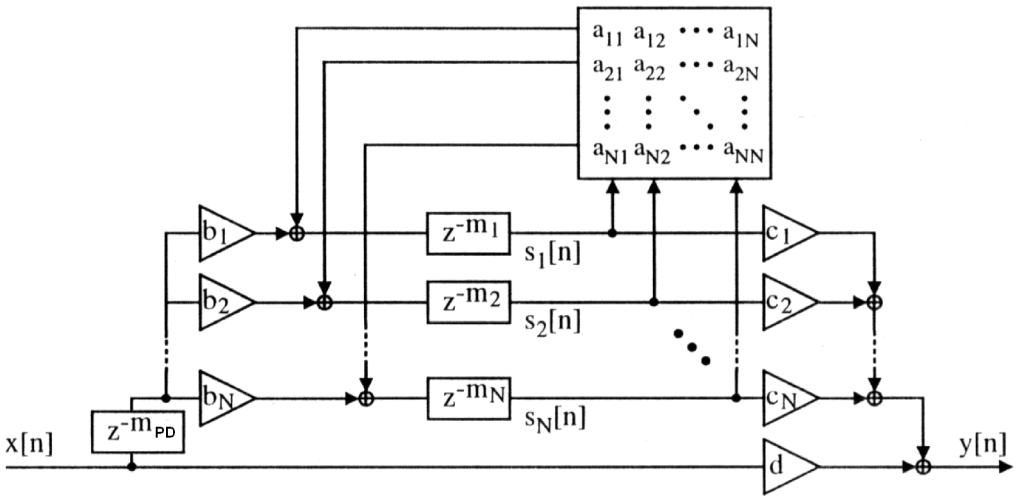


Figure 3.28 FDN with Predelay

As shown in figure 3.27, the input signals of the Feedback Delay Network are the early Reflections. To model the late reverberation at the right position in the reflectogram, the FDN shown in figure 3.24 has to be predelayed as it is shown in figure 3.28.

Figure 3.29 shows the pre-delayed impulse response of the original and the permuted FDN. The predelay was set to be 10000 samples.

The entire reflectogram is shown in figure 3.30. If we compare it to figure 3.16 we can se that the late reverberation was modelled onto it, exactly after the last 4^{th} order reflection.

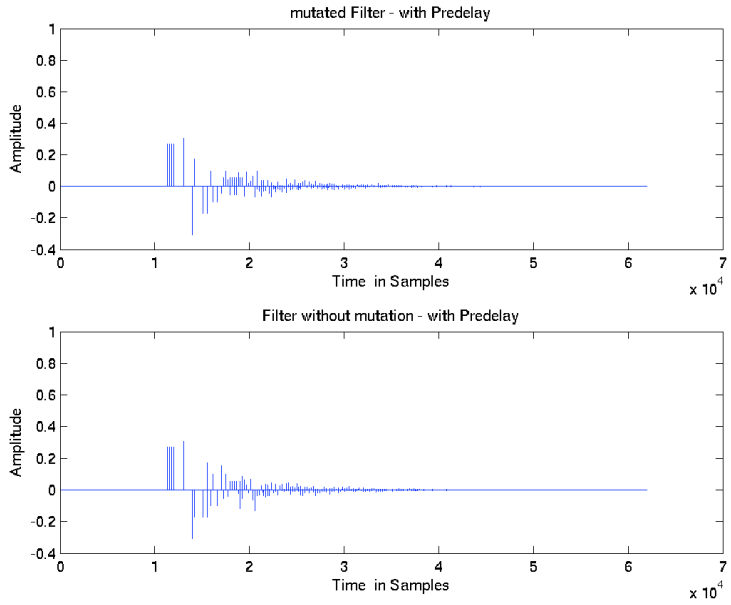


Figure 3.29 Pre-delayed impulse response of the original and the permuted FDN

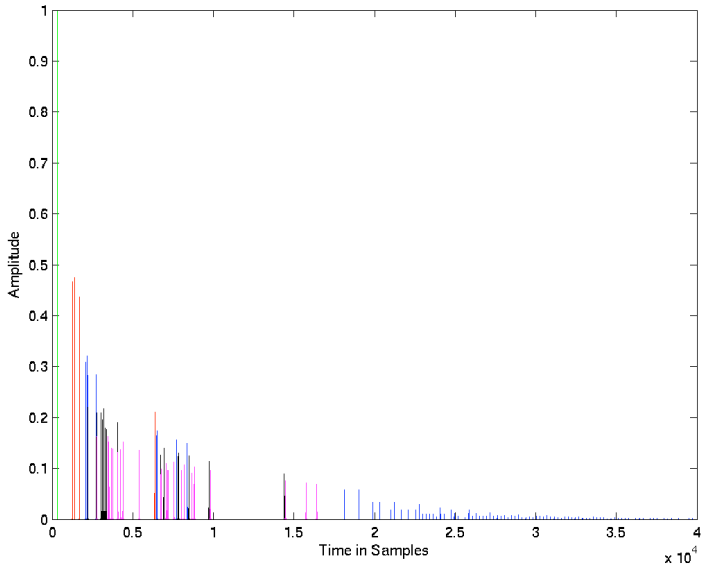


Figure 3.30 Reflectogram with late reverberation

In this system the FDN of Jot is used because it sounds much more realistic than the reverberator of Schroeder. Moreover it is easier to control the reverberation time that should be varied with the size of the chamber.

To make it possible to control the reverberation time, the reverberation time by Sabine is calculated:

$$R_{T60} = 0.16 \cdot \frac{V}{A} \tag{3.25}$$

with the volume (V) of the enclosure and the total surface absorption

$$A = S \cdot \alpha \tag{3.26}$$

where S is the sum of all the surface areas in the room and α are their respective absorption coefficients.

Then the control parameter g is set empirically and the reverberation time of the room is calculated from the Schroeder integration curve (backward integration of the squared impulse response). After setting the control parameter g, the calculated reverberation time and the empirically optained reverberation time should be the same.

Chapter 4
Modelling and Interpolation of HRTFs

4.1 Head-Related Transfer Function Modelling

4.1.1 Common Acoustical Pole and Zero (C.A.P.Z) Modelling

There are many ways to model head-related transfer functions. An all-zero model is a typical HRTF model. Strictly speaking, the n-th order all-zero model has implicitly n poles at z = 0. It simply uses the impulse response of the measured HRTF as its coefficients and can be constructed using a finite impulse response (FIR) filter. When using the all-zero model, however, many parameters (filter coefficients) are needed to simulate HRTF's corresponding to various source directions. This is because a low-order-zero model (low-order FIR filter) is not able to represent steep frequency structures [16]. Furthermore, the impulse response of a HRTF strongly depends on the source direction. So, many FIR filter coefficients for every source direction must be utilized.

To reduce the number of parameters, pole-zero modelling has been studied by different authors like M. A. Blommer and G. H. Wakefield. The pole-zero model can be constructed using an infinite impulse response (IIR) filter. An IIR filter uses more instructions when implemented on digital signal processors (DSP's) and is more sensitive to numerical errors in the fixed-point arithmetic than a same order FIR filter. Nevertheless, the model is more efficient than the all-zero model, because the poles can represent the long impulse response caused by resonances with fewer parameters.

In conventional pole-zero modelling, however, because both the poles and the zeros are estimated for every source direction of the HRTF, the estimated poles depend on the source direction, even though the physical poles of the HRTF do not. Therefore, a different set of poles and zeros is used to represent the HRTF for each different source direction.

A more efficient way of modelling acoustic transfer functions is the common acoustical pole and zero (CAPZ) model. This model expresses the acoustic transfer function by using the common acoustical poles, which are independent of the source and receiver positions, and the zeros, which depend on those positions. An HRTF represented by the CAPZ model has two parts: a direction-independent part (common acoustical poles) and a direction-dependent part (zeros) [16]. The common acoustical poles are estimated as values common to the measured HRTF's for the various source directions – they correspond to the physical resonance system of an ear canal.

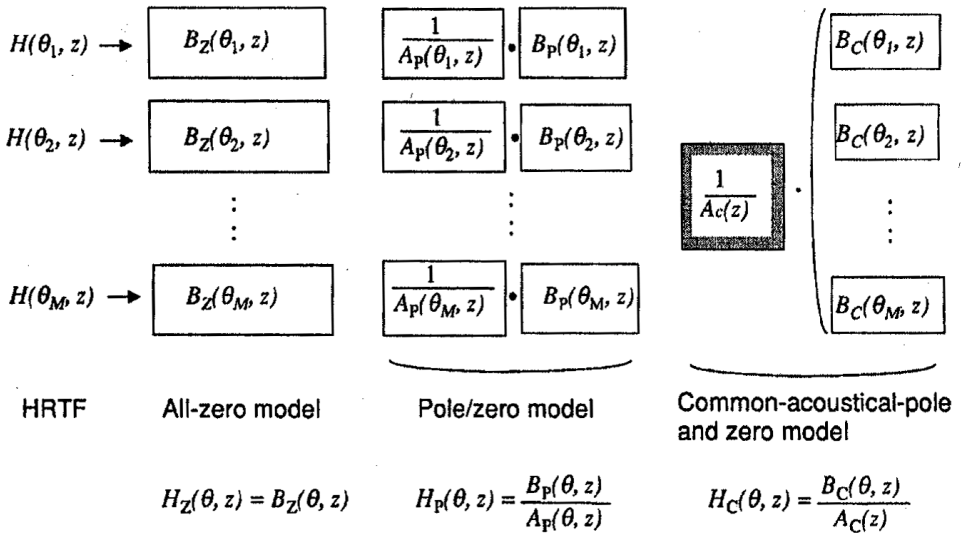


Figure 4.1 Differences between models [16]

Because the CAPZ model represents the directional dependence of the HRTF using only the zero variations, it requires fewer parameters than the conventional all-zero model or the pole-zero model. Furthermore, it can extract even the zeros that are missed in conventional models due to pole zero cancellation. This means that the CAPZ model can well trace the zero variations due to changes in the source direction and thus could offer a new characterization method for HRTF variations due to changes in the source direction.

An HRTF contains a resonance system composed of an ear canal. The resonance frequencies and Q factor of this system can be assumed to be independent of the source direction [16]. The poles of the HRTF represent these resonance frequencies and Q factors, and are therefore independent on source directions. These directional-independent poles are called common-acoustical-poles, because they are commonly included in the HRTF for any source direction.
When the HRTF $H(\theta,z)$ is modelled using the CAPZ model, the poles are independent of the source direction θ, but the zeros depend on it. Therefore, the CAPZ model of HRTF $H(\theta,z)$ is expressed as

$$H(\theta,z)=\frac{B(\theta,z)}{A(\theta,z)}=\frac{C\,z^{-Q_1}\prod_{i=1}^{Q_2}\left[1-q_i(\theta)\,z^{-1}\right]}{\prod_{i=1}^{P}\left(1-p_{ci}z^{-1}\right)}=\frac{\sum_{i=0}^{Q}b_i(\theta)\,z^{-i}}{1-\sum_{i=0}^{P}a_{ci}\,z^{-i}} \qquad (4.1)$$

where p_{ci} denotes the common acoustical poles independent of source direction θ, and $q_i(\theta)$ denotes the zeros dependent on the source direction θ. The P and Q (Q = $Q_1 + Q_2$) are the orders of the poles and of the zeros. The coefficients a_{ci} (i = 1,.....,P) denote the common autoregressive (AR) coefficients corresponding to the common acoustical poles p_{ci}, and the coefficients $b_{i(\theta)}$ (i = 0,.....Q) denote the moving-average (MA) coefficients corresponding to the zeros $q_i(\theta)$.
Figure 4.1 compares the all-zero model, the pole-zero model and the CAPZ model, for modelling the multiple HRTF's corresponding to M source directions. Because all parameters of the all-zero model depend on the source direction θ, a completely different set of parameters is required for each source direction. Therefore, the all-zero model requires many parameters to represent HRTF's for various source directions.

The pole-zero model divides an HRTF $H(\theta,z)$ into its pole function $A_P(\theta,z)$ and its zero function $B_P(\theta,z)$. Both the pole $A_P(\theta,z)$ and the zero function $B_P(\theta,z)$ are estimated for each source direction of the HRTF $H(\theta,z)$. As a result, although the poles in the HRTF are physically invariant for source direction θ, the estimated poles $A_P(\theta_m,z)$ (m = 1,2,....,M) vary with the source direction θ_m because of the interference from the direction dependent zeros. Therefore, both poles in $A_P(\theta_m,z)$ and zeros $B_P(\theta_m,z)$ have to be kept for all source directions ($\theta_1 to \theta_M$) to represent M-source direction HRTF's.

In contrast, the CAPZ model represents HRTF $H(\theta,z)$ by using the common-acoustical-pole function $A_C(z)$, which is independent of the source direction, and by the zero function $B_C(\theta,z)$, which depends on the source direction.

As the CAPZ model expresses the source directional dependence of the HRTF by using only the zero variations, the number of parameters that depend on the source direction is reduced. Another remarkable feature of the CAPZ model is that it can extract the zeros missing due to pole-zero cancellation [16].

The common acoustical poles are physically included in the HRTF for any source direction. They cannot be estimated using a single HRTF measured for an arbitrary source direction because they are usually strongly affected or cancelled by *-th* direction-dependent zeros. That is, the poles estimated using a single measured HRTF cannot be regarded as common poles. Therefore, common acoustical poles are estimated by using an entire set of HRTF's measured for different source directions.

Practically, the common acoustical poles are estimated as the common AR coefficients a_{ci} using HRTF's for several source directions. Equation 4.1 can be modified as

$$H(\theta,z)\cdot\left[1-\sum_{i=1}^{P} a_{ci}\, z^{-i}\right]=\sum_{i=0}^{Q} b_i(\theta)\, z^{-i} \tag{4.2}$$

Taking the inverse z transform of equation 4.2, the impulse response $h_C(\theta,k)$ of the CAPZ model $H(\theta,z)$ can be described in the time domain as

$$h_c(\theta,z)=\sum_{i=1}^{P} a_{ci}\, h_c(\theta,k-i)+\sum_{i=0}^{Q} b_i(\theta)\,\delta(k-i) \tag{4.3}$$

where $\delta(k)$ is the unit pulse function.

The output error $e_{out}(\theta,k)$ between the impulse response $h(\theta,k)$ of the original (measured) HRTF $H(\theta,z)$ and the impulse response $h_c(\theta,z)$ of the CAPZ model is defined by

$$e_{out}(\theta,k) = h(\theta,k) - h_c(\theta,z) \tag{4.4}$$

$$= h(\theta,z) - \sum_{i=1}^{P} a_{ci}\, h_c(\theta,k-i) - \sum_{i=0}^{Q} b_i(\theta)\,\delta(k-i)$$

However, finding values of a_{ci} and b_i that minimize the mean square of the output error $e_{out}(\theta,k)$ is known to be difficult, so equation 4.5 $e_{eq}(\theta,k)$ is used:

$$e_{eq}(\theta,k) = h(\theta,z) - \sum_{i=1}^{P} a_{ci}\, h(\theta,k-i) - \sum_{i=0}^{Q} b_i(\theta)\,\delta(k-i) \tag{4.5}$$

The common AR coefficients are determined so as to minimize cost function J_{eq}, which is defined as the square sum for time index k, and source direction index m:

$$J_{eq} = \sum_{m=0}^{M} \sum_{k=0}^{N+P} {e_{eq}}^2(\theta_m,k) \tag{4.6}$$

where M is the number of HRTF's and N is the length of the original impulse response $h(\theta,k)$.

The coefficients that minimize the cost function J_{eq} using the least squares method can be represented in vector form 4.7 [19]:

$$x = \left(A^T A\right)^{-1} A^T h_a, \tag{4.7}$$

with the formulation in vectors 4.19 to 4.24

$$x = \left[a^T, b^T(\theta_1), ..., b^T(\theta_M)\right]^T, \tag{4.8}$$

$$a = \left[a_1, a_2, ..., a_p\right]^T, \tag{4.9}$$

$$b(\theta_m) = \left[b_0(\theta_m), b_1(\theta_m), ..., b_Q(\theta_m)\right]^T, \tag{4.10}$$

$$h_a = \left[h^T(\theta_1), h^T(\theta_2), ..., h^T(\theta_M)\right]^T, \tag{4.11}$$

$$h(\theta_m) = \left[h_0(\theta_m), h_1(\theta_m), ..., h_{N-1}(\theta_m), 0, ..., 0\right]^T, \tag{4.12}$$

$$D = \begin{bmatrix} 1 & & & 0 \\ \cdot & \cdot & & \\ \cdot & & \cdot & \\ \cdot & & & \cdot \\ 0 & & & 1 \\ 0 & & & 0 \\ 0 & & & 0 \\ \cdot & \cdot & & \\ \cdot & & \cdot & \\ \cdot & & & \cdot \\ 0 & & & 0 \end{bmatrix} \tag{4.13}$$

$$A = \begin{bmatrix} H(\theta_1) & D & 0 & & & 0 \\ H(\theta_2) & 0 & D & & & \\ \cdot & & & \cdot & & \\ \cdot & & & & \cdot & \\ \cdot & & & & & \cdot \\ H(\theta_M) & 0 & 0 & & & D \end{bmatrix} \tag{4.14}$$

$$H(\theta_m) = \begin{bmatrix} 0 & 0 & & & 0 \\ h_0(\theta_m) & 0 & & & 0 \\ h_1(\theta_m) & h_0(\theta_m) & & & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & & \cdot \\ h_{P-1}(\theta_m) & h_{P-2}(\theta_m) & \cdot\ \cdot\ \cdot & & h_0(\theta_m) \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ h_{N-1}(\theta_m) & h_{N-2}(\theta_m) & \cdot\ \cdot\ \cdot & & h_{N-P}(\theta_m) \\ 0 & h_{N-1}(\theta_m) & \cdot\ \cdot\ \cdot & & h_{N-P-1}(\theta_m) \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & & \cdot \\ 0 & 0 & & & h_{N-1}(\theta_m) \end{bmatrix} \tag{4.15}$$

The orders P and Q are determined as follows. First output error index J_{out} , which is the average output error energy normalized by the impulse response energy, and which corresponds to the accuracy of the modelling, is defined as:

$$J_{out} = 10 \cdot \log_{10} \left[\frac{1}{M} \sum_{m=1}^{M} \frac{\sum_{k=0}^{N} e^2{}_{out}(\theta_m, k)}{\sum_{k=0}^{N} h^2(\theta_m, k)} \right] \tag{4.16}$$

Then, the desired value J_{out} is predetermined. Next, P and Q pairs are determined so as to minimize the sum of P and Q that satisfies the predetermined value of J_{out} .

The pole and zero orders P = 20 and Q = 40 in the CAPZ model were chosen to achieve a J_{out} of -10 dB. The length N of the impulse response of the original HRTFs (Kemar database) was 128. The common acoustical poles were estimated as the common AR coefficients a_{Ci} by using 12 HRTF's corresponding to the source direction at every 30˚ from 0˚ to 330˚. The common acoustical poles can be estimated from such a relatively small set of HRTF's because the poles cancelled by the zeros are different when the source directions are largely different.
Figure 4.2 shows the left and right HRTF (30˚ azimuth, 0˚ elevation) in the time and frequency domain. The original Kemar functions in this figure are plotted blue and the modelled ones are red. The Filter of the right ear is pictured in the upper line – the filter of the left ear in the bottom line.

In figure 4.2 the modeled curve and the original Kemar curve of 0˚azimuth/0˚elevation are compared in time and frequency domain. A mean squared error of $8.2 * 10^{-5}$ is detected.
However, not all of the dips are well traced in frequency responses. For example the 8-kHz dips between 95˚ and 110˚ cannot be traced very well as seen in figure 4.4, the left and the right modelled HRTF set as a function of time and frequency. This manifests in a source jump at exactly 95˚ -105˚ azimuth. As zeros create dips (notches) in the frequency response, the analysis of the zero variation must be useful. In figure 4.3 the modelling problem at 95˚ azimuth at the contralateral ear can be noticed. That can be explained by the longer delay that has to be modelled. Because the model has 40 zeros and the delay is 28 samples there are not enough zeros left to model the rest of the function. That can be detected easily in figure 4.3. The modelling error begins exactly at the 40th sample.

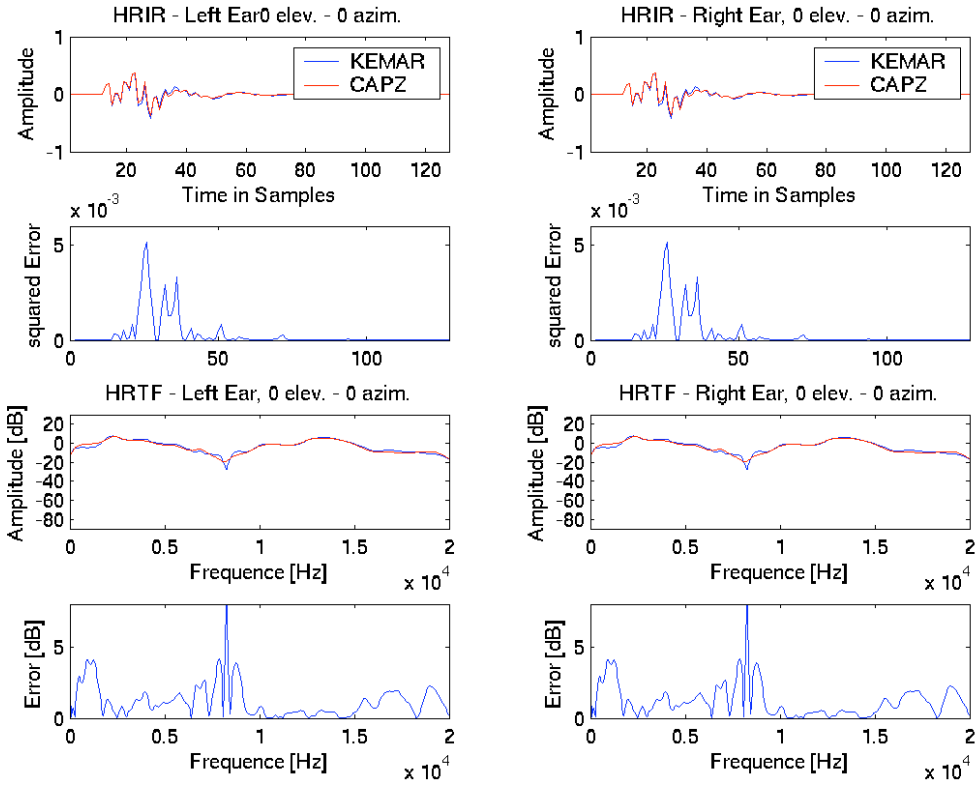


Figure 4.2 CAPZ Model compared to the desired Kemar curve of 0˚azimuth/0˚elevation - with 40 zeros -

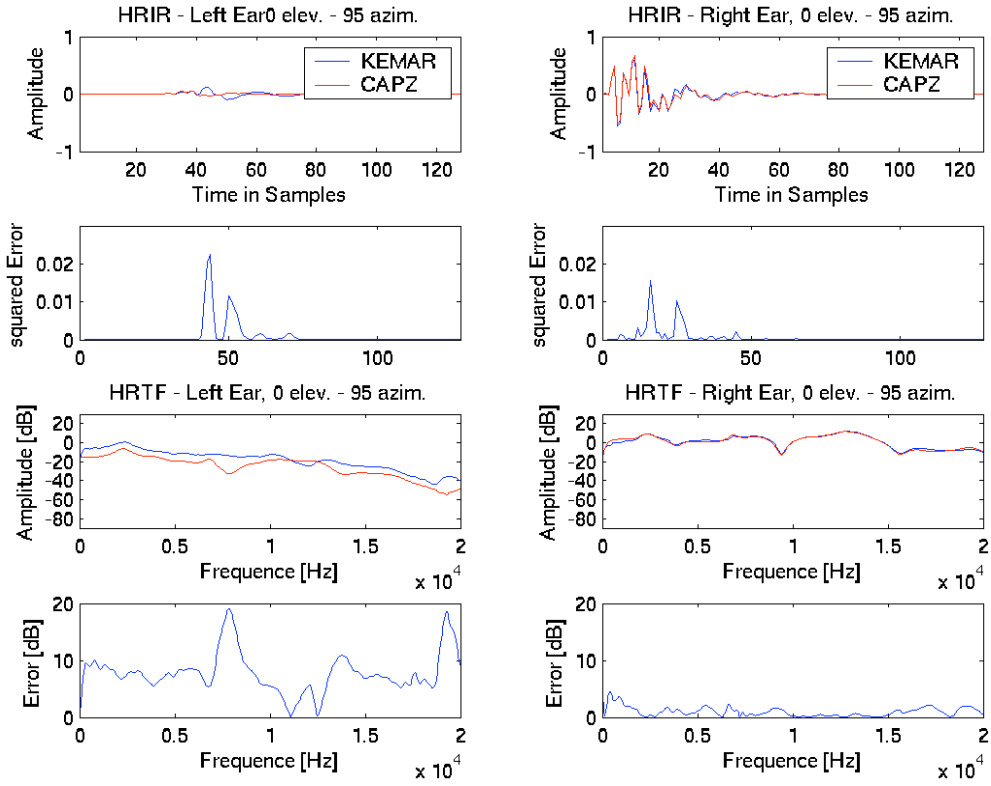


Figure 4.3 CAPZ Model compared to the desired Kemar curve of 95˚azimuth/0˚elevation - with 40 zeros -

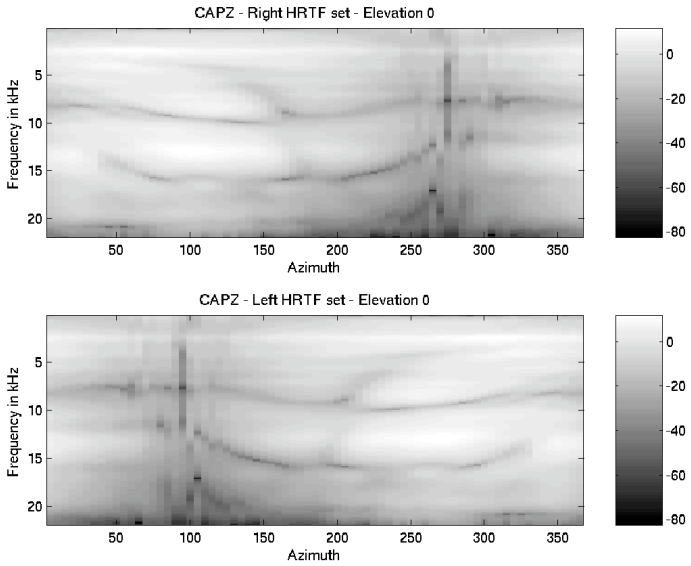


Figure 4.4 Modelled KEMAR set from 0˚-360˚ azimuth/0˚ elevation as a function of time and frequency

The first upgrade of the model is the increment of the number of zeros. The number is set to be 50 and the critical angle is observed, shown in figure 4.5 for 95˚azimuth and 0˚elevation.

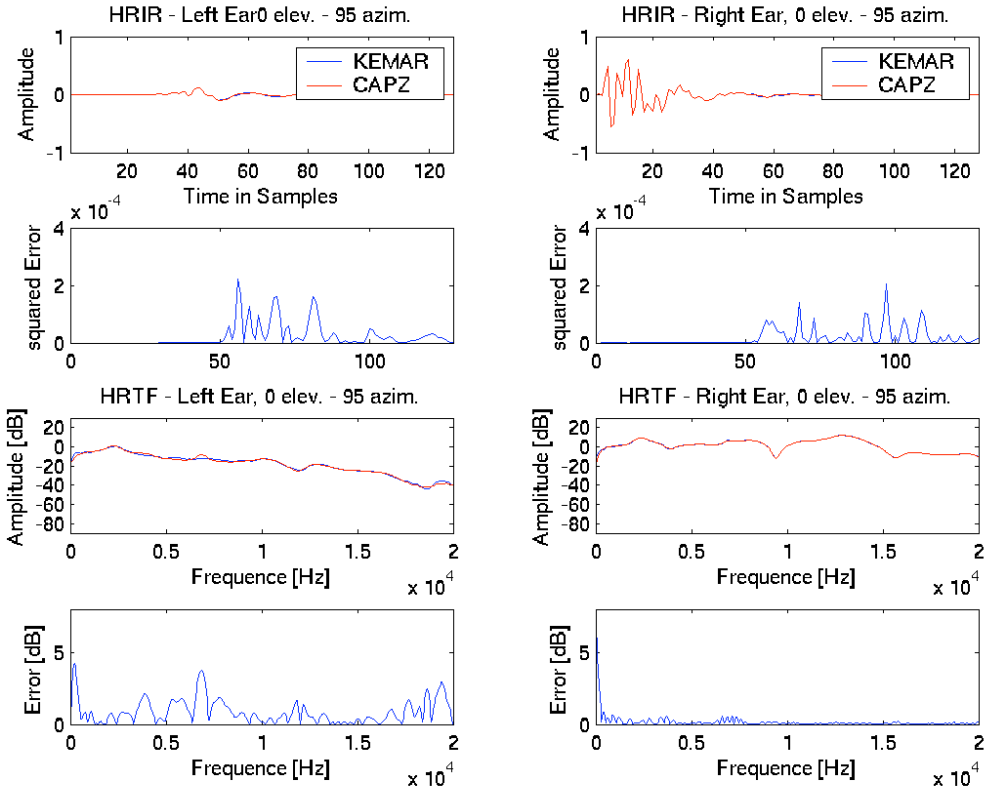


Figure 4.5 CAPZ Model compared to the desired Kemar curve of 95˚azimuth/0˚elevation - with 50 zeros –

It can be noticed that the squared error of the CAPZ curve modelled with 50 zeros and the desired KEMAR curve is 100 times smaller than the squared error of the CAPZ curve modelled with 40 zeros and the desired KEMAR curve.
However, the goal of modelling transfer functions is to minimize the number of coefficients more or less at the factor 10. The idea to reach this goal is to cut the delay of each HRIR, to model them and at last to zeropad every HRIR with the same number of samples that has been cut before.
The number of samples is calculated with the centroid of the energy of the bandbassfiltered impulse response. First, the impulse response is twice filtered with a FIR bandpass filter with 200 coefficients and a passband from 1 kHz to 4 kHz. One time foreward and the second time backward to provide a phase distortion. Then, the centroid of the energy of this filtered signal is calculated:

$$T(n,z,\theta) = \frac{\sum_n n \cdot {h_F}^2(n,z,\theta)}{\sum_n {h_F}^2(n,z,\theta)} \tag{4.17}$$

where ${h_F}^2$ is the power of the filtered HRIR and n is the sample number. The centroid of the energy of a signal has the same value as the mean group delay but the integral of the phase of a signal does not always have a defined value. That is why the centroid of the energy of a signal is calculated.
From the centroid of the energy of the filtered signal 17 samples are subtracted to ascertain to cut important information of the signal. The part from the first to the calculated value of the function is cut. The remaining, relevant part of the impulse response is modelled and at last the number of the cut samples is zeropadded. Figure 4.6 shows the result of this method.

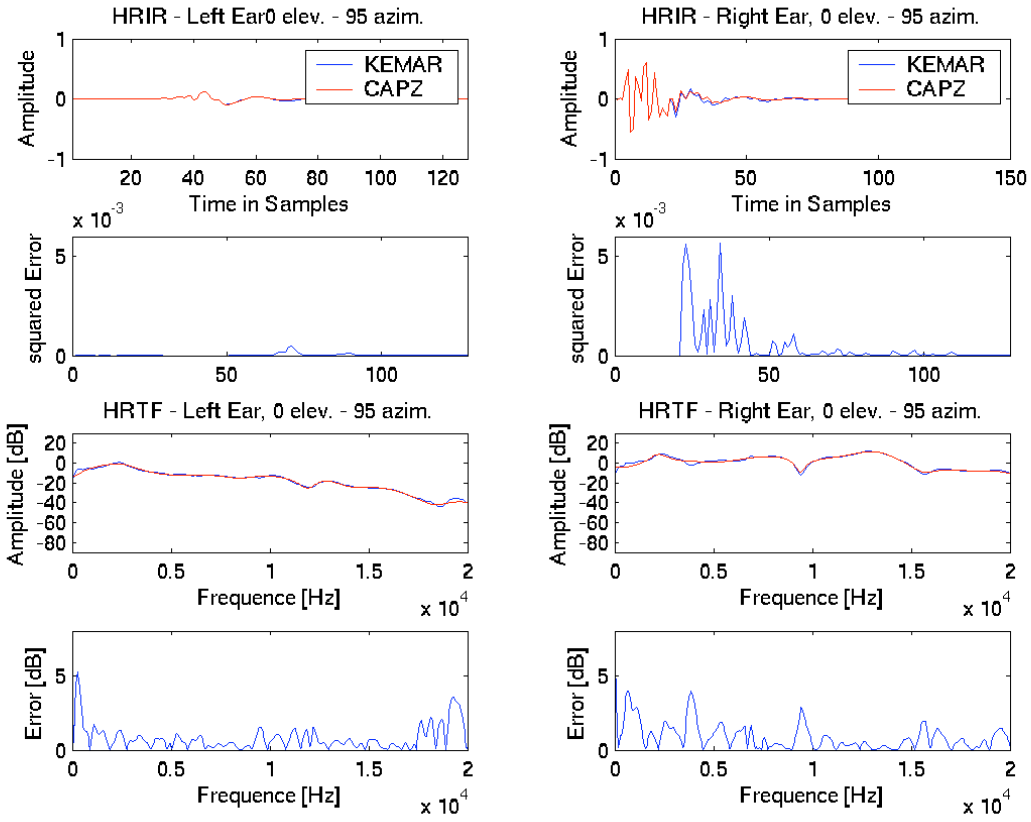


Figure 4.6 CAPZ Model compared to the desired Kemar curve of 95˚azimuth/0˚elevation - with 20 zeros and zeropadding -

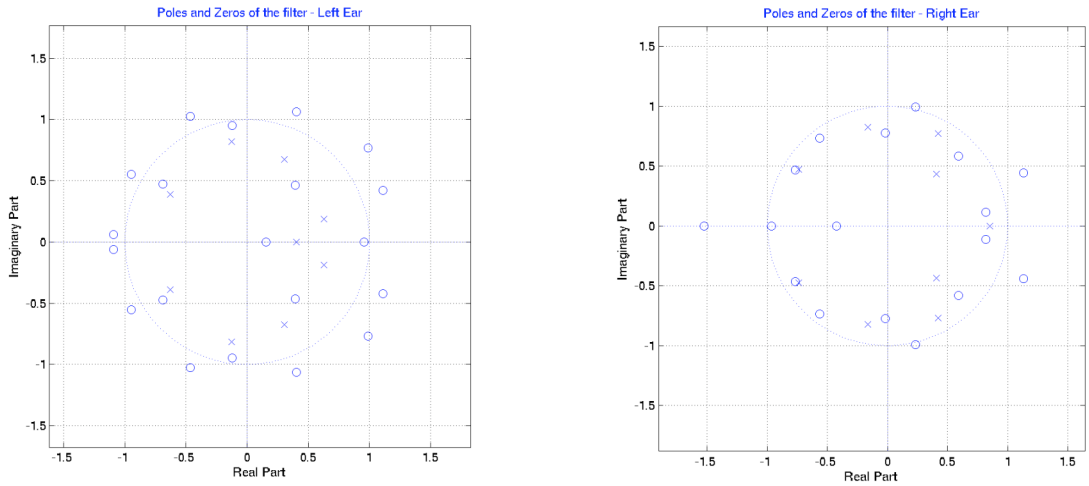


Figure 4.7 Poles and zeros of a CAPZ model, left and right ear (0˚ azimuth, 0˚ elevation)

In figure 4.7 the pole-zero diagram of the direction 0˚ azimuth/0˚ elevation is pictured. It can be seen that the relative source – listener distance of the left and the right ear is the same, because the zeros are not changing. As the poles in the diagram are inside the unit circle, the system is a stable one.
Poles and zeros are very close together and that is why there is no pole-zero cancellation. Therefore zero variation in function of directional variation can be studied more efficiently. Because there are zeros outside of the unit circle, this filter is called a nonminimum phase one. A nonminimum phase zero is generally generated when indirect sound, or reflected sound, which arrives after the direct sound, has greater energy than the direct sound. When the several reflected sounds arrive at a receiver in a room at the same time, the indirect sound has greater energy than the direct sound. In this case, nonminimum zeros are produced by the pinna reflections.

Figure 4.8 shows the Kemar HRTF set from 0˚ - 355˚ in 5˚ steps (72 HRTFs) as a function of time and frequency. In this plot the typical azimuth notch can be seen. Compared to figure 4.9, the modelled HRTF set, it can be seen that there are hardly any differences. That is why this model in connection to the method metioned above works real well.
The modelled Kemar set as a function of time and frequency is shown in figure 4.9. Compared to figure 4.4, the same plot but modelled with 40 zeros and without cutting off the delay, it can be perceived that at the angle 90˚-110˚ azimuth the plot is much more continuous. Moreover, also the source- jump in this region is solved.

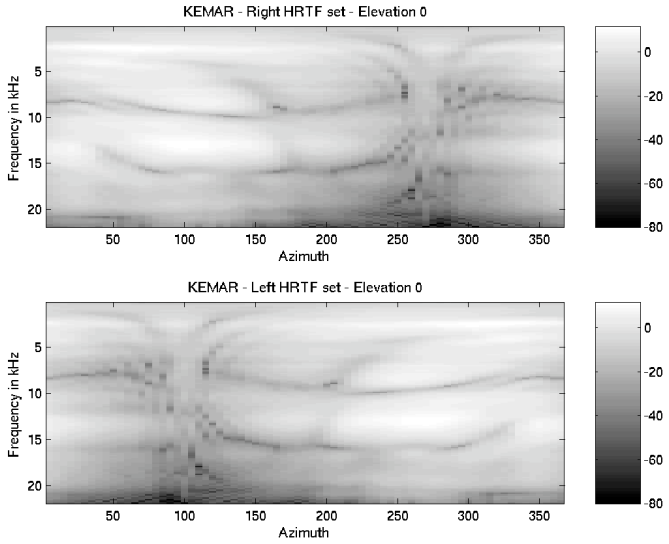


Figure 4.8 Kemar HRTF set (0˚ - 355˚ azimuth, 0˚ elevation)

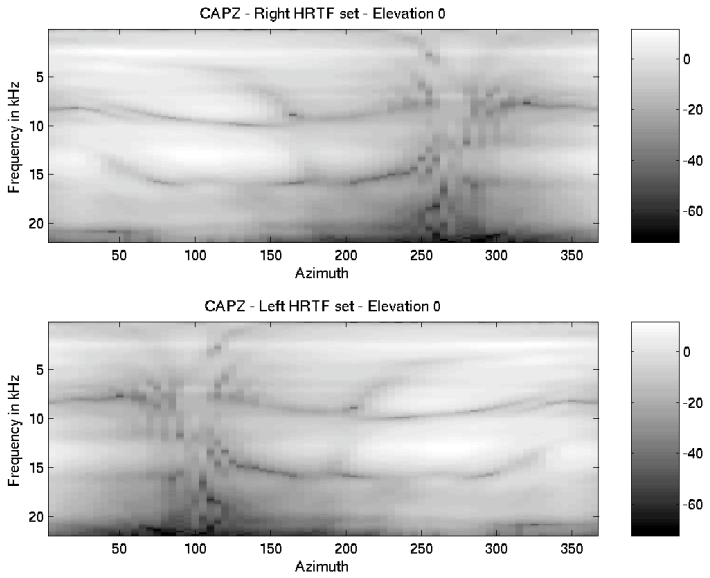


Figure 4.9 CAPZ HRTF set (0˚ - 355˚ azimuth, 0˚ elevation) as a function of time and frequency

4.2 Head-Related Transfer Function Interpolation

In practical systems only a finite number of HRTFs is available. They are usually obtained from Head-Related Impulse Responses (HRIR) measured under controlled conditions on a spherical surface of constant radius for only pre-defined set of elevation and azimuth angles, with respect to the ears of a target head at its centre. As a consequence, estimating the HRTF associated with any desired source location that has not been previously measured asks for some interpolation scheme. When realizing moving sound, especial care must be taken to avoid audible discontinuities along the required path. Additionally, in real time systems, the computational cost to perform HRTF interpolation must be kept to a minimum without degrading the final perceptual result. In the following subsections two different HRTF - interpolation schemes are discussed: The Vector-Base-Amplitude Panning (VBAP) and the Interpolation by Residues.

4.2.1 Vector-Base Amplitude Panning (VBAP)

In VBAP the amplitude panning method is reformulated with vectors and vector bases. The reformulation leads to simple equations for amplitude panning, and the use of vectors makes the panning methods computationally efficient. In Ville Pulkki [17], two and three dimensional VBAP methods are presented. For this work, just the 3D method is used and described.

To perform the HRTF interpolation with three dimensional amplitude panning, the typical two-channel stereophonic listening configuration [17] is extended with a third loudspeaker placed in an arbitrary position at the same distance from the listener as the other loudspeakers. However, the loudspeaker should not be placed on the two dimensional plane defined by the listener and the two other loudspeakers. The virtual source can now appear within a triangle formed by loudspeakers when viewed from the listener's position, as illustrated in figure 4.8. The term three dimensional amplitude panning denotes a method for positioning a virtual sound source, which are driven by coherent electrical signals with different amplitudes.

The direction of the virtual source is dependent on the relation of the amplitudes of the emanating signals. If the virtual sound source is moving and the loudness should be constant, the gain factors that control the channel levels have to be normalized. The sound power can be set to a constant value C, whereby the following approximation can be started:

$$g_1^2 + g_2^2 + g_3^2 = C \tag{4.18}$$

The virtual source can thus be placed on the surface of the three dimensional sphere, the radius of which is defined by the distance between the listener and loudspeakers. The region on the surface of the sphere onto which the virtual source can be positioned is called the active triangle, shown in figure 4.10.

As detectable in figure 4.10, the loudspeakers are positioned on the surface of a three dimensional unit sphere, equidistant from the listener. The three dimensional unit vector $l_1 = \left[l_1 \; l_2 \; l_3 \right]^T$, the origin of which is the centre of the sphere, points to the direction of loudspeaker 1. The unit vectors l_1, l_2 and l_3 then define the directions of loudspeakers 1, 2, 3, respectively. The direction of the virtual sound source is defined as a three dimensional unit vector $p = \left[p_1 \; p_2 \; p_3 \right]^T$. A sample configuration is presented in figure 4.10.

The virtual source vector p is expressed as a linear combination of three loudspeaker vectors l_1, l_2 and l_3, analogically to the two dimensional case [16], and expressed in matrix form:

$$p = g_1 l_1 + g_2 l_2 + g_3 l_3 \tag{4.19}$$

$$p^T = g L_{123} \tag{4.20}$$

Here g_1, g_2 and g_3 are gain factors, g = [g_1 g_2 g_3], and $L_{123} = \left[l_1\ l_2\ l_3 \right]^T$. Vector g can be solved,

$$g = p^T L_{123}^{-1} = \left[p_1\ p_2\ p_3 \right] \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix}^{-1} \tag{4.21}$$

if L_{123}^{-1} exists, which is true if the vector base defined by L_{123} spans a three dimensional space. The components of the vector g can be used by as gain factors after scaling it with C.

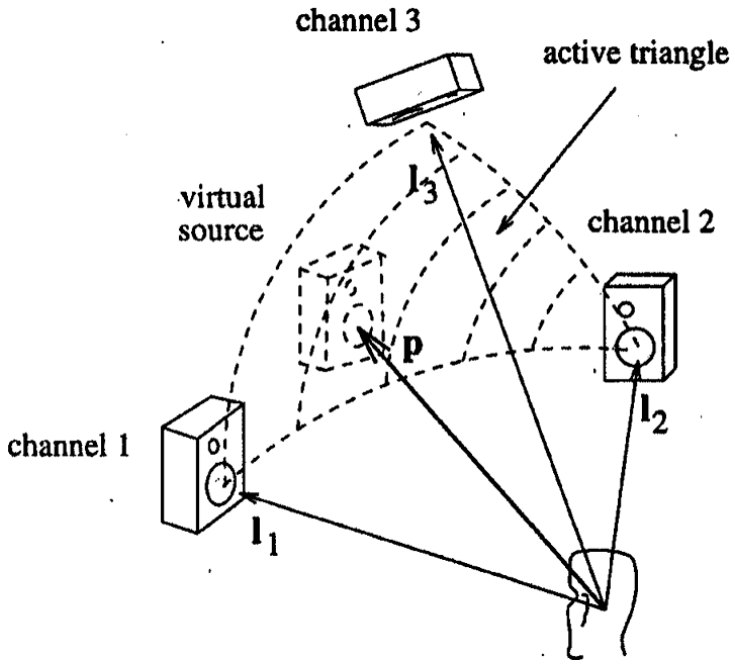


Figure 4.10 The region on the surface of the sphere – The active triangle

In the VBAP, as in all amplitude panning methods, the virtual source cannot be positioned outside the active arc or region. This holds even if the listener is in an arbitrary position. Thus the maximum error in virtual source localization is proportional to the dimensions of the active region. Therefore when good localization accuracies on a large listening area are desired, the dimensions of the active regions must be decreased. This is done by applying more loudspeakers on the desired region on the sound field, such as around and behind the screen in movie theatres.

In this work we extend the concepts of the VBAP explained so far to the case of HRTF interpolation. Notice that the HRTF are measured in fixed positions. This observation let us to adopt the idea that we can obtain a virtual transfer function in the desired position. Now, loudspeakers are replaced by HRTFs. The active triangle is defined by the three closest transfer functions relative to the desired position. Therefore equation 4.19 becomes

$$h(\theta,\varphi) = g_1 h_1(\theta_1,\varphi_1) + g_2 h_2(\theta_2,\varphi_2) + g_3 h_3(\theta_3,\varphi_3) \tag{4.22}$$

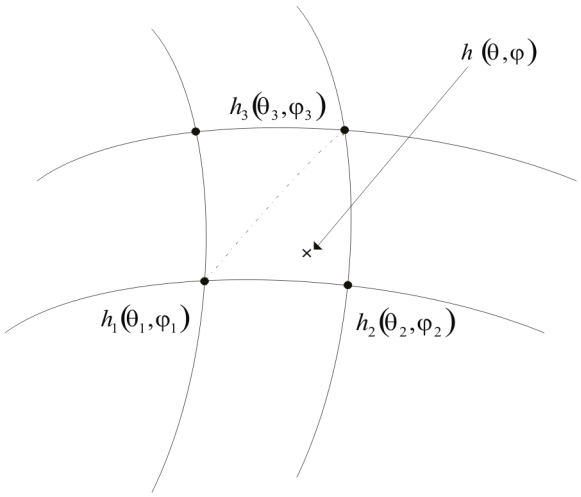


Figure 4.11 The active triangle of HRTF Interpolation

Where $h(\theta,\varphi)$ is the desired HRTF and $h_1(\theta_1,\varphi_1)$, $h_2(\theta_2,\varphi_2)$ and $h_3(\theta_3,\varphi_3)$ are existing HRTFs defining the active triangle. In figure 4.11 a graphical example is shown.

VBAP has three important properties:

1) If the virtual source is located in the same direction as any of the loudspeakers, the signal emanates only from that particular loudspeaker, which provides maximum sharpness of the virtual source.

2) If the virtual source is located on a line connecting two loudspeakers, the sound is applied only on that pair, following the tangent law. The gain factor of the third loudspeaker is zero.

3) If the virtual source is located at the centre of the active triangle, the gain factors of the loudspeakers are equal.

The interpolation method has been tested with a KEMAR HRTF raster of 10˚ and 30˚ azimuth and elevation.

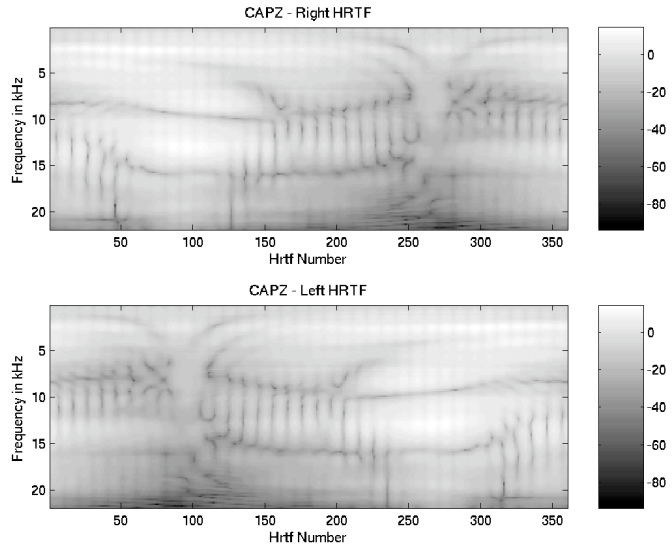


Figure 4.12 360 VBAP - interpolated HRTFs (0˚ - 359˚ azimuth, 0˚elevation) as a function of time and frequency, using a 10˚ raster of azimuth and elevation

Figure 4.12 shows 360 VBAP - interpolated HRTFs (0° - 359° azimuth, 0°elevation) as a function of time and frequency, using a 10° raster of azimuth and elevation. If we compare figure 4.12 with figure 4.8, we can see that the transitions between the different HRTFs of figure 4.12 are much smoother. Furthermore, the Vector-Base-Amplitude Panning is very efficient in computation.

Figure 4.13 shows 360 VBAP - interpolated HRTFs (0° - 359° azimuth, 0°elevation) as a function of time and frequency using a 30° raster of azimuth and elevation. If we now compare figure 4.13 with figure 4.8, we can see that still the transitions between the different HRTFs of figure 4.13 are much smoother and that the main structure of the azimuth cues still exists. Furthermore, the Vector-Base-Amplitude Panning with 30° is even more efficient in computation than the interpolation using a 10° raster because the same source movement can be managed with a smaller number of HRIRs.

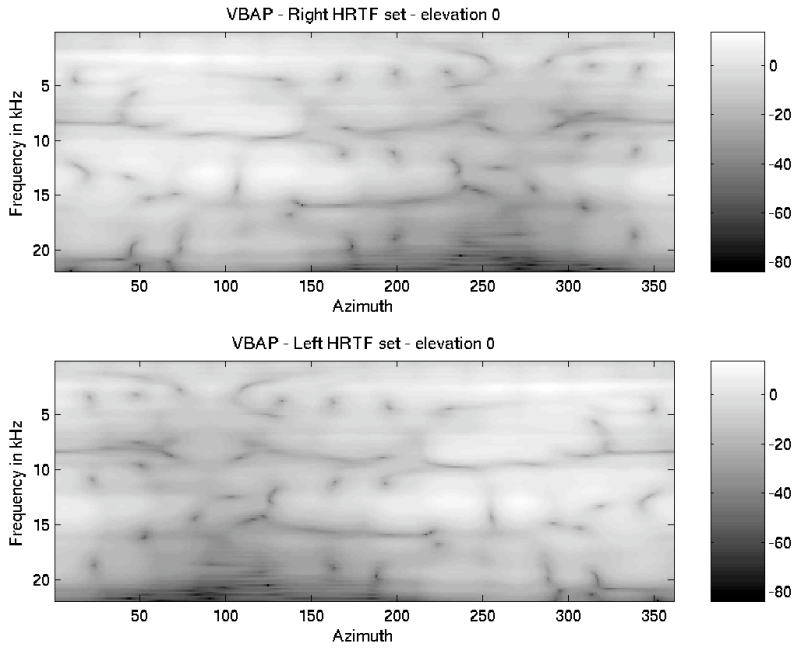


Figure 4.13 360 VBAP - interpolated HRTFs (0° - 359° azimuth, 0°elevation) as a function of time and frequency, using a 30° raster of azimuth and elevation

4.2.2 Common-Acoustical-Pole and Residue Interpolation (CAPR)

Another interpolation scheme is based on the Common Acoustical Pole and residue model. It becomes clear that this model uses the common acoustical poles p_{Ci}, which correspond to the resonance frequencies and damping factors of the ear channel. The CAPZ model, discussed in chapter 4.1.1, is represented as:

$$H_{CAPZ}(z,\theta) = \frac{C_C z^{-Q_1} \prod_{i=1}^{Q_2} \left[1 - q_i(\theta) z^{-1}\right]}{\prod_{i=1}^{P} (1 - p_{Ci} z^{-1})} \tag{4.23}$$

where C_C is a gain constant and θ represent the direction of the source. The poles p_{Ci} are direction-independent and the zeros $q_i(r_S, r_0)$ in equation 4.23 are direction-dependent.

The basic of this model is that a Head Related Transfer functions can be expressed by using eigenfrequencies (resonance frequencies) and their eigenfunctions, as discussed in Yoichi Haneda, Yutaka Kaneda, and Nubuhiko Kitawaki [18]. That leads to the following equation:

$$H(z,\theta) = \sum_{i=1}^{P/2} \left[\frac{A_i(\theta)}{1 - p_{Ci} z^{-1}} + \frac{A_i^*(\theta)}{1 - p_{Ci}^* z^{-1}} \right] \tag{4.24}$$

where P is the number of poles in the objective frequency band and function $A_i(\theta)$ is a residue function. The superscript * denotes the complex conjugate. In this model, the common acoustical poles p_{Ci} and their residues $A_i(\theta)$ are generally complex numbers.

Like explained in [16], the specific residue value $A_i(\theta)$ for the i-th common acoustical pole p_{Ci} at the source direction θ can be calculated using:

$$A_i(\theta) = \frac{C_C z^{-Q_1} \prod_{i=1}^{Q_2} \left[1 - q_n(\theta) z^{-1}\right]}{\prod_{\substack{n=1 \\ n \neq i}}^{P} (1 - p_{Cn} z^{-1})} \tag{4.25}$$

The Common Acoustical Pole and Residue interpolation method is outlined in figure 4.14. There, the HRIR $h(x_{IN})$ at the source direction x_{IN} is interpolated by using, as an example, the four impulse responses $h(x_1)$ to $h(x_4)$ measured at locations x_1 to x_4. The number of impulse responses is required to exceed the number of parameters in the residue functions. First, the common acoustical poles p_{Ci} are estimated from the measured impulse responses. Then the parameters of the residue functions are determined based on the calculated residue values $A_i(x_m)$ at the four positions. The residue function $\hat{A}(x_m)$ is thus expressed by this parametric model. Residue value $\hat{A}_i(x_{IN})$ of the direction x_{IN} is calculated by evaluating the estimated residue function $\hat{A}_i(x)$ for $x = x_{IN}$. These steps are repeated for all i (i = 1,2,...,P). Finally, using all of the estimated residue values $\hat{A}_i(x_{IN})$ (i = 1,2,...,P) and the common acoustical poles p_{Ci} (i = 1,2,...,P), the interpolated HRIR $\hat{h}(x_{IN})$ at x_{IN} is obtained.

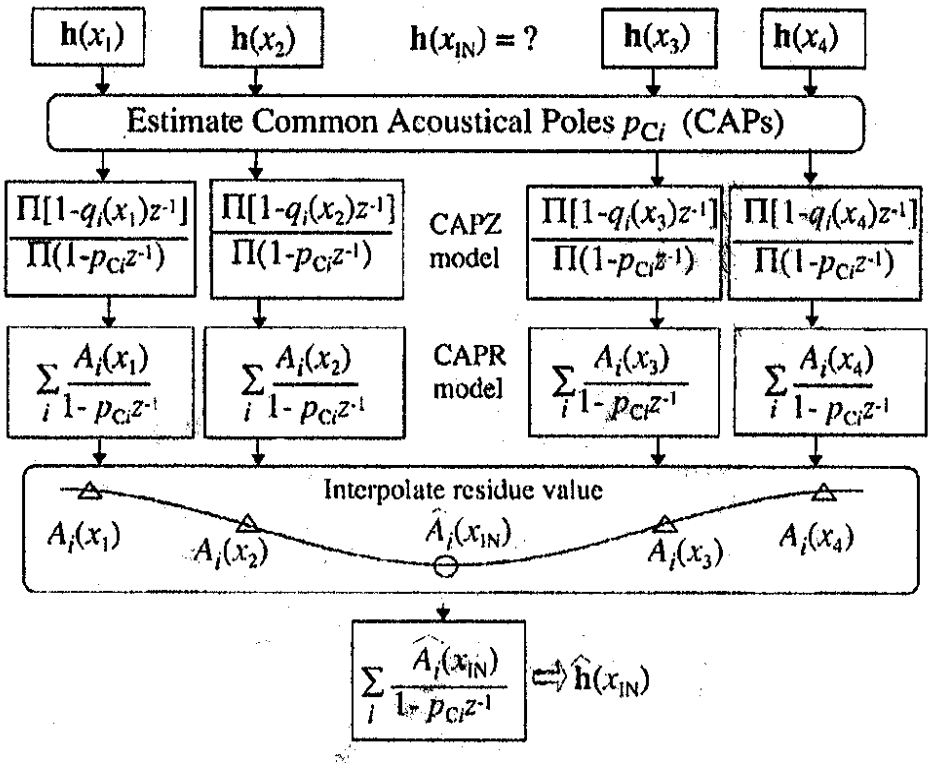


Figure 4.14 The Common Acoustical Pole and Residue interpolation method [18]

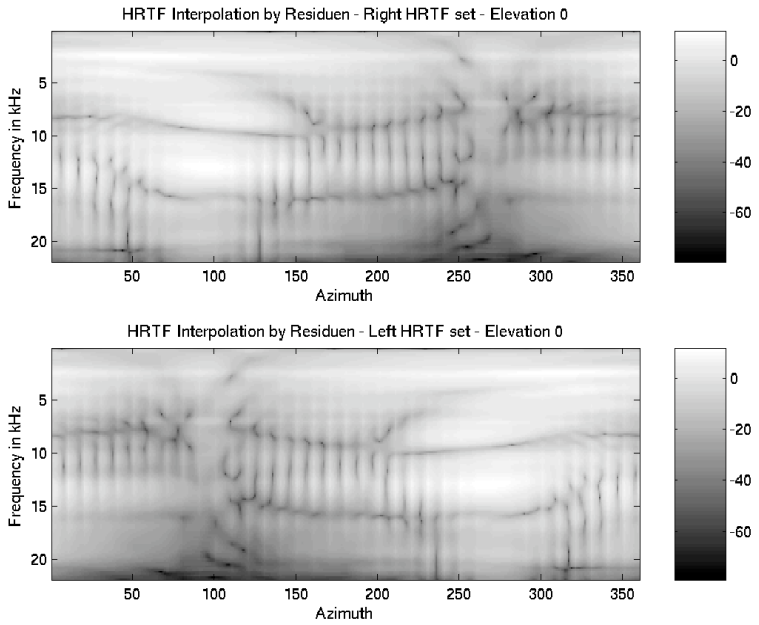


Figure 4.15 Residue interpolated HRTF set (azimuth 0˚- 359˚, elevation 0˚) as a function of time and frequency - 10˚ raster -

Like the VBAP, described in 4.2.1, the CAPR interpolation method has been tested for a HRTF database raster of 10˚ and 30˚ azimuth and elevation. The desired HRTF is interpolated, as with VBAP (chapter 4.2.1), by using three transfer functions.

In figure 4.15, the Residue interpolated HRTF set as a function of time and frequency is pictured. The raster of the CAPZ directions is chosen to be 10˚ in azimuth and elevation. Compared to figure 4.8, the KEMAR HRTF set as a function of time and frequency, it can be detected that the azimuth cues are well interpolated. The source movement can be realized realistically too. The disadvantage of this kind of interpolation is that the computation is more inefficient than the Interpolation with Vector-Base Amplitude Panning.

To perform the CAPZ more efficient in computation the raster size is set to 30˚ azimuth and elevation. Figure 4.16 shows the Residue interpolated HRTF set as a function of time and frequency using a 30˚ CAPZ raster. It can be noticed that compared to figure 4.13, the VBAP-interpolated HRTF set as a function of time and frequency with the same raster, the patterns of the two plots are quite the same, but the contrast, the amplitude of the interpolated HRTFs changes, especially at the angle between 250˚ and 360˚ at the contralapseral ear. That leads to distance distortion in exactly that region and could be explained by the linear interpolation of the residue. Like at VBAP the active triangle, three different residues are interpolated to get the desired residue. However, if the virtual source is located on a line connecting two HRTFs,

the sound is not applied only on that pair. That is why a gain factor of the third HRTF is not zero and dependent on the relative distance to the virtual sound source a percentage of this third HRTF is added to the interpolated HRTF.

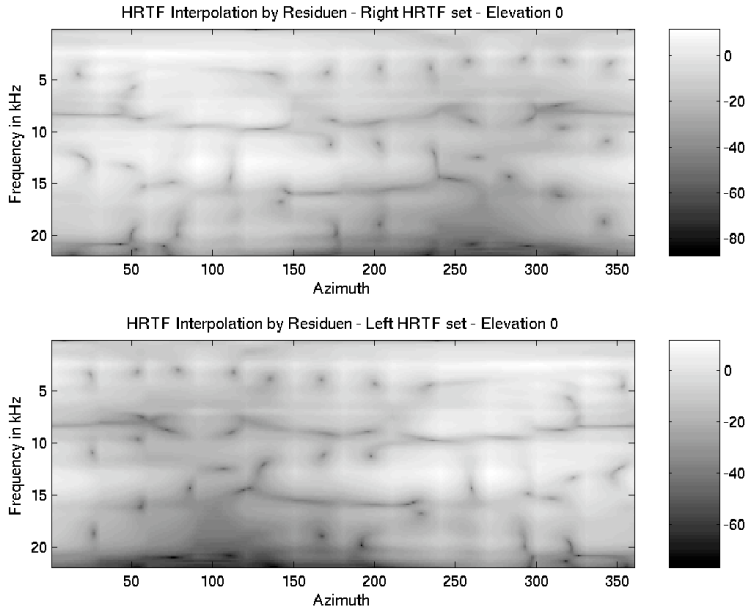


Figure 4.16 Residue interpolated HRTF set (azimuth 0˚- 359˚, elevation 0˚) as a function of time and frequency - 30˚ raster -

Chapter 5
The System

5.1 The Signal Routing of the 3D System

Figure 5.1 shows the signal routing of the System. In the figure, the input signal shown at the top represents the individual object sound that is to be spatially processed to create the scene. The input signal is monophonic.

Primary, the monoaural input signal is split into two monoaural signals: One for the direct path of the system and the other one for the reverberant path.

The signal of the direct path is processed through the 1/d attenuation, the air absorption and then the 3D spatial effect (HRTF filtering), labelled “3D cues“ in the figure. The air absorption effect is controlled by the distance between the sound object and the listener, as explained in chapter 3.1.1. The 3D spatial effect is controlled by the position of the sound object relative to the listener. The 3D spatial effect creates a stereophonic (two channels) output.

In the reverberant path of the system, the signal first goes throught the early reflection filter. There, first upto fourth order reflections are calculated. The signal is retarded with the delay of the reflection, is diminished with 1/d, goes throught the air absorption filter, is attenuated by the wall absorption filter, that is defined by the material of the wall, at which the reflection

occurs. The order of the reflections defines the number of wall absorption filters. The still monoaural signal now is processed with the 3D cues. In other words, filtered with the HRIR of its direction. The last step of the Early Reflection filter is the summation of every left ear signal of the first reflections and the summation of every right ear signal.

The next step of the reverberant path is the Late Reverberation unity. This part of the system is implemented with a feedback delay network, explained in chapter 3.2.2. To place the late reverberation at the right spot of the impulse response, the feedback delay network is predelayed. It is important to know that the feedback delay network of the right-ear signal is mutated to provide inside head localisation.

Finally, the left signal output of the direct path is mixed with the left output signal of the reverberant path and the right signal output of the direct path is mixed with the right output signal of the reverberant path. At the output of the system, a binaural signal is obtained.

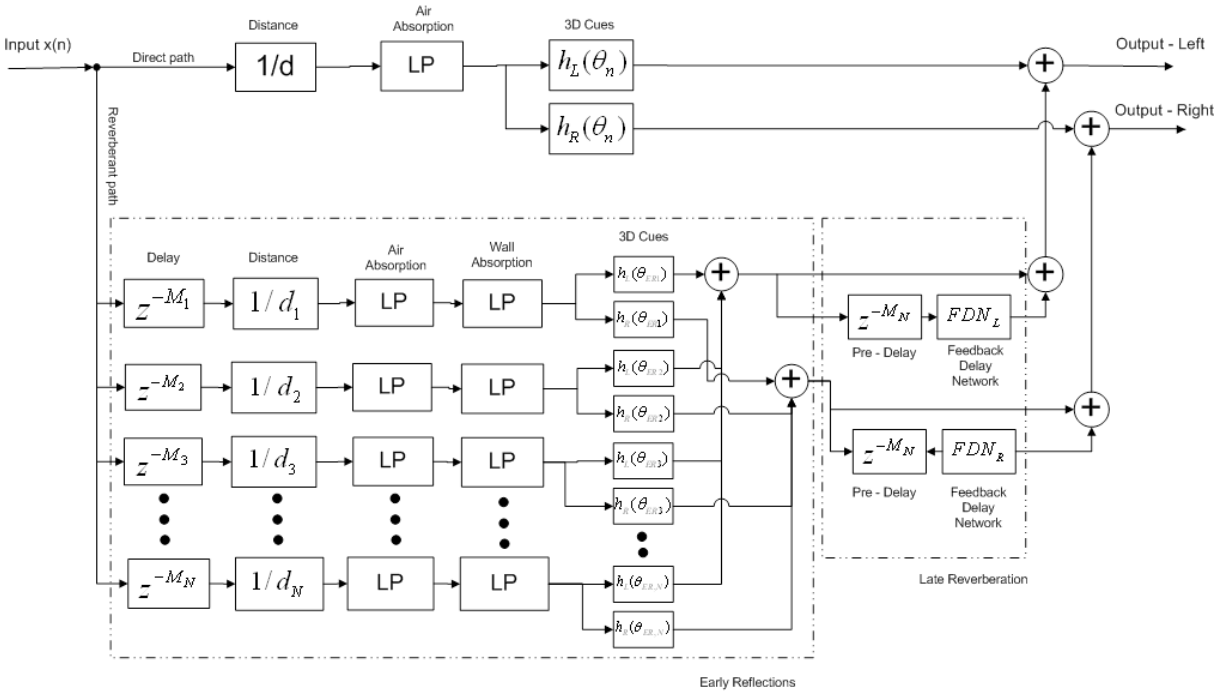


Figure 5.1 The Signal Routing of the 3D System

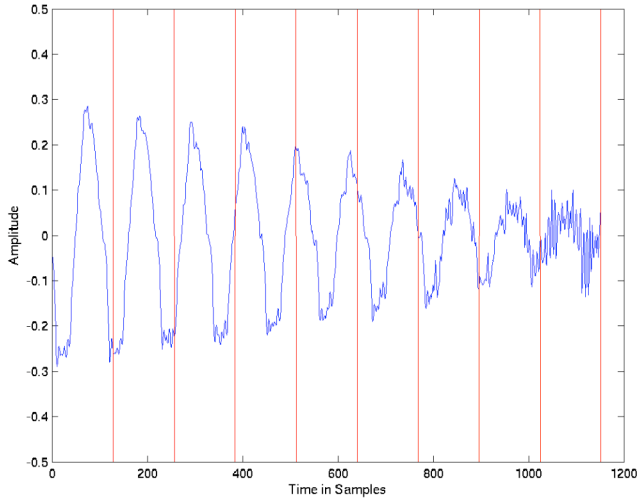


Figure 5.2 Input signal is divided into blocks of the length of 128

To render binaural, moving sources, the input signal is divided into blocks of the length of 128, as shown in figure 5.2. Each block is processed by the system, shown in figure 5.1. A trajectory that defines the path direction of the sound is implemented and each of its positions is transferred to the 3D spatialisation unit. If a direction does not exists in the set of HRTF curves, they are interpolated as mentioned in chapters 4.2.1 and 4.2.2. The distance (Δd) of two contiguous points on the trajectory is defined by the velocity (v) of the source, the block-size (BS) and the Sampling Frequency (Fs):

$$\Delta d = \frac{BS}{Fs} \cdot v \tag{5.1}$$

Finally the computed blocks are composed by the overlap and add method [21].
If the directional interpolation unit that calculates a new HRTF dependent on the direction is not utilized, just HRTFs placed at the disposal of the database can be used. In this work the KEMAR compact HRTF set with a maximum azimuth resolution of 5° at 0° elevation is chosen. Thus, clicks in the output signals are perceived, as the switch from one direction (e.g. 0° azimuth) to the next one (5° azimuth) is too 'hard'. Therefore, to provide these clicks, two kinds of block filter techniques have been studied:
The first one is pictured in figure 5.3. The blocksize is set to be 1000 sampels and weighted with a hanning window. The hopsize is the half of the blocksize. Then, each second block is

filtered with a different HRTF, to create a fade out of the previous filtered sigal and a fade out of the current filtered signal in the transition area, shown in figure 5.3. The cyan curve represents the summation of the hanning windows and is pictured with an offset of 1. It can be recognized that this block-filter technique avoids amplitude modulation. The disadvantage of this method is that if the blocksize is chosen to be too small (e.g. 128 sampels), the fade in and fade out time is too short and clicks are still perceived. A small blocksize is desired to increase the maximum possible source-velocity.

To avoid these clicks with a small blocksize of 128 samples, rectangular windows instead of hannings windows are used, shown in figure 5.2. The first 5 blocks in the transition area are filtered with a linear combination of the two existing transfer functions, as it is shown in figure 5.4.

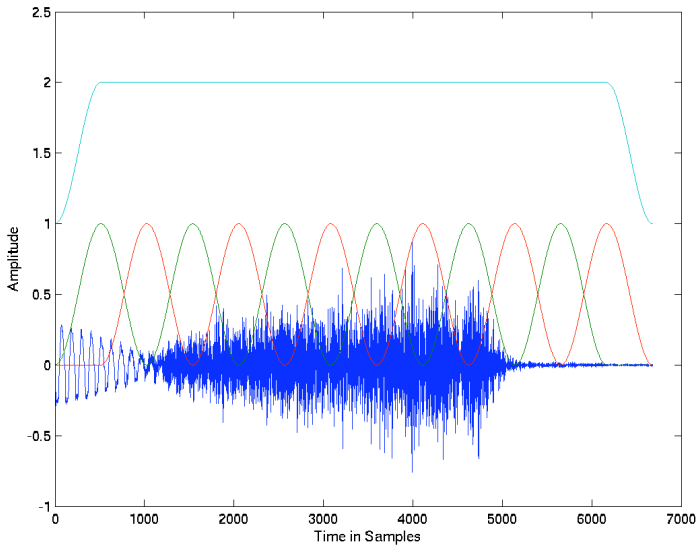


Figure 5.3 Block filter technique using hanning windows

Blocknumber	Previous HRTF	Current HRTF
1	80%	20%
2	60%	40%
3	40%	60%
4	20%	80%
5	0%	100%

Figure 5.4 Linear interpolation of HRTFs in the transition area

As mentioned above, this block-filter technique avoids clicks in the signal even with a small blocksize.

$$5 \text{ (interpolation blocks)} \cdot 128 \text{ (samples)} = 640 \text{ samples} = 14.5 \text{ ms}$$

If the blocksize is set to be 128 samples and the sample frequency is 44100 Hz, a velocity of 5˚ per 14.5 ms (around 360˚ per second) can be realized.
Therefore, without using the directional interpolation unit, this method is utilized.

5.2 The Manual:

To get started:

Copy the desired folder, Windows Version or Mac Version from the attached CD to your computer. Check the directory of the folder named 'KEMAR' and change the fifth line of the following m files by editing the new directory:
Capzred.m - line 5
Matr.m - line 5

To start the program, execute the function **Main.m** in Matlab. The main interface, shown in figure 5.5 should appear:

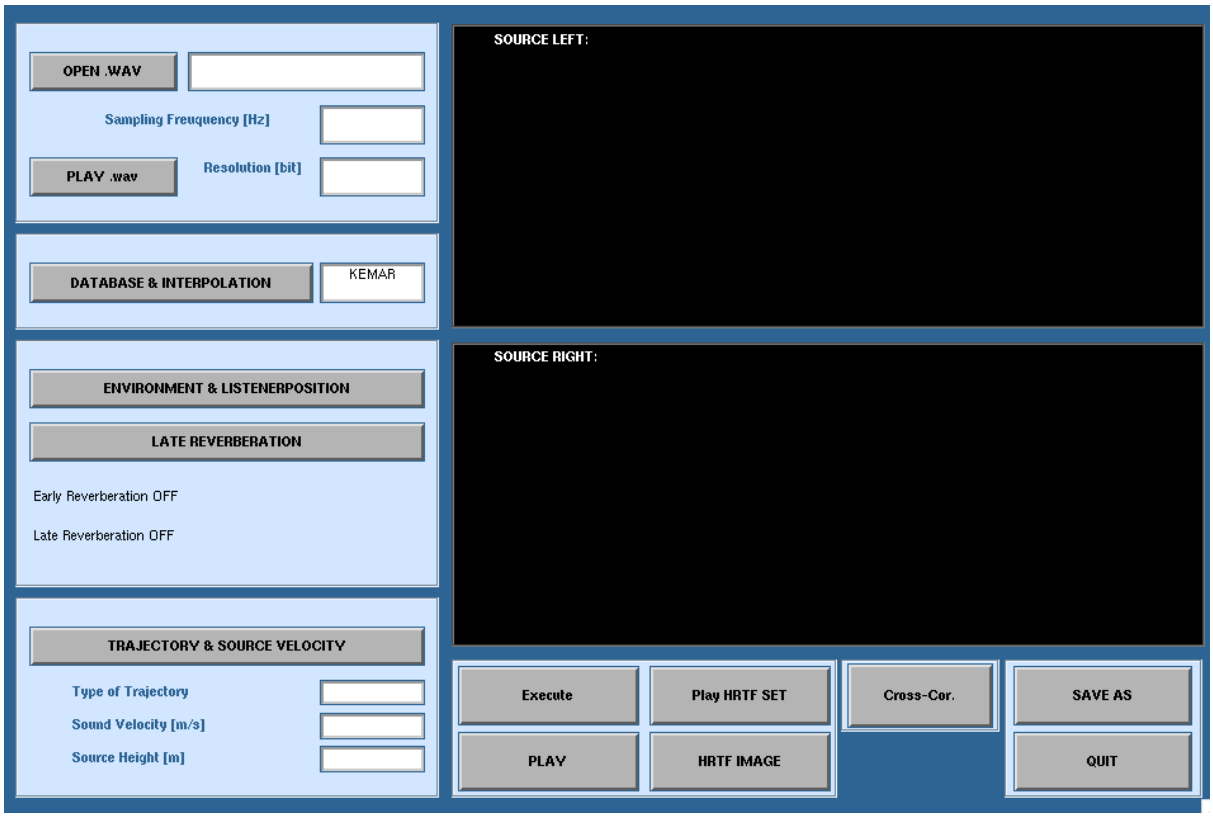


Figure 5.5 The Main Interface

There are six main parts:

1) **open .wav**
2) **database**
3) **environment and late reverberation**
4) **trajectory**
5) **control panel**
6) **signal plot**

1) The open .wav

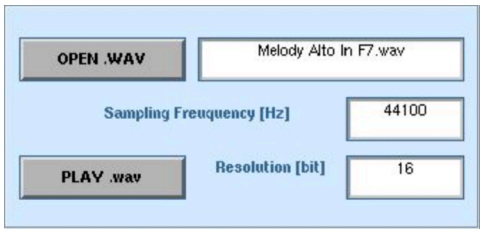


Figure 5.6 The open .wav panel

If you click **OPEN .WAV** the following window, shown in figure 5.7 appears:

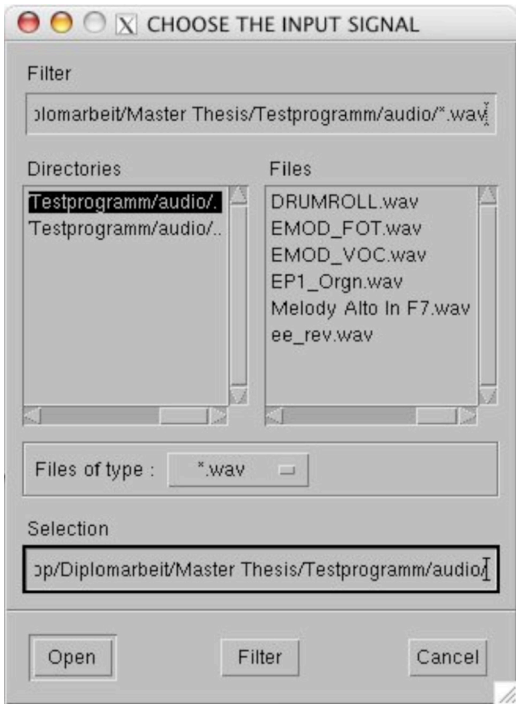


Figure 5.7 Choose the input signal

Choose a wav file and click **Open**. If you are using Matlab for Macintosh, please do not click **Cancel** or close this window. Otherwise an error occurs and you have to enter **cd ..** at the Matlab command window

By activating the Open button, the Sample Frequency and the Resolution of the wav file is displayed in the Main interface.

2) The database & interpolation

Figure 5.8 The database and interpolation panel

If you click **DATABASE** the following window, shown in figure 5.9 appears:

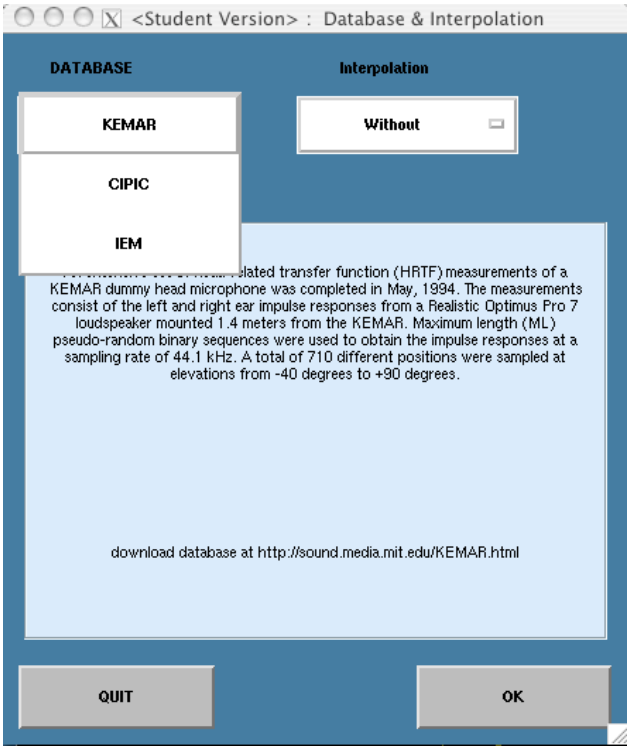


Figure 5.9 Choose database popupmenu

Here, the database can be chosen. This program only uses the *KEMAR database*. If you want to implement the CIPIC or IEM database, the matlab file database1.m has to be mutated by specifying the existing elevation and azimuth angles of the database.
You don't have to choose the database. By loading the **Main.m** file, the *Kemar database* is already chosen.

In the Interpolation popupmenu the program user has the choice between three types of interpolation: The first one, in the interface called '*Without*', interpolates the HRTF direction by replacing the nearest existing HRTF of the database.
The second and the third, VBAP and CAPR are explained in chapter 4.

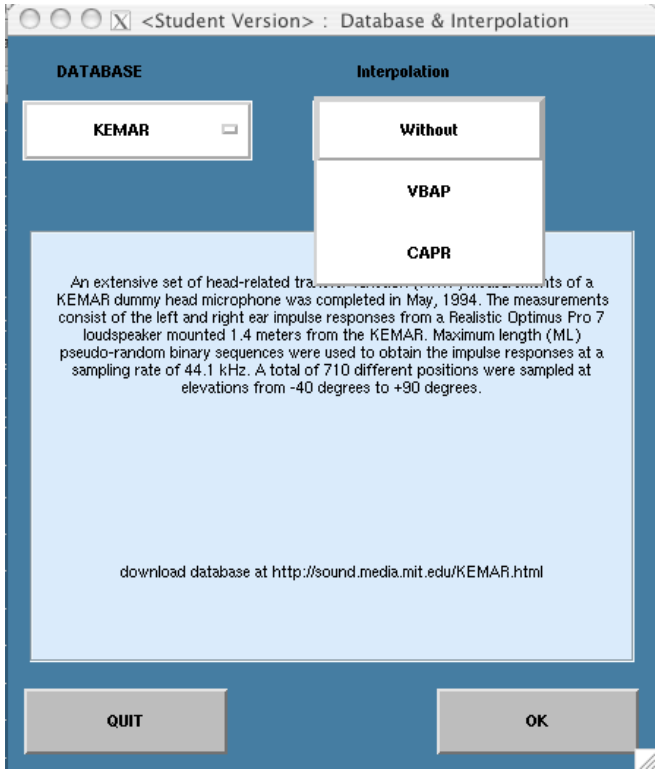


Figure 5.10 The choose interpolation popupmenu

3) Environment and late reverberation

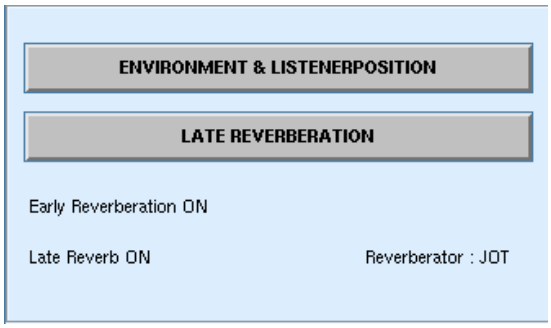


Figure 5.11 The environment and late reverberation panel

If you click **Environment & Listener position** the following window, shown in figure 5.12 appears:

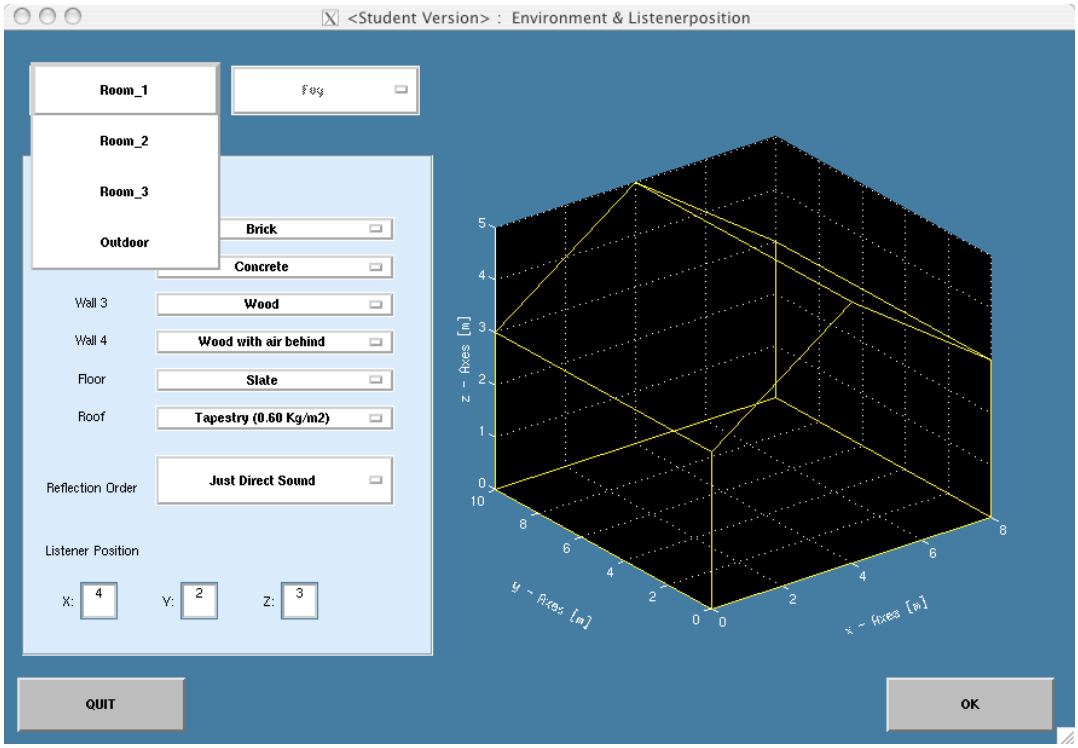


Figure 5.12 The choose the environment popupmenu

In the first popupmenu the room itself can be chosen. There are three simple rooms pre-programmed. Choose the floor, wall and roof material of the environment by clicking the popupmenues in the light-blue frame, like it is shown in figure 5.13. In the same frame, the reflection order and the listener position can be chosen.

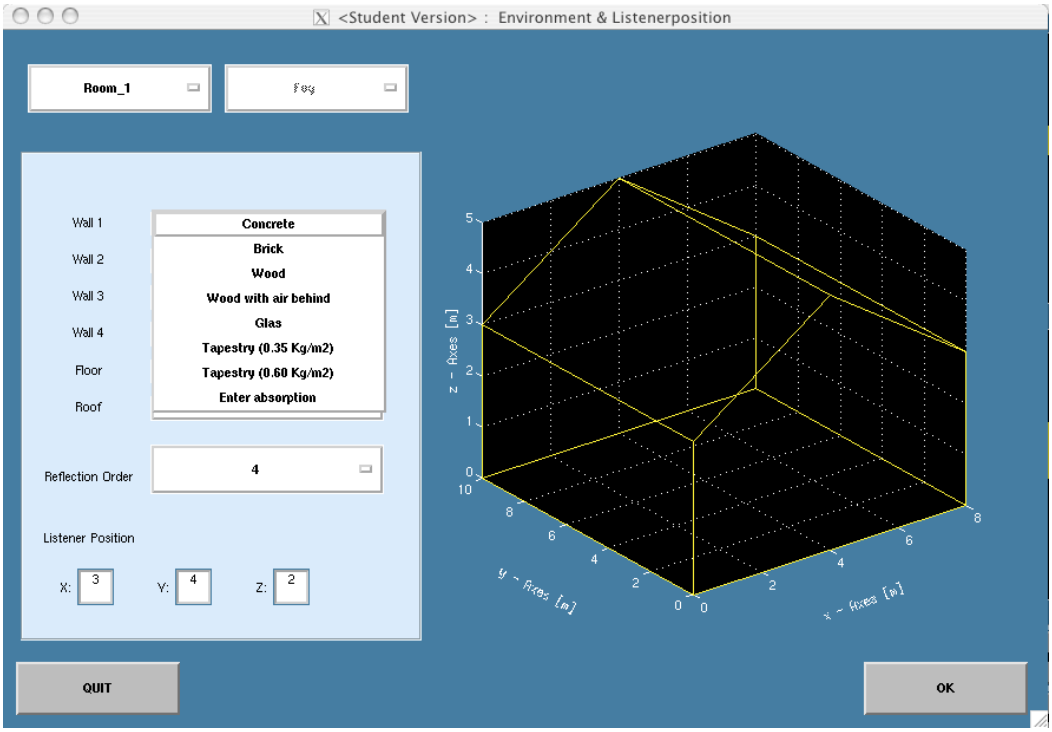


Figure 5.13 The choose the absorption material popupmenu

The fourth environment is the outdoor propagation with just the floor reflections. If this environment is selected, there is the choice to select the type of the outdoor environment as shown in figure 5.14.

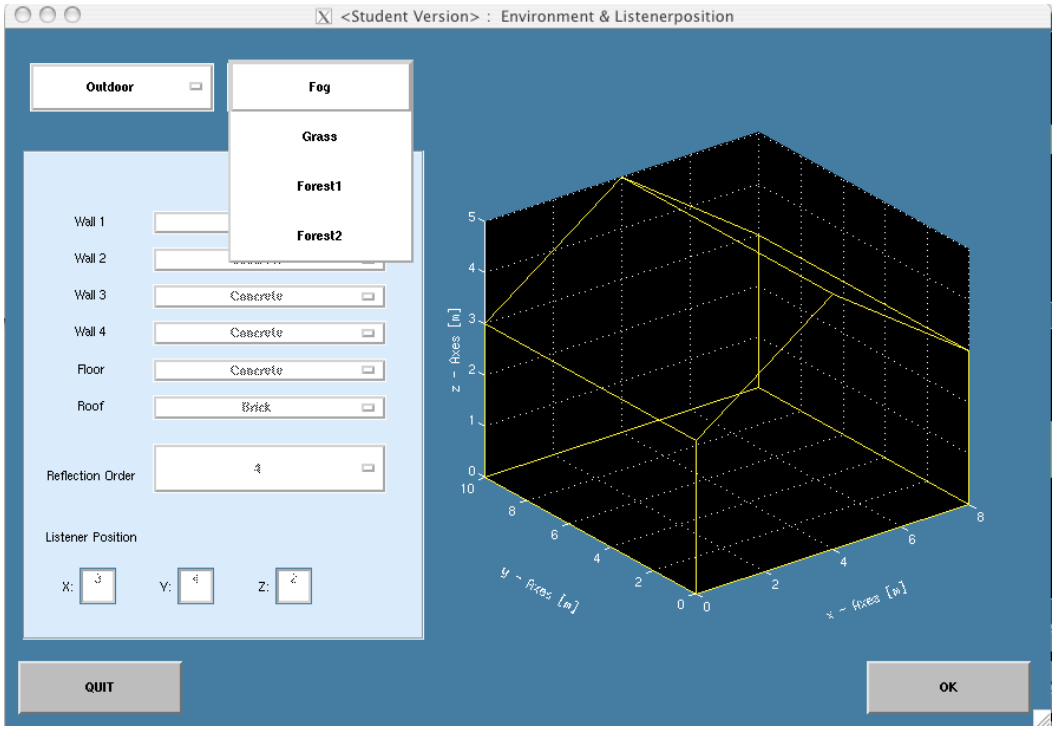


Figure 5.14 The choose the absorption material popupmenu

If you want to add an environment, you have to create a .mat file with the following pattern:
The first number in the first line of the .mat file defines the number of planes of the environment. The second and the first column of the first line are filled with zeros. In the second line of the matrix, the first plane is defined. The first element of this line stands for the number of vertices of the layer and the second one defines if the layer is the floor, the roof or a wall. 0 stands for floor, 1 for wall and 2 for roof. The next lines of the file are the vertices of the plane. The first element stands for the X-, the second one for the Y- and the third for the Z-axis of one vertex of the plane. It is important to specify the order of the vertices clockwise or counter clockwise. This is repeated for every plane.
The following Matrix shown in figure 5.15 represents such a pattern for a quadratic room - a cube:

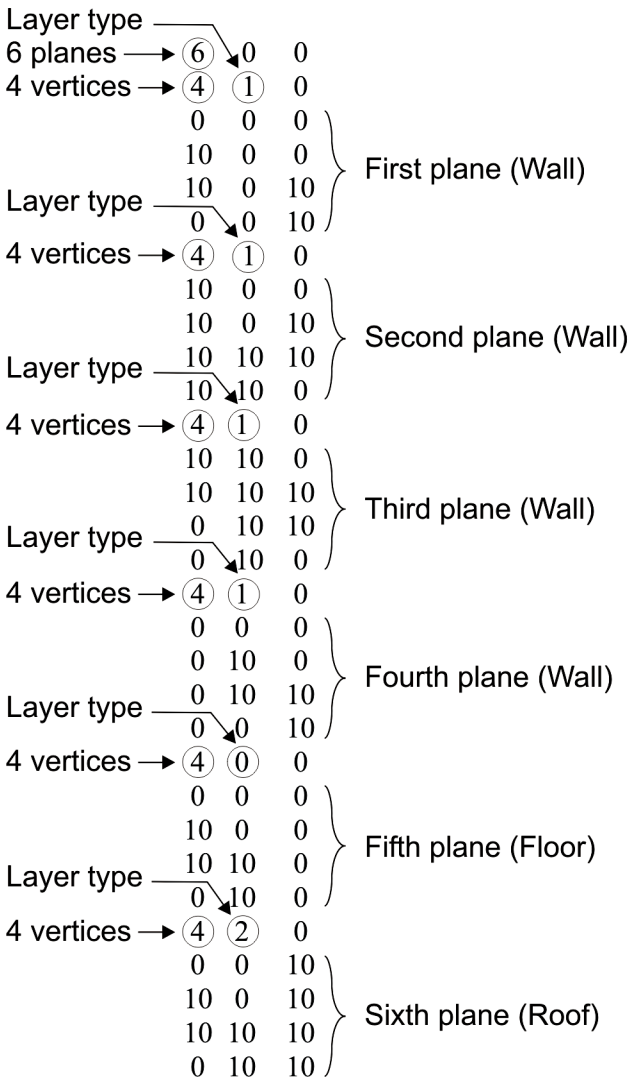


Figure 5.15 The Matrix pattern for a cube

If you click **Late - Reverberation** a Late reverberator and its parameters can be selected, as shown in figure 5.16.

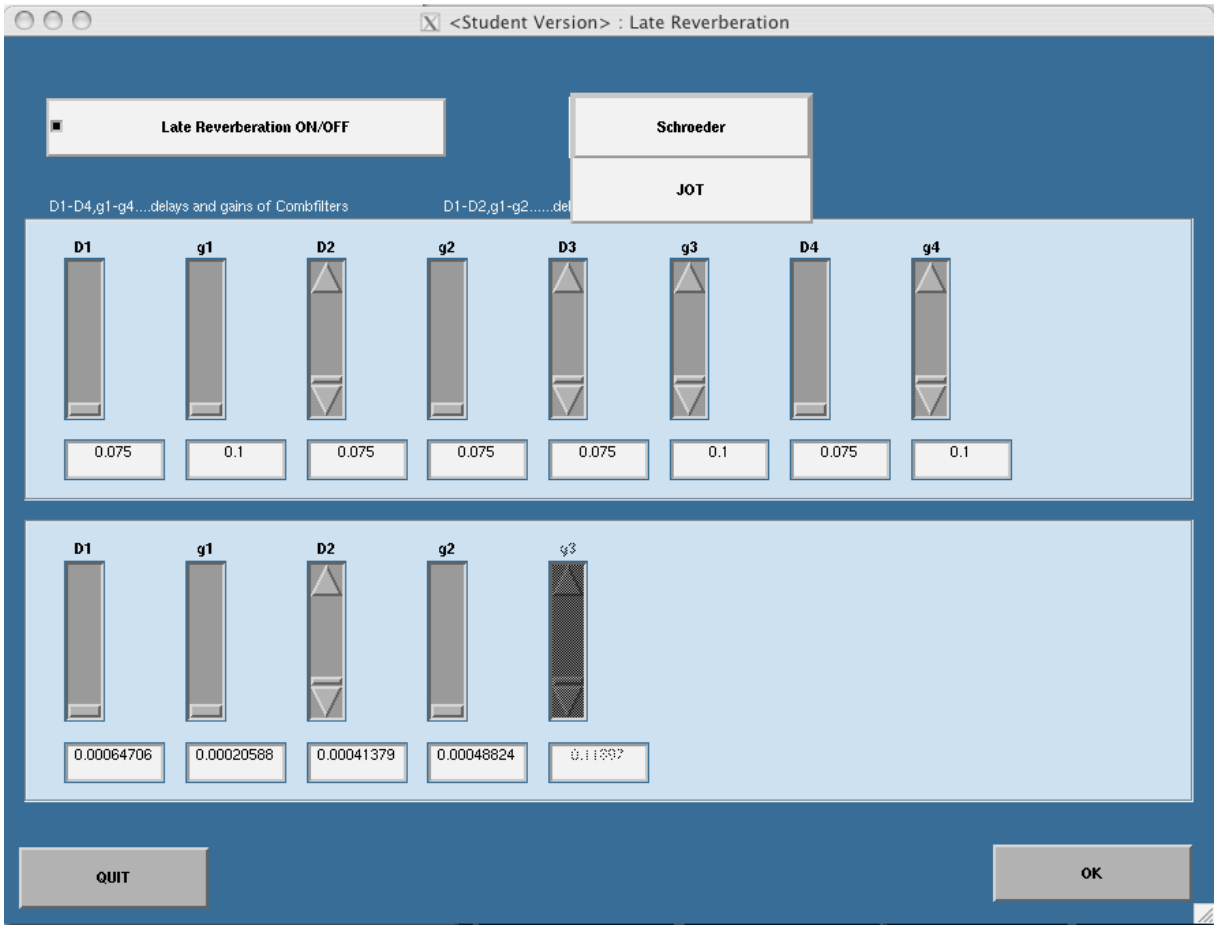


Figure 5.16 The Late reverberation panel

4) The trajectory

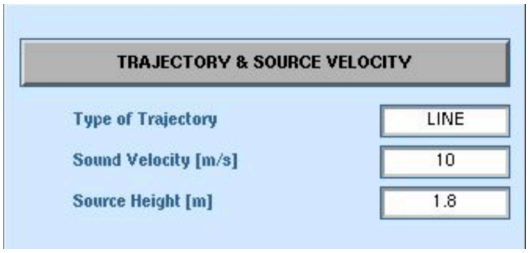


Figure 5.17 The trajectory & source velocity

If you click **Trajectory & Source Velocity** the following window, shown in figure 5.18 appears:

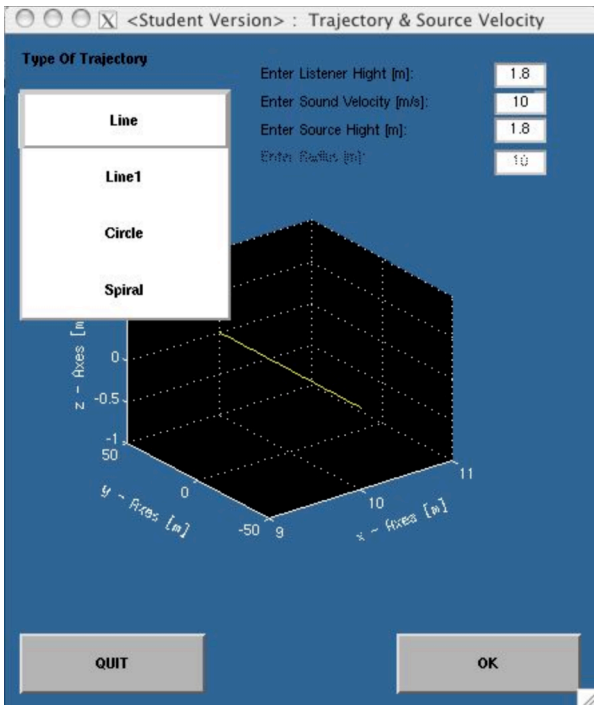


Figure 5.18 The type of trajectory popupmenu

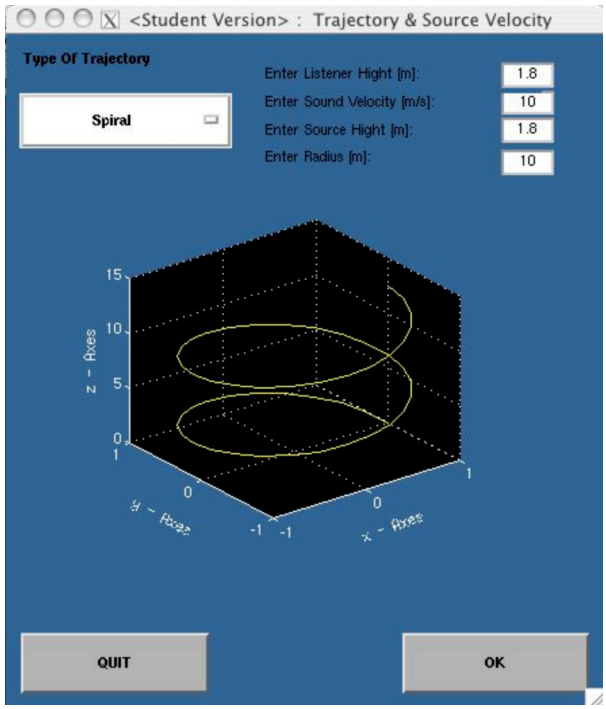


Figure 5.19 The parameters of the trajectory popupmenu

Choose a trajectory and confirm with **OK**. If *Circle* or *Spiral* is chosen, there is the opportunity to enter a radius, as shown in figure 5.16. Enter and confirm with **OK**.

5) The control panel

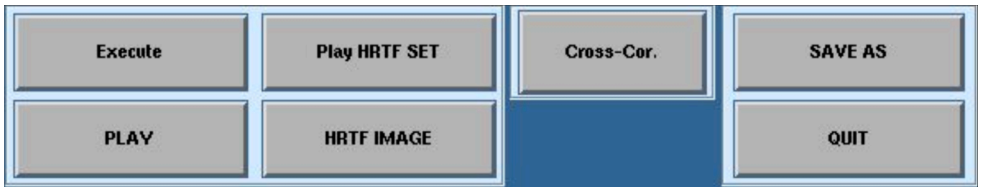


Figure 5.20 The control panel

If you click **Execute** the following waitbar, shown in figure 5.21 should appear:

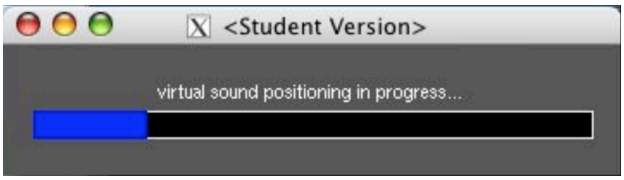


Figure 5.21 The waitbar

If you click **Play HRTF Set**, the used HRTF set is plotted. For instance, the figure 5.22 shows the HRTF set of a *Line Trajectory*. Close this window by clicking the **Close** button.

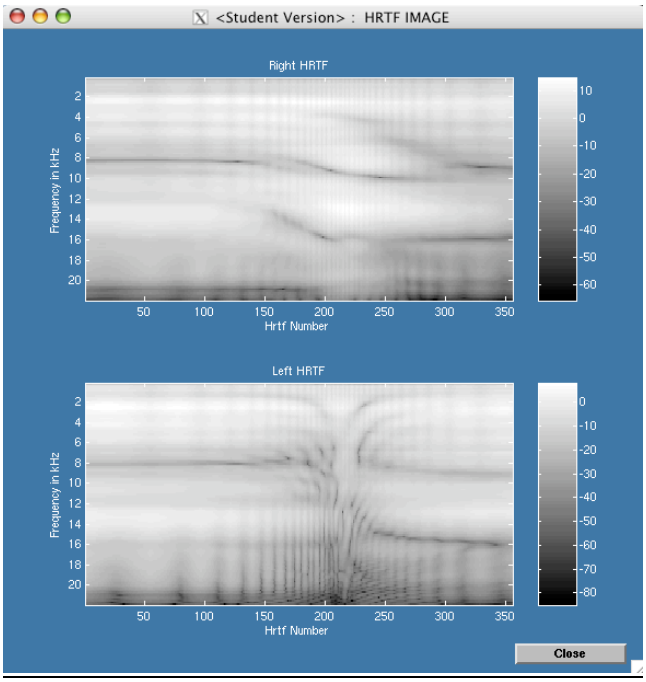


Figure 5.22 The HRTF Image window

If you click **Play HRTF Set**, the HRTF set (Hz over dB) is shown.

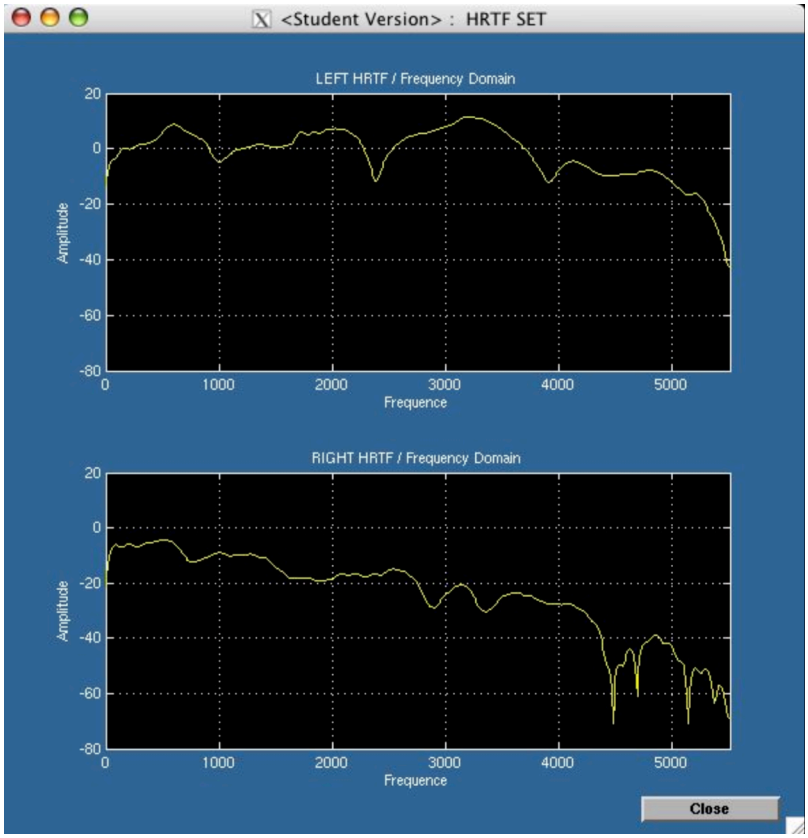


Figure 5.23 The HRTF SET window

Close this window by clicking the **Close** button.

If you click **Cross–Cor.** The following window, shown in figure 5.24 appears:

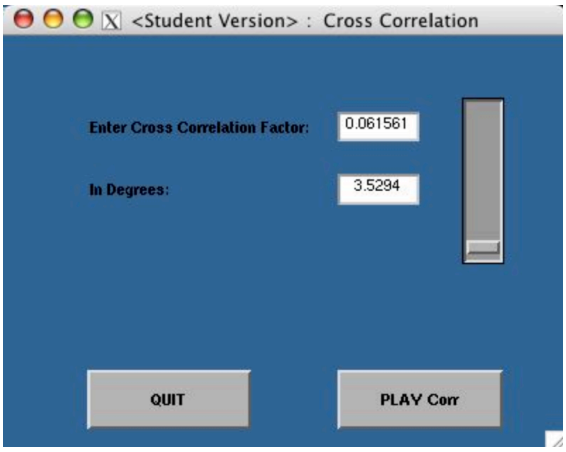


Figure 5.24 The Cross Correlation window

You can choose an *angle* and listen to the cross correlated sound by clicking **Play Corr**. Close this window by clicking the **Close** button.

If you click the **Save As** button the following window, shown in figure 5.25 appears:

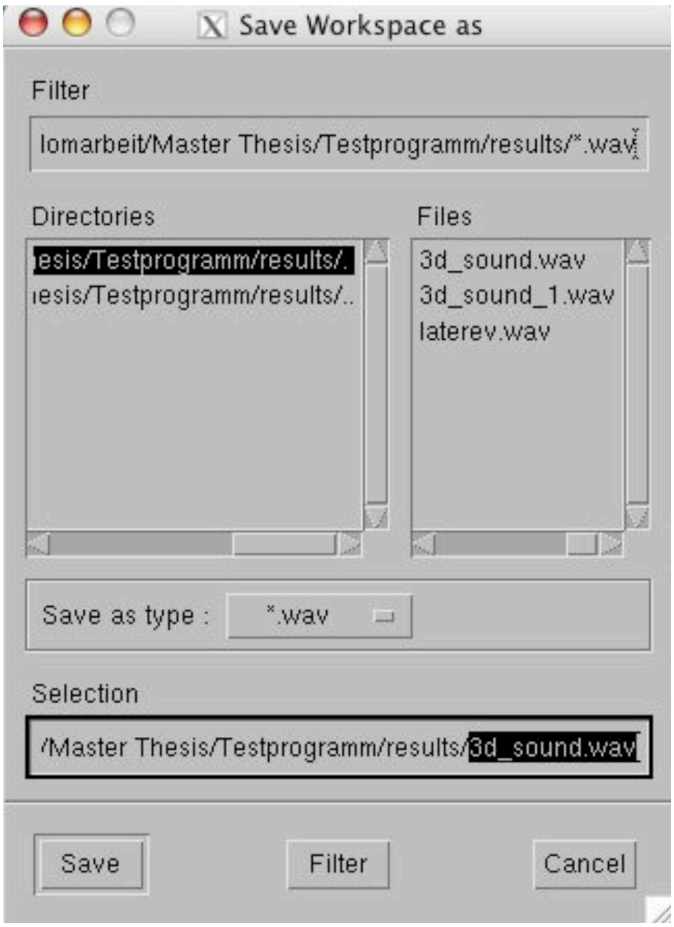


Figure 5.25 The "Save Workspace as" window

Enter a *file name* and confirm with **Save**.

If you click **PLAY** you can listen to the executed 3D sound.

The button **Close** finishes the program.

6) The signal plot

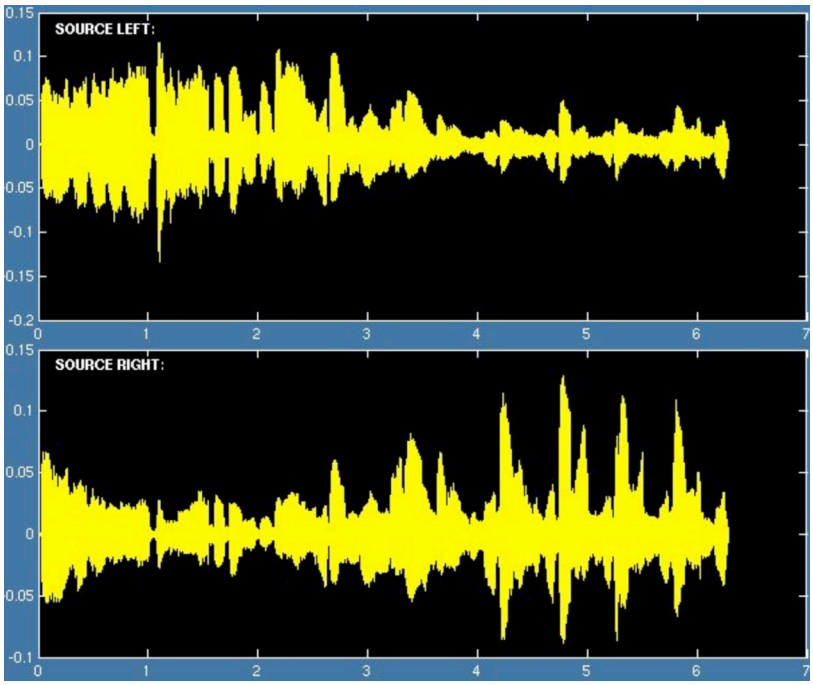


Figure 5.26 The Signal Plot

The X - axis is displayed in sec and the Y – axis in values between -1 and 1.

Error Messages:

If you don't enter enough parameters before executing the program the error message, shown in figure 5.27 will appear:

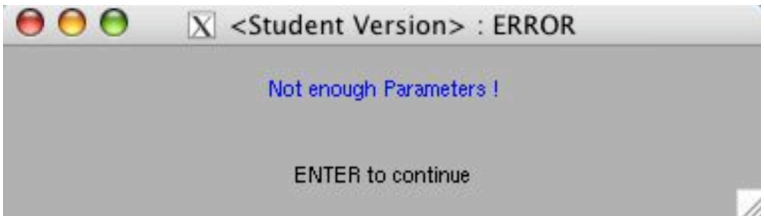


Figure 5.27 Error Message "Not enough Parameters"

If you don’t enter enough late reverberation parameters before confirming the reverberation panel, the error message, shown in figure 5.28 will appear:

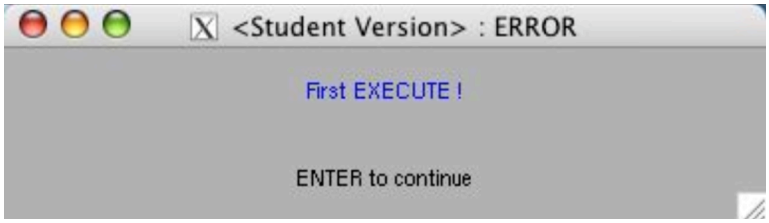


Figure 5.28 Error Message “First Execute“

References

[1] Begault D. R.: " 3-D Sound For Virtual Reality And Multimedia ", AP Professional USA 1994

[2] Blauert J.: " An Introduction of Binaural Technology ", S. Hitzler Verlag Stuttgart 1994

[3] Kahrs M. and Brandenburg K.: "Applications of Digital Signal Processing To Audio and Acoustics", Chapter 3, Kluwer Academic Publisher 1998

[4] Gardner W. G: " 3-D Audio Using Loudspeakers ", Massachusetts Institute of Technology 1998

[5] Zölzer U.: " DAFX – Digital Audio Effects ", John Wiley & Sons 2002

[6] Algazi V. R. and Duda R. O.: "Approximating the head-related transfer function using simple geometric models of the head and torso", University of California 2002

[7] Zotkin D. N, Hwang J., Duraiswami R., Davis L. S.: "HRTF Personalization Using Anthropometric Measurements", Institute For Advanced Computer Studies

[8] Algazi V. R., Duda R. O., Thompson D. M.: "The Use of Head and Torso Models For Improved Spatial Sound Synthesis", AES Convention Paper 2002

[9] Farina A.: "Simultaneous Measurement of Impulse Response and Distortion with a Swept Sine Technique", University of Parma

[10] E. Zwicker, H.Fastl: "Psychoacoustics, second updated edition, Springer 1999

[11] Gardner B. and Martin K.: "HRTF Measurement of a KEMAR Dummy-Head Microphone", MIT Media Lab Perceptual Computing – Technical Report #280, 1994

[12] Algazi V. R., Duda R. O. and Thompson D. M.: " The CIPIC HRTF Database", U.C. Davis CIPIC Interface Laboratory

[13] Kahrs M., Brandenburg K.: " Applications of Digital Signal Processing to Audio and Acoustics, chapter 3 (Gardner W. G.), Kluwer Academic Publisher 1998

[14] Heinrich Kuttruff: " Room Acoustics ", Third edition 1991, E & FN Spon

[15] J.S. Lamancusa: "Outdoor Sound Propagation", Penn State 2000

[16] Yoichi Haneda, Shoji Makino, Yutaka Kaneda, and Nubuhiko Kitawaki: " Common-Acoustical-Pole and Zero Modelling of Head-Related Transfer Function, IEEE Transactions on speech and audio processing, vol. 7. No. 2, march 1999

[17] Ville Pulkki: "Virtual Sound Source Positioning Using Vector Base Amplitude Panning", Laboratory of Acoustics and Audio Signal Processing. Helsinki University of Technology, FIN-02015 HUT, Finland

[18] Yoichi Haneda, Yutaka Kaneda, and Nubuhiko Kitawaki: " Common-Acoustical-Pole and Residue Model and Its Application to Spatial Interpolation and Extrapolation of a Room Transfer Function", IEEE Transactions on speech and audio processing, vol. 7. No. 6, November 1999

[19] K. Kammeyer, K. Kroschel: "Digitale Signalverarbeitung", 5.Auflage, Teubner 2002

[20] Ming Li: "Implementation of a model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction", Department of Electrical Engineering and Information Technology, University of Kaiserslautern

[21] A. Oppenheim, R.Schafer: "Discrete-Time Signal Processing", Prentice Hall 1999