Report T96:03

ISRN: SICS-T--96/03-SE
ISSN: 1100-3154

# VIRTUAL AUDIO

## Three-Dimensional Audio in Virtual Environments

*by Daniel Adler*

DCE-Group; 1996-06-17

dce@sics.se

Swedish Institute of Computer Science
Box 1263, S-164 28 KISTA, SWEDEN

Also Published: TRITA-NA-E9631

# ABSTRACT

Three-dimensional interactive audio has a variety of potential uses in human-machine interfaces. After lagging seriously behind the visual components, the importance of sound is now becoming increasingly accepted.

This master's thesis mainly discusses background and techniques to implement three-dimensional audio in computer interfaces. A case study of a system for three-dimensional audio, implemented by the author, is described in great detail. The audio system was moreover integrated with a virtual reality system and conclusions on user tests and use of the audio system is presented along with proposals for future work at the end of the thesis.

The thesis begins with a definition of three-dimensional audio and a survey on the human auditory system to give the reader the needed knowledge of what three-dimensional audio is and how human auditory perception works.

***Virtuellt ljud – Tredimensionellt ljud i virtuella världar***

Tredimensionellt ljud har en mängd potentiella användningsområden i människa-maskin gränssnitt. Efter att ha varit försummat gentemot de visuella komponenterna, så har ljudets betydelse nu börjat uppmärksammas allt mer.

Denna exjobbsrapport behandlar i huvudsak bakgrund och metoder för att implementera tredimensionellt ljud i datorgränssnitt. Ett system för tredimensionellt ljud, implementerat av författaren, beskrivs utförligt. Systemet integrerades dessutom med ett ”virtual reality”-system. Resultat och slutsatser av användartester och begagnandet av ljudsystemet ges tillsammans med förslag på fortsatt utveckling av systemet i rapportens avslutning.

Rapporten börjar med en definition av tredimensionellt ljud samt en orientering i det mänskliga hörselsystemet, för att ge läsaren de nödvändiga förkunskaperna.

# ACKNOWLEDGEMENTS

This master's thesis is the result of a project performed at the group for Distributed Collaborative Environments, the DCE-group, at the Swedish Institute of Computer Science (SICS).

The idea for this master's thesis was born when the author saw the virtual reality system DIVE developed at the DCE group and realised that it lacked three-dimensional audio. This in conjunction with the author's personal interest in sound resulted in this thesis.

Many people have been victims of my questions about all sorts of subjects, related and unrelated to the topics of this master's thesis. I would especially like to thank Olof Hagsand for the architectural aspects on the integration of the audio system with DIVE, as well as for proof-reading this thesis. A special thanks also goes to Emmanuel Frécon and Mårten Stenius for answering all the tedious questions on how to implement things in the DIVE environment. Thanks also to Erland Lewin for proof reading the thesis with greatest care.

I would also like to thank my supervisor at the Royal Institute of Technology (KTH), Kai-Mikael Jää-Aro, for the thorough and repetitive proof-reading sessions on this thesis as well as giving me pointers to literature of interest. Finally I would like to thank my supervisor at SICS, Lennart Fahlén, for supplying me with interesting papers and books on the subject.

# TABLE OF CONTENTS

# CHAPTER ONE
# THREE-DIMENSIONAL AUDIO

'Three-dimensional audio' – isn't that just another tautology?

That depends very much in what context you consider it. All the everyday sounds around us certainly are three-dimensional in the sense that they have a *spatial position* that we, more or less, are able to judge. However, when considering sound images reproduced on TVs, home stereos, computer speakers, and similar equipment, the three-dimensional image collapses into a single sound source right in front of us or, as in stereo, to a point on a line between our loudspeakers.

***Spatial position:*** *A position in the space around us.*

However, techniques and technologies to inexpensively produce three-dimensional audio are emerging, and it is expected that we soon will have them in our homes.

I will open this master's thesis with a definition of what 3D audio is. I continue by describing some theory on hearing, followed by a presentation of some models for implementing 3D audio. I will also describe how the implementation of 3D audio in the virtual reality system DIVE was done. Finally, I present some results and conclusions along with proposals for future work.

## 1.1. WHAT IS THIS 3D AUDIO?

In recent years, several technologies have been presented to the market as '3D audio' equipment. This has led to a considerable confusion as to what this term means. As a start it, is beneficial to bring clarity to some terms and concepts around the notion of 3D audio. The contents of this summary are mainly from Schneider (1996).

### *Mono and Stereo Enhancers*

The aim of the enhancement technologies is to create a more *spacious* sound field out of an existing mono or stereo *soundtrack or mix*. This is especially useful on narrowly located speakers, like on TV sets and multimedia computer systems. The effect can for example be achieved by adding a very small delay on one channel and mix it in at a reduced volume in the other channel and vice versa.

***Spacious:*** *Here: giving the illusion that the room enclosing the sound source and the listener is bigger than it really is.*

***Soundtrack or mix:*** *Here as examples of a complex sound, the parts of which you cannot control individually.*

The enhancement effects are sometimes added at the mixing stage of a soundtrack or song, but are usually found as add-on systems. The effects are user controllable, in similarity with loudness and balance controls. This technology is sold today in many forms as ‘3D sound’ or ‘multi-dimensional sound’.

### *Surround Sound*

The Surround Sound system, which today is found on almost every new home stereo system, is also referred to as “Dolby ProLogic”. Although not labelled as a ‘3D system’, it is often referred to in discussions, because of its ‘multi-dimensional’ characteristics.

The ProLogic technique is an encoding/decoding scheme designed to add an extra, ambient, channel to the left and right channels found on traditional multi listener environments like movies and TVs. The scheme was mainly created to enable storage and broadcast of the extra channel over the two ordinary stereo channels. The ambient surround channel is encoded into the normal stereo mix at mastering, and decoded at playback.

When encoded, the extra channel is superimposed over the two other channels. The trick is that the superimposed ambient channel cannot be heard if the encoded channels are played on a stereo that does not support the ProLogic scheme. This means that the superimposed channel is only heard in the ambient speakers if played through a ProLogic-equipped stereo. Thus, the left and right channels are always kept intact, resulting in a preserved *frontal sound image*.

***Frontal sound image:** The left and right frontal channels of a movie normally include the dialogue, visually related sound effects, and background music. Put in the ambient surround channel are sounds whose only purpose is to make the sound image more convincing, like birdsong and rain (see Begault, 1994, p. 22).*

The effect of surround sound is a more immense and convincing auditory display than normal stereo. Though the overlaying results in an ambient channel that is not completely independent of the two frontal channels, it is enough to convey the effect of enhanced depth in a sound image compared with ordinary stereo.

### *Binaural Audio and Interactive 3D Audio*

A further step is to be able to control the positioning of a sound source in three dimensions, i.e. its angular location and its distance relative to the listener. Such a process takes mono audio signals as input and produces a two-channel sound stream as output. The left and right channels, usually played over headphones, are what the listener would hear if the sound source was placed in that position in the real world. These left and right channels are usually referred to as ‘binaural audio’. It is also possible to process the signals further to enable them to be played over loudspeakers with a conserved 3D sound image.

When updating the positions in real-time, or at least at a rate of 20 Hz, the binaural audio goes interactive. Every time the process is run, the new positions of the sound sources are taken into account and a sense of movement is achieved. The high update rate also makes it reasonable to add velocity dependent effects like *Doppler shifts*, which add even more realism to the sound image produced. The position and angle of the listener can also be accounted for by attaching a head-tracking device.

***Doppler shift:** The phenomenon perceived when for example a police car is driving by and the siren’s frequency becomes shifted.*

The processing of multiple sound sources is done by processing each sound source separately. Then all the sound sources’ individual outputs are mixed together before finally being played to the listener. All this enables the control of the position of every sound source individually and ultimately creating a complex sound image.

Throughout the rest of this document, I will refer to binaural audio or interactive 3D audio when using the term 3D audio.

# CHAPTER TWO
# OVERVIEW OF AUDITORY PERCEPTION

Through the years of evolution, the human *auditory perception system* has become a very impressive sensory system. Its early function was probably as an omnidirectional warning system but it has developed to be one of the key components of human communication.

***Auditory perception system:** Consists of the external, middle, and internal ear, and the neurological parts in the brain that deals with auditory tasks such as localisation, recognition, and discrimination of different sounds.*

This chapter begins with an introduction to the physiology of the ear and the functions of its different parts. The two following sections discuss the basic interaural cues and some troubles when only using these to spatialise sounds. The section thereafter discusses the more advanced head related transfer function (HRTF) approach and how it can enhance localisation further from the interaural cues. The chapter is concluded with two sections discussing the distance and environmental cues like echoes and reverb.

## 2.1. HEARING

The wide range of frequencies that humans can perceive stretches from about 20 to 20,000 Hz, and is sensitive to fluctuations of circa 0.2%. Even more impressive is the auditory system's dynamic range with a capability up to about 110 dB(A), comparable to a starting jet at a distance of 50 meters. The ability to discriminate differences in fractions of a decibel in combination with the large dynamic range, is something that only highly specialised electronic systems can emulate.

The auditory perception system also has an enormous strength in its adaptability (learning capabilities). Take for example a person that has been able to hear all of his life and that over a short time loses his hearing due to some complication, but where the function of the nerve from the cochlea (see Figure 2.1) still is intact. He can now have a so called *cochlea implant* that picks up the sounds around him using an external microphone. The implant processes the sound and stimulates the nerve ends in the cochlea resulting in giving the neurological components some input. This input, however, is not the same stimulus as the brain is used to. The point is that the person is able to over again learn how to hear with the new stimuli provided by the implant. Naturally not as well as before, but some people can even communicate over a telephone with this kind of implant. This is made possible due to the qualities of the nervous system in the brain. The interested reader can find more on the adaptability of the nervous system in

***Cochlea implant:** An implant consisting of a microphone, some electronics and nerve end stimulator. There is to some extent a discussion in the deaf society about the benefits of this kind of implants, but that is beyond the scope of this text.*

Reichert (1992, Ch. 9), and can definitely satisfy his hunger on the physiology of the ear in Sullivan (1996).

Finally as a last example of the impressive discriminating capabilities of the auditory perception system there is the cocktail party effect. Imagine yourself in a room full of thirty people or so, standing in small groups talking to each other. You might be deeply involved in a conversation, but you will certainly notice if your name is uttered in one of the other groups.

*Physiology of the Ear.*

Below is a cross-section of the ear. It is divided into three anatomical regions: the external ear, the middle ear and the internal ear.

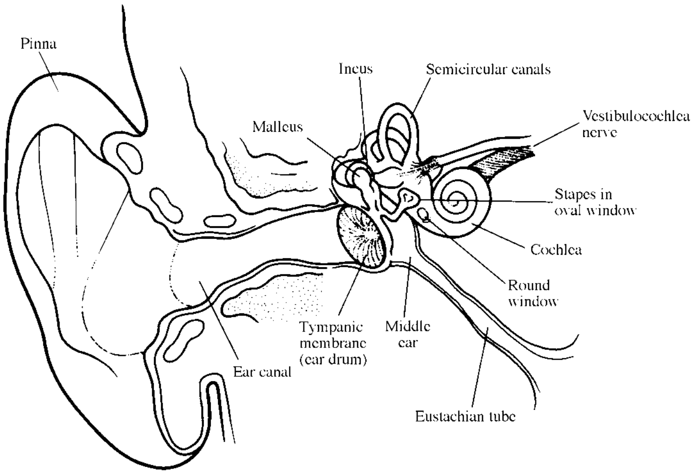


**Figure 2.1** Anatomy of the ear (adapted from Kalawsky, 1993).

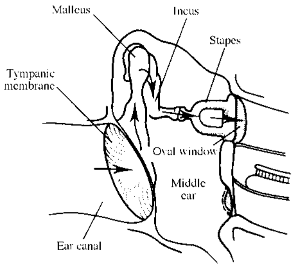


**Figure 2.2**
Impedance transformation
(adapted from Kalawsky, 1993).

The external ear collects the sound waves with the pinna and guides them through the ear canal onto the ear drum (tympanic membrane). The ear drum is linked to the cochlea's oval window by the malleus, incus and stapes. Together they work as an impedance transformer (Figure 2.2), converting the waves in the low impedance air to movements in the high impedance liquid in the cochlea. The amplification needed for this impedance transformation is achieved due to the ear drum having an area about 22 times larger than the area of the oval window.

The vibrations that originated from the sound waves have now been transferred to the cochlea. In the cochlea, the movement of the liquid causes the hairs of the basilar membrane to move as well. They in turn are connected to nerve ends causing nerve impulses to be sent to the low level auditory sections of the brain (cerebral cortex). The low level information of the nerve signals from the ear is processed and results in some higher level information passed on to other parts of the brain. The higher level information can for example be of the type "what object caused the sound" (recognition, discrimination) or "from where did the sound originate" (localisation).

*The Sound Waves' Interaction with the Pinna*

A sound signal becomes distorted in many ways when interacting with the pinna. The pinna acts as a linear filter whose transfer function is dependent upon direction and distance to the sound source. Hence, the pinna is encoding the sound’s spatial parameters into temporal and spectral attributes. The physical properties of the pinna include interference, *diffraction, refraction*, masking, reflection and resonance. All the spatial encoding is actually taking place outside the ear canal. Møller (1992) shows that the air pressure ceases to be spatially dependent even a couple of millimetres outside the ear canal.

***Diffraction, refraction:*** *Respectively, the bending and the breaking up of sound waves as they pass by some physical object. See further in section 2.6.*

All these effects can be summarised into something called head related transfer functions. These are described later in section 2.4.

## 2.2. Basic interaural cues

The direction of a sound source can be divided into a horizontal plane component and an elevation component. A human’s ears are symmetrically placed in the horizontal plane, thus giving him a much better localisation capability for sound sources in the horizontal plane. The physical properties described above are mainly affecting the elevation component.

When perceiving directions in the horizontal plane, the two primary cues are the *interaural* time and intensity differences. When a sound is played to the right it arrives sooner to the right ear than to the left, causing the time difference. Also, the head shadows the left ear, causing the sound to be louder at the right ear.

***Interaural:*** *Relating to the combined effects of listening with two ears.*

### *ITD – Interaural Time Difference*

As mentioned, the interaural time difference is caused by the sound source being closer to one of the ears. The obvious mathematical expression for this is $\Delta t = \Delta d / c$, where $\Delta t$ is the time difference, $c$ is the velocity of sound in air (approx. 343 m/s at 20°C), and $\Delta d$ is the distance difference.

In the simplest case the distance difference is calculated by considering the direct paths from the sound source to the ears, thus excluding the head (see Figure 2.3). In the figure, D is the diameter of the head and $\theta$ is the angle of incidence (also called azimuth) of the sound waves, which are considered planar. The distance difference is given by $\Delta d = D \cdot \sin\theta$.

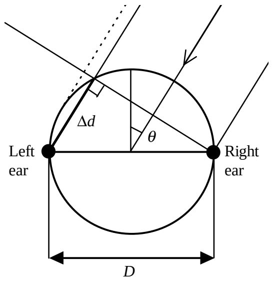


**Figure 2.3**
Direct path calculation.

A more elaborate approach is to consider the distance from the ear along the arc up to the point where the sound source becomes visible (the dotted line). Furthermore, three different cases must be considered. First, the case just described; a distant sound source implying planar sound waves. In this case the difference in distance becomes $\Delta d = D(\theta + \sin\theta)/2$. The other cases are a very close sound source with both paths to the ears bent and a close sound source where one of the ears is visible. The distance difference calculations for these two cases are left as an exercise, or can be found in Linderhed (1991).

The time difference is zero for sounds directly ahead and behind, and about 0.63 ms (Burgess, 1992) for sound sources to the far right or left. If one uses the first (simplest) formula given above and inserts 0.18 m for head diameter (the approximate size of a normal head), one will get 0.53 ms, giving a clue of the error degree in that formula. The time difference varies as a sinusoid with azimuth but is also dependent on frequency because of the head diffraction. This make sounds below 1.6 kHz manifest as a time difference, but above that, sounds exhibit envelope delays.

*IID – Interaural Intensity Difference*

The intensity difference comes from the head shadowing the sound, i.e. the head being in the way of the sound waves (see Figure 2.4). This effect begins to operate at frequencies above 1.5 kHz and becomes more apparent with higher frequencies. Relative intensity levels can approach 40 dB in one ear compared to the other.

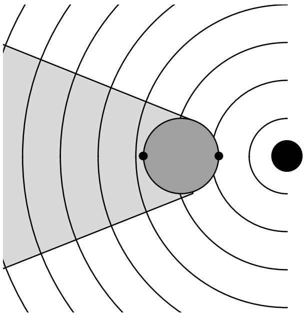


**Figure 2.4** Head shadow.

The increasing attenuation with higher frequencies is due to the higher frequencies not being able to diffract (see section 2.6) as effectively as lower frequencies.

For instance, a 3 kHz sine wave at one's far side (90 degrees left or right) will be attenuated by about 10 dB. A 6 kHz sine wave will be attenuated by about 20 dB and a 10 kHz sine wave by about 35 dB (Begault, 1994, p. 41).

Variations in the overall difference between right and left intensity levels at the eardrum are interpreted as changes in sound source position, independent of frequency content. Consider an ordinary stereo recording played over headphones, where the only way to position a sound is by twisting the balance knob on the mixer board. The balance controls interaural intensity differences without regard to frequency, yet it works for separating sound sources in most applications. This is because typical sounds usually include frequencies both above and below the theoretical frequency limits, and listeners are sensitive to IID cues down to at least 200 Hz (Blauert, 1983).

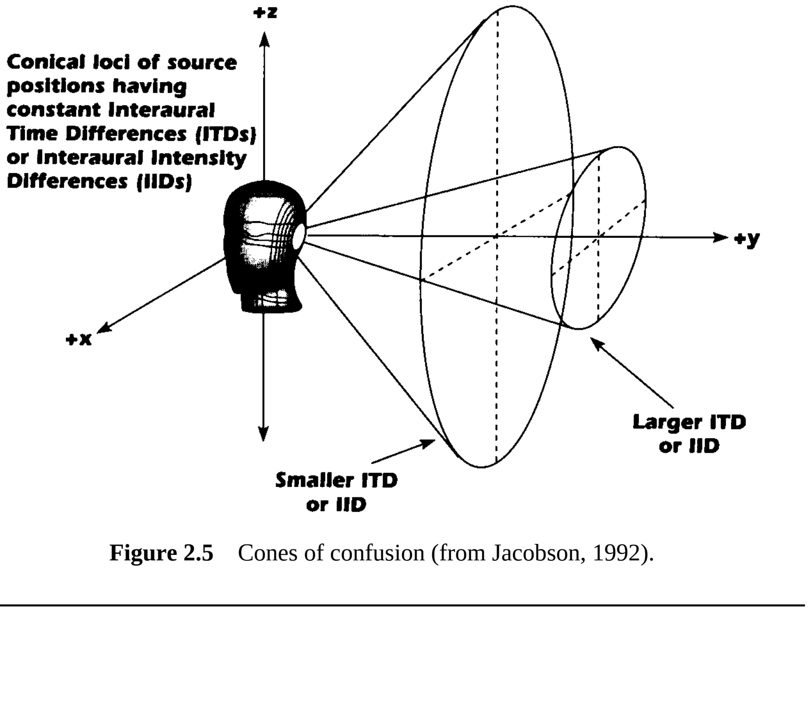
## 2.3. Ambiguities in hearing

When synthesising spatial positions of sound sources, there are some errors that listeners are prone to do. What one must bear in mind is that a listener can make these errors in a real situation as well, and that he will not perform any better in a test than he does in reality.

*The Cone of Confusion*

A problem that arises when spatialising sounds with only ITD and IID cues, is that there is a lot of ambiguity in the spatial information. The ITD and IID values are constant on a cone around the *interaural axis*. This cone is called the cone of confusion (see Figure 2.5).

***Interaural axis:*** *The axis passing left to right through our head and ears.*



+z
Conical loci of source positions having constant Interaural Time Differences (ITDs) or Interaural Intensity Differences (IIDs)
+y
+x
Larger ITD or IID
Smaller ITD or IID

**Figure 2.5** Cones of confusion (from Jacobson, 1992).

For instance, if you increase the intensity or time difference, the sound source's position may tend to move farther out, but its true three-dimensional position is still ambiguous. When the interaural time and intensity are your only location cues, any point on the entire conical surface is a possible position of the sound source.

The cone of confusion is especially noticeable for sound source positions mirrored in the plane separating front from back. This means that a position in front of you could also be a position at your back. This ambiguity can result in something called front-to-back reversals, which is quite usual when synthesising 3D sound.

There are some means to lessen the extent of this ambiguity. The means are based on better models of the head and/or ears but include the basic interaural cues. Examples are HRTFs (section 2.4) or some algorithmic model considering some of the properties imposed by the head shadowing like diffraction (section 3.2).

### *Inside the Head Localisation*

When listening to a sound with headphones it is quite normal to hear the sound inside your head, in other words it is not externalised. If you modify the interaural time and intensity difference values, the sound seems to be moving around inside your head, somewhere between your ears.

The lack of externalisation is due to the sound reaching our ears not being consistent enough with what it should be when coming from an external source. One spatial component that is clearly missing when playing a sound through headphones is reflections from objects around us. It has been shown that a sound containing *reverberant* information is perceived to be externalised to a much higher extent than a sound without reverb, but that this also implies that the localisation accuracy is decreasing (Begault, 1992).

***Reverb:** Room echo; reflections coming off the walls in a room providing the listener with information about the qualities of the room.*

Another method to increase the externalisation is to provide the listener with a better model of his pinnae (outer ears) by synthesising the head related transfer function (see section 2.4).

### *Head Movement and Moving Sound Sources*

When we wish to localise a sound in an everyday situation, we move our head in order to minimise the interaural differences. We use our head as a kind of pointer to take bearing on the heard sound source. Some animals use movable pinnae for this purpose.

Studies have shown that allowing a listener to move his head improves the localisation ability and lessens the amount of reversals (Begault, 1994). When the listener moves his head, the differences in interaural cues tell him that he is turning in either the wrong or the right direction, thus eliminating the interaural ambiguities.

A problem with synthesised 3D sound is that some kind of tracking device must be attached to the head, correcting for head movements, otherwise the cues of head movement expected by the listener will confuse him instead of help him.

Just as head movement provides the listener with a dynamic change for fixed sources, a moving sound source will cause dynamic change for a fixed head. This assists the listener in localisation in the same manner as head movement does. Another cue provided by moving sound sources is the Doppler effect. It is a shift in frequency following the relative velocity of the sound source. This cue is very basic to our perception system and thus it is firmly rooted as a cognitive cue as well (see below).

*Visual and Cognitive cues*

Whether we provide the listener with an incredibly sophisticated model or just a mono sound, the listener's mind is always in charge over the perceived location and distance. Expectation and memory highly influence the judgement of localisation. Visually acquired stimuli can also modify auditory localisation.

Cognitive cues are even more persuasive. There are examples of demo tapes where it was claimed that the producers had overcome the front-back reversals and which included a sound of someone drinking a glass of water. It really sounded as if it was in front of you, but when was the last time you tried to drink a glass of water from behind your head? The point is that if you are expecting a sound to originate from some direction, either because your experience says so (memory) or because a visual cue indicates it, then you have already decided the direction of the sound and only really convincing auditory cues will make you change your standpoint.

As a last example of the strength of cognitive cues, there is a 3D demo of race cars zipping by, conveying the impression of sound movement. There is just one minor objection one can have to this, and that is that this demo *always* works – even with mono systems! The examples were taken from Jacobson (1992).

## 2.4. Head related transfer functions

The term head related transfer function (HRTF) refers to the spectral filtering mainly caused by the outer ears (pinnae). The HRTF can be thought of as frequency- and direction-dependent amplitude and time-delay differences primarily resulting from the complex shape of the pinnae (see Figure 2.6).

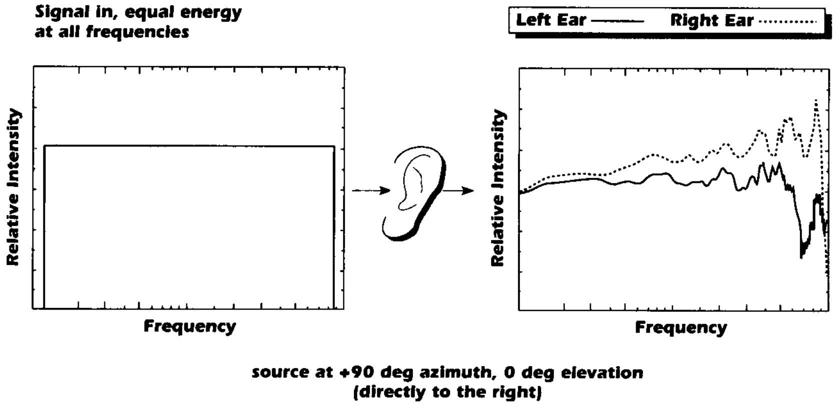


**Figure 2.6** Spectral shaping by the pinnae (from Jacobson, 1992).

*HRTF Components and their Characteristics*

The most important element of the HRTF components is the pinna. The asymmetrical complex shape of the pinnae causes microsecond delays, resonances and diffractions, that transform every direction to a unique *spectral filter*. These unique filters are what a listener to some degree recognises as a spatial cue.

**Spectral filter:** *Description of how to alter the amplitudes of different frequencies in a sound. This is exactly what an equalizer on a home stereo system does; emphasising some frequencies and attenuating other. In Figure 2.6, the pinnae are functioning as spectral filters.*

The largest resonant area of the pinna is called the cavum conchae (Figure 2.7) and is asymmetrically located around the entrance to the ear canal. The asymmetric location of the cavum conchae causes the delay of the first reflection from the cavum conchae's walls into the ear canal to be a different amount of time for different directions.

Other components of the HRTF are head diffraction and reflection, reflections from shoulders and from the torso. All of these parts are operating in different frequency ranges because of their different sizes affecting different wavelengths. Below is a list of what range each element is most likely to affect (from Begault, 1994). All figures are (naturally) approximate.

- Pinnae and cavum conchae reflections: 2–14 kHz.
- Head shadow and diffraction: 0–20 kHz.
- Shoulder reflection: 0.8–1.2 kHz.
- Torso reflections: 0.1–2 kHz.

### *A Do-It-Yourself Experiment*

To get some hands-on experience of what the pinna does for us, let us try a couple of exercises that some of us might already have tried. The best effect is achieved sitting closely to a broadband sound source. A broadband sound is a sound with a wide frequency content, for instance blank-channel noise on a TV or fan noise from a computer at your desk. Close your eyes and turn your head to focus the sound source to be right in front of you. This is the normal condition with the pinnae unblocked. Try the following exercises and listen what happens with the sound's spectral contents and its position.

1) With both hands slightly cupped and fingers together, create a "flap" in front of both ears shadowing sound from the front. The result is as if you had large reversed pinnae focusing sound from the rear.
2) Blocking one ear's opening and turn your unblocked ear towards and away from the sound source.
3) Flatten your pinnae back against your head using your fingers.
4) Cup your hands like in 1) but focus forward instead, thus enlarging your pinnae. Also try to turn your hands a bit, hence altering the auditory focus.

You should clearly hear the tone colour change between normal listening and the exercises, especially some muffling (exclusion of high frequencies) in case one and two and some sharpening (emphasis of high frequencies) in case four. Case two is a very good example of the head shadow effect.

Regarding the position of the sound source it might be somewhat hard to make the "cognitive leap of faith" since you already know where the source is. However, some people do observe spatial effects when trying these exercises. Condition one can move the sound source to the rear. Number three might spread the position of the sound source causing its location to be more diffuse, and number four makes the sound louder and can thereby reduce its perceived distance. The moveable pinnae of case four can be seen on some animals, cats for instance, who bring the sound source into focus in this way. The test was adapted from Begault (1994), except for 2) which was the author's own example.

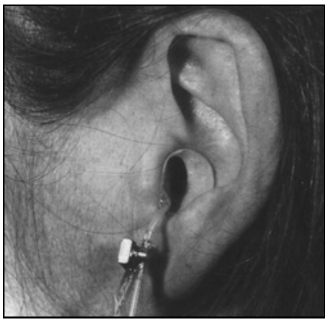


**Figure 2.7** Placement of probe microphone in the cavum conchae (from Wightman et al., 1989a).

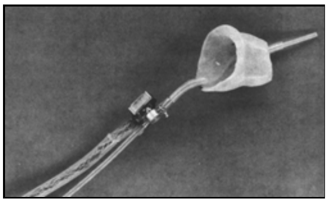


**Figure 2.8** Close up of the probe microphone used in Figure 2.7.

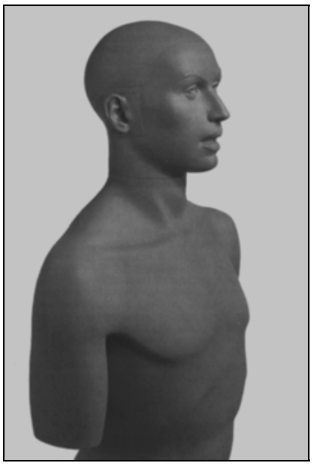


**Figure 2.9** The KEMAR mannequin head and torso (from Begault 1994, originally from Knowles Electronics).

### *Measuring HRTFs*

The collecting of data for HRTFs is a highly time-consuming and difficult process. Also, there is no perfectly accurate method for measuring HRTFs.

When performing a measurement, there are basically two ways of doing this. The first is to use probe microphones inserted into a subject's ear (see Figure 2.7 and 2.8). The other is to use a mannequin head and/or torso with built in microphones at the end of the artificial ear canal (see Figure 2.9). The prime advantage of the mannequin is its exchangeable pinnae, thus making it possible to measure different pinnae on the same subject, so to speak. The possibility of changing the pinna also makes it possible to measure two different pinnae at the same time. Additionally, the mannequin will not become tired and will consequently remain still.

The procedure of collecting HRTF data is as mentioned quite cumbersome. The objective is to find out what the HRTF is for different spatial positions. The subject's head is placed in a position regarded as the centre. Now, a stand with speakers mounted equidistantly at different elevations can be used. The stand can be moved on an arc around the subject. Possibly, a single speaker can be used and moved around, or perhaps a permanent sphere with equidistantly positioned speakers all around it. Now a known broadband sound is played through one speaker at a time and recorded by the probe microphone. This is done for all the positions one wants to measure. Normally the distance of the speakers from the head is about 1.5–3 m and the angular displacement about 15–20 degrees.

Since the interesting component of the recorded signal is the head related transfer function, the spectral characteristics of the speakers and the probe microphone imposed on the recorded signal must be removed. The probe microphone, for instance, has a very bad response for lower frequencies, and the speaker can have just about any response curve. This means that the response curves of the microphone and speakers must be measured in advance in order to be corrected for at this point. For a discussion on how the filtering (both the HRTF filtering and the correction of microphone and speaker) is done, see section 3.3.

### *HRTF Directional Characteristics*

When inspecting HRTF diagrams there are certain characteristics that stay the same between different subjects (see Figure 2.11 for an example of how different two subjects' HRTFs can be). The distinguishable characteristics are for example the effects imposed by the head shadow at the turned away ear for sounds coming from behind and sounds coming from below.

The head shadow mainly affects the higher frequencies from about 4 kHz and up by attenuating them. The filtering on sounds from behind are due to the pinnae being focused forward. On these sounds to the rear, there seems to be no single characteristic, but more that each individual learns how their own pinnae work. This means that there is no common characteristic that provides us with the ability to discriminate sounds originating from the front and back. See Figure 2.10 for a graph of the difference for two sounds being on the cone of confusion for one individual. The graph indicates that HRTFs might solve the troubles of the cone of confusion that arise when only using ITD and IID cues.

For sounds coming from below, there is notable attenuation around 4-10 kHz. This is probably due to the pinnae and to some extent the torso reflecting and absorbing some frequencies. For a thorough discussion on this subject see Wightman and Kistler (1989a).

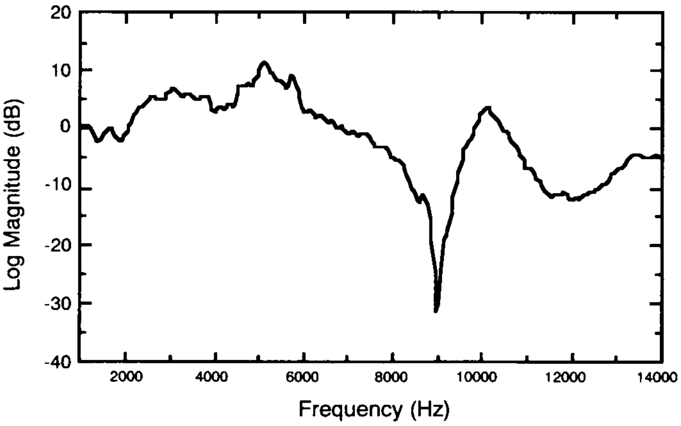


**Figure 2.10** Difference in spectra between two front-back sources on a cone of confusion, located at 60 and 120 degrees on the horizontal plane (from Begault, 1994).

*Localisation with Individual and Generalised HRTFs*

Wightman and Kistler (1989b) showed that when simulating *free-field* listening over headphones with a listener's own HRTFs, the localisation accuracy was quite good. The measuring of the HRTFs was done in an anechoic chamber (a room with very good sound absorption), thus simulating the free-field situation.

***Free-field:*** *Referring to an environment where sound waves can spread without being obstructed, as in a large flat field coated by grass.*

A problem with this is that every user has a unique set of HRTFs (Figure 2.11) and in many situations, as in multi-user systems, it is not feasible to have every user listening through his own set, which is mainly due to the problems related in measuring the HRTFs. This is why some research focuses on the question whether it is possible to obtain a set of generalised (non-individualised) HRTFs where a majority of the users perform adequately.

Studies have shown that there are "good" localisers and that there are "bad" localisers. In Wenzel et al. (1988), for instance, there is a comparative study between two good localisers and one bad localiser. The good localisers showed good accuracy in judging elevation and horizontal position both when listening to real sounds and when listening to synthesised stimulus through their own HRTFs. The bad localiser showed little ability to determine elevation in either case. When using non-individualised HRTFs, the accuracy of the good localisers was only slightly degraded, as long as the non-individualised HRTF was derived from another good localiser. Large errors in judging source elevation were made by a good localiser when listening to synthesised stimuli from the bad localiser's HRTFs (i.e. listening through the ears of the bad localiser). However, the converse was not true. The poor localiser was not able to improve his performance in elevation accuracy by listening with a good localiser's HRTFs.

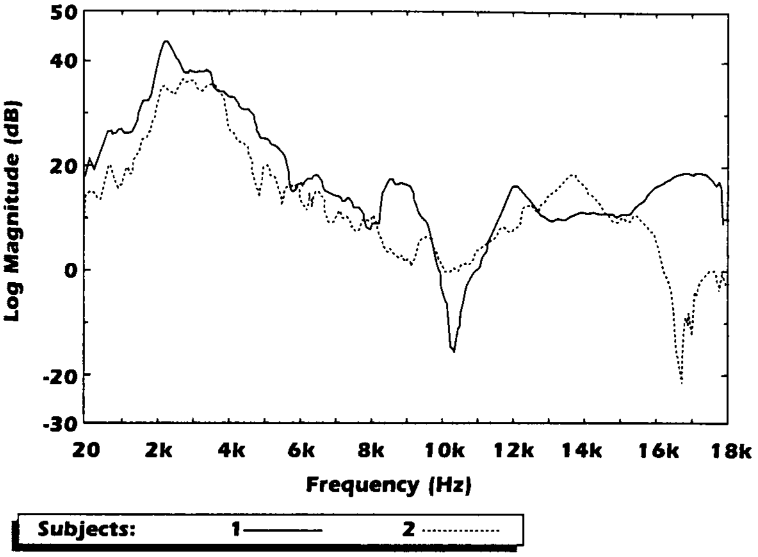


**Figure 2.11** Two people's HRTFs measured at the same position in an anechoic chamber (from Jacobson, 1992).

Another study by Wenzel et al. (1993) shows that front-back and up-down confusions increased significantly when subjects were listening through non-individualised HRTFs. The report suggests that while interaural cues to horizontal location are robust, the spectral cues considered important for resolving location along a particular cone of confusion are distorted when produced by a synthesis using non-individualised HRTFs.

The question is how to produce non-individualised HRTFs with a good result. Begault discusses this in Jacobson (1992). If one averages the difference of many people's HRTFs, then the distinct individual features of the HRTFs are removed and one ends up with something in the middle that does not provide any spectral cues. In Begault (1994, p. 140) it is suggested that the best solution is probably to use a single non-individualised HRTF that is statistically validated and thus proves to give a majority of users an acceptably accurate localisation rate.

## 2.5. Distance cues

One important cue for distance used by a listener is the intensity of a sound source. Further cues used for distance perception are reverberation and the acoustics of the room the listener is in. These cues and how listeners perceive them will be discussed in this section.

### *Intensity and Loudness Cues*

Without other acoustic cues, the intensity, and its interpretation as loudness, is the primary cue for distance. Auditory distance is nothing that can be judged from the beginning, but it is learned through visual-aural observations throughout life. This means that the intensity cue to distance plays a more important role in unfamiliar environments than in familiar ones. For instance, the bus passing by outside the window might be louder than the clock on the table, but despite this we know that the bus is not coming through the window into the house.

Normally the sound source is considered to be an *omnidirectional* point source. Under anechoic conditions, the inverse square law can be used to predict sound intensity reduction with increase in distance from an omnidirectional source. The inverse square law states that the intensity will be ¼ of the intensity for each subsequent doubling of the distance. The decrease in intensity corresponds to about 6 dB.

***Omnidirectional:** The emission of sound equally intense in all directions. The omnidirectional model is almost always used in virtual reality simulations. In noise-control applications on the other hand, like when modelling a freeway, a line source can be used. The sound intensity from a line source falls with about 3 dB for each doubling of distance.*

The judgement of distance based on intensity alone is difficult. A study by Gardner (1969) shows that expectation plays a big role. In the study, subjects were to discriminate from what distance certain sounds came out of four possible positions in straight line ahead of them. When the sound was whispering, the subjects always underestimated the distance, and when the stimulus was shouting, the distance was overestimated. The opposite should have been true if intensity was the relevant cue.

There are some spectral cues to distance as well. These cues are quite insignificant at short distances around ten meters, but around a hundred meters the attenuation in the 4 kHz area is around 7 dB. The attenuation is due to the absorption of high frequencies in the air and is a function of humidity and temperature.

*Reverberation*

Sounds, however, are not usually heard in anechoic environments, but in conjunction with reverberation. Reverberation is sound waves that reach the listener indirectly, i.e. reflected from surfaces in the space surrounding the sound source and the listener. In an ordinary room, the intensity level does not decrease more than 3 dB even though the distance has been doubled three times (Begault, 1994). Hence, reverberation precludes the inverse square law in these kinds of contexts. In a reverberant environment the R/D ratio (explained below) is a much stronger cue to distance than intensity scaling.

If recording a loud impulsive noise, such as a starter pistol being fired, in an acoustic system (e.g. a room), the recording will show us the system's *impulse response*. An impulse response of a classroom can be seen in Figure 2.12.

***Impulse response:** When an acoustic system is being triggered by a short pulse (Dirac-pulse), the system will answer in a certain way. This answer to the impulse is the system's fingerprint and is called its impulse response.*

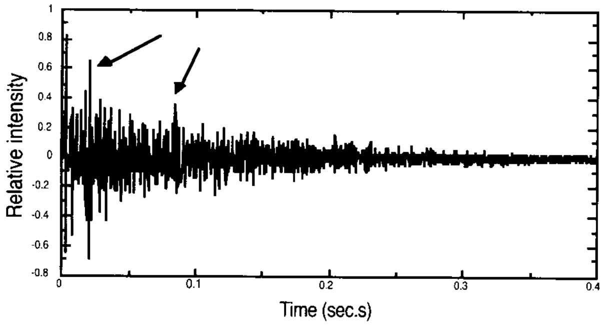


**Figure 2.12** An impulse response of a classroom. Arrows indicate significant early reflections (from Begault, 1994).

A particular reflection in an impulse response diagram (reflectogram) is usually categorised as an early reflection or as late reverberation (i.e. late reflections). The category is dependent on the time of arrival to the listener after the arrival of the direct sound.

The early reflections help us to judge the size of the room. We do not perceive these early reflections as individual sounds, but instead the auditory perception system interprets them as additional information to the initial direct sound.

The point where reflections are beginning to be categorised as late reverberation is about 80 msec after the direct sound has reached the listener. The late reverberation help us judge the acoustic qualities of the room (reverberant as in a church or sound absorbed like an office). This is a quite arbitrary value, but corresponds to and has its origin from a term called the reverberation time, or t60. The t60 is the point in time where the level of the reflections has dropped below 60 dB of the direct sound. At this point the reverberation is no longer distinguishable from the ambient noise in the room. Figure 2.13 shows a simplified reflectogram that is sometimes produced when simulating 3D environments. In the reflectogram, the early reflections might be calculated using ray-tracing and the late reverberation is just a dense synthetic (no physical basis) reverb.

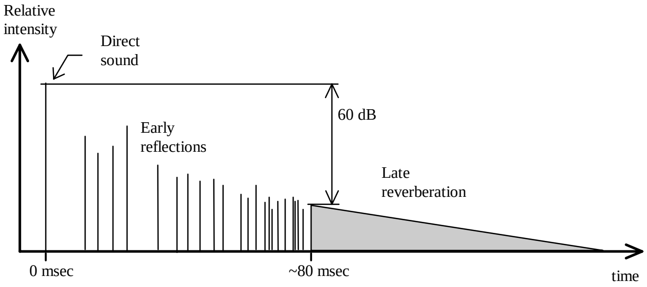


**Figure 2.13** Simplified reflectogram of an impulse response.

As stated earlier, the intensity cue is not enough to judge distance accurately, and some further cues are needed. One further cue is the R/D ratio, i.e. the reverberant-to-direct sound ratio. The R/D ratio refers to the level ratio between the direct and the reverberant sound. In studies it has been found that if the intensity is kept constant, but the R/D ratio is altered, then this will be perceived as distance changes. The cue, however, is not very robust, because in an acoustically treated room the R/D ratio will vary between narrower limits than in, for instance, a gymnasium. It seems that no single distance cue is sufficient, but that it is a combination of the different cues that makes distance judgement as accurate as it can be.

## 2.6. Some environmental effects

As sound waves travel, they can be obstructed in many different ways. Different kinds of obstacles hinder the sound waves in different manners, and impose their own type of spectral shaping to the sound. Below some aspects are explained.

*Diffraction*

When sound waves approach an object that is small in comparison to their wavelength, the sound waves start to bend on the rear side of the object. This phenomenon is called diffraction, the sound wave bends around the obstacle and progresses on the other side (see Figure 2.14).

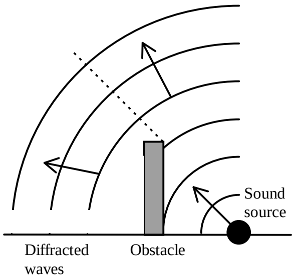


**Figure 2.14** Diffraction.

Sounds with a long wavelength (low frequency) easily pass corners and pillars, while sounds with a high frequency require free sight between the source and the listener. This is why you do not want to end up behind a pillar at a concert, apart from the fact that you will not be able to see anything either.

*Reflection*

All sound waves are reflected against obstacles if the obstacles are big enough, i.e. the sound waves cannot diffract around them. In practice these obstacles are usually walls in a room, thus providing "infinite" reflection area. The walls can consist of different materials and each material has its own reflection characteristics.

The opposite of reflection is absorption. The absorption coefficient is one minus the reflection coefficient. Technically, the absorption coefficient represents a combination of true absorption (sound energy being converted to heat), and the transmission and dissipation of sound energy to another volume (e.g. the other side of the wall).

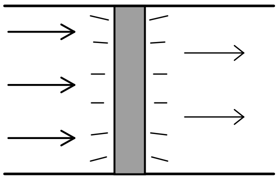


**Figure 2.15**
A wall absorbing sound energy (see text for explanation).

A room with poured concrete walls, for example, reflects close to 99% of the sound energy at all frequencies ranging from 100 Hz up to 5,000 Hz. This in comparison with velour draperies whose reflection coefficient is 93% at 125 Hz but only 30% at frequencies at 1–2 kHz.

A final example is an ordinary window, reflecting 95% at 4 kHz but reflecting only 70% at 125 Hz. This high absorbing capacity of glass at low frequencies is due to the window's measures being in the range of the sound's wavelength. Consequently, the window starts to vibrate in resonance with the sound, hence absorbing it efficiently since all the sound energy is transferred to motion energy in the window glass. All the reflection coefficients were taken from Hall (1990, p. 324).

*Resonance*

When a sound is initiated in a closed or half closed volume (again a room for example), the room starts to amplify certain frequencies. The phenomenon is called standing waves. The dimensions of the room determine which frequencies are amplified. A large room amplifies lower frequencies than a small room. The lowest frequency that becomes a standing wave is the frequency that has a wavelength two times the room's length (for a room with four walls, roof and floor).

A room with the length of 3 m will then have a standing wave with a frequency around 100 Hz. This is the reason why bathrooms in general encourages men to sing in the shower. They will be rewarded with a standing wave by just opening their mouths, also the walls in a bathroom often have very low absorption coefficients.

*Transmission*

In the case above where a wall picked up sound energy and did not convert all of it to heat, the wall dissipated some of the energy to adjoining rooms. Sometimes this can be very disturbing, especially in combination with standing waves.

An example of this is an apartment house built of poured concrete. In this construction load-bearing walls, floors and ceilings are tightly attached together. These points of attachment act as nodes in the standing waves that arise when the walls are actuated by sounds produced by people living in the house. The measures of the walls, floors and ceilings are in the magnitude that they convey low frequency sounds, thus the rumbling type of sounds your neighbours seem to cause.

Another aspect of transmission is the speed of sound in diverse media. In air the molecules are rather free and unordered and consequently the speed of sound is quite low. In water for example the speed of sound is about 1,400 m/s and in materials where the molecules have a rather ordered structure, like steel, the speed of sound is about 5,000 m/s.

# CHAPTER THREE
# THE REALITY OF 3D AUDIO MODELLING

This chapter will discuss the implementation issues of 3D audio. It will describe both spatialisation and the modelling of environmental context. The chapter will begin with an introduction to digital signal processing, making it easier to understand the rest of the chapter. The following two sections describe the two principal approaches to spatialisation, the computationally cheap algorithmic approach and the more expensive HRTF-approach. A section on how to compensate for presentation over loudspeakers is included and also a section on four channel spatial audio. Following is a section covering the issues concerning the modelling of the environmental context. The chapter is closed with a section on topics for integrating a 3D audio system in a virtual reality system followed by a section on example applications of 3D audio.

## 3.1. INTRODUCTION TO DIGITAL SIGNAL PROCESSING

This section will cover the absolute basics of digital signal processing (*DSP*) and is by no means a comprehensive summary on the subject. The interested reader can find a very detailed and thorough discussion covering digital signal processing and digital filters in Proakis and Manolakis (1992).

***DSP:*** *Can mean both Digital Signal Processing as a concept and a Digital Signal Processor, the silicon chip.*

Digital signal processing today is done mainly by highly specialised inexpensive silicon chips that can usually perform very few operations but execute them very fast. The processing might also be done in software on standard PCs, but then several magnitudes in speed/cost are lost compared with the DSP chips. On the other hand, the processing can be integrated in the normal computer environment and no additional hardware is needed.

### *Analogue and Discrete Representation of Signals*

A continuous, analogue, function like $x(t) = \sin(\omega t + \alpha)$, is a function of continuous time. When the signal produced by a continuous function is to be stored and processed in a digital system, it must first be converted to discrete form. This is done by sampling the signal, i.e. acquiring the value of the analogue signal at discrete intervals of unit time, the sampling rate. The discrete representation of the continuous signal is written $x(n) = \sin(\omega n + \alpha)$. In other words, $x(n)$ denotes the

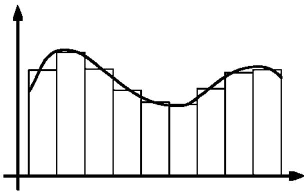


Figure 3.1 An analogue signal being sampled at discrete time intervals.

sampled version of the signal $x(t)$, with $n$ being the sample index (the sequential number of the sample), see Figure 3.1. The amplitude of the analogue signal also has to be converted to a discrete value, i.e. quantised. The sampled amplitudes can be stored as integers or floating point numbers. For a further explanation on the representation of numbers in digital form, any basic book in computer science will do.

*The Impulse Response*

In the previous chapter it was explained that the impulse response was an acoustic system's answer to being triggered by a short pulse, the Dirac-pulse. The Dirac-pulse is a signal that starts with a one and with the rest of the samples being zero. The starter pistol in section 2.5 was meant to be a fair approximation of the Dirac-pulse.

*The Frequency Domain and The Fourier Transform*

A signal can be represented both as a function in the time domain (as above), and as a function in the frequency domain. In the frequency domain the signal's magnitudes in different frequencies are shown. To transform a signal from the time domain to the frequency domain and vice versa, one can use the Fourier transform and the inverse Fourier transform respectively. In this text the Fourier transform will be regarded as a tool for converting a signal back and forth between the two domains. For a deep coverage of the Fourier transform in conjunction with DSP see Proakis and Manolakis (1992).

Below in Figure 3.2, a sketch of what spectral effects an unknown (acoustic) system might impose on a Dirac-pulse is shown. The system is denoted $h(n)$ and the transforms of the functions are denoted with their respective letters in capitals. The output $y(n)$, from the system is in this case, where the Dirac-pulse is the input to the system, the impulse response and consequently is $y(n) = h(n)$. Note the spectral content of the Dirac-pulse including all frequencies.

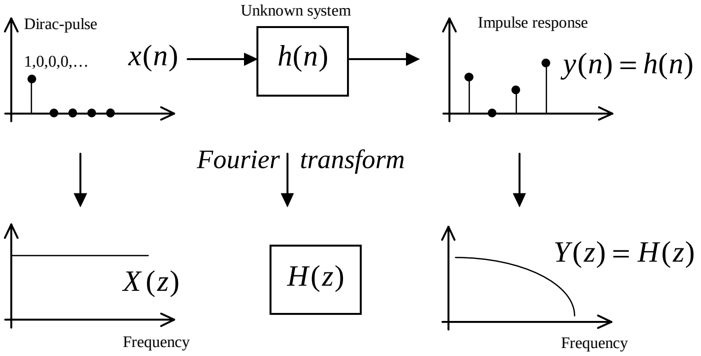


Figure 3.2 Impulse response from a system both in time and frequency domain.

*Some DSP Operations*

In audio, DSP algorithms are usually complex combinations of simple elements including multiplication, addition and delaying of a signal. The elements' schematic and mathematical representations are shown in Figure 3.3–3.5.

Figure 3.3 shows a multiplication element. The factor $g$ can be any value, even negative. If the product of the multiplication is too large to fit in the assigned computer memory, the value will be clipped, which becomes audible as distortion. For a series $x(n)=\{1,3,0,-2,0,\ldots\}$ and $g=2$, the output will become $y(n)=\{2,6,0,-4,0,\ldots\}$.

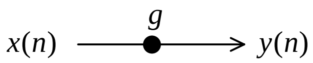


**Figure 3.3** Multiplication of a signal: $y(n)=g\cdot x(n)$.

The second operation, delaying of a signal, is shown in Figure 3.4. The delay operation holds a value for a certain amount of time, stated in samples. Say that you want to build a system that delays the signal by one second and that the sample rate is 50 kHz, then the delay buffer would have to be 50,000 samples long (i.e. be able to store 50,000 samples). For a series $x(n)=\{1,3,0,-2,0,\ldots\}$ and D=2, the output is $y(n)=\{0,0,1,3,0,-2,0,\ldots\}$.

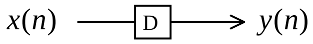


**Figure 3.4** Delaying of a signal, D=1: $y(n)=x(n-1)$.

Finally, there is the digital summation operation, schematised in Figure 3.5. The summation operator can theoretically take any number of inputs, but will presumably be implemented by summing the inputs one by one. The same goes for the sum as for the product in the multiplication operation above; it will be clipped if it becomes too large. For the two input sequences $x_1(n)=x_2(n)=\{1,2,3,4,\ldots\}$, the output is $y(n)=\{2,4,6,8,\ldots\}$.

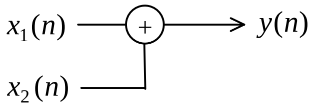


**Figure 3.5** Addition of two signals: $y(n)=x_1(n)+x_2(n)$.

*Filtering and Convolution*

The basis of 3D audio processing lies in imitation of the spatial cues present in natural spatial hearing. In natural spatial hearing, the HRTF imposes spatially dependent spectral modifications and time delays on the incoming sound waves, making it possible for the hearing system to judge the direction of the sound source.

The spectral modifications caused by the HRTFs can be simulated by digital filtering. What filtering does is multiplying the spectra (curves in the frequency domain) of two signals. If one of the signals is the sound and the other signal is the filter, then the filtering results in a filtered signal, see Figure 3.6.

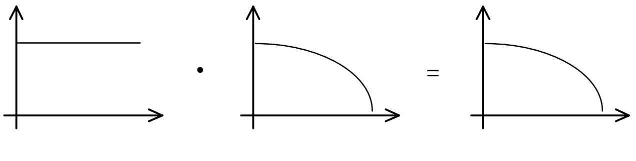


**Figure 3.6** Filtering of a signal in the frequency domain, by multiplying the Dirac-pulse to left with a low-frequency passing filter, in the middle.

The multiplication in the frequency domain is equivalent to something called convolution in the time domain and is denoted $*$. The equivalence can mathematically be expressed as $x(n)*h(n) \Leftrightarrow X(z)\cdot H(z)$. The convolution can be though of as a multiplication and summation operation performed on two numerical series (arrays). Its mathematical expression is

$$y(n) = x(n)*h(n) = \sum_{k=0}^{n} h(k)x(n-k), \text{ where } x(n){=}h(n){=}0 \text{ for } n{<}0.$$

Note that to produce one new output sample, $y(n)$, it takes $p$ multiplications and summations, where $p$ is the length of the filter $h(n)$, since $h(n)=0,$ for $n \geq p$, and thus is pointless to calculate.

This means that to convolve a signal at 32 kHz in real time with a filter length of 1,000 samples (31.25 ms), it takes 1,000 operations per out-sample multiplied with 32,000 output samples per second which results in $32\cdot 10^6$ multiply-and-accumulate operations per second. The high calculation demand is the reason why the specialised DSP chips are so commonly used.

When convolving audio signals with recorded HRTFs, i.e. filtering audio signals to simulate the natural spectral characteristics of the HRTFs, the HRTF filters are commonly about some hundred samples long. In commercial products aimed at the mass-market (like room simulators for home stereos and 3D audio equipment for computers), the filters are seldom any longer than hundred samples, hence keeping down the computational needs and the cost of the hardware.

## 3.2. Algorithmic modelling

The algorithmic model is a computationally cheap implementation of the spectral cues provided by the HRTF. It aims at extracting the stable and generally applicable spectral and time-delay features, and implement them with as simple elements as possible.

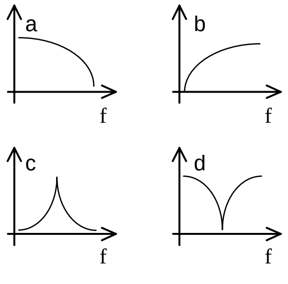
### *Which are the General Cues?*

In the list below there are some of the more powerful and stable cues. Included are of course the interaural cues, but also distance cues and some spectral cues regarding the head shadow and pinnae. The contents of the list are mainly from Pope and Fahlén (1993).

- Interaural time difference.
- Interaural intensity difference.
- Sound intensity according to distance.
- *Low-pass* filter for sounds behind the listener.
- Low-pass filter for the ears farthest away from the listener, due to the head shadow.
- Low-pass filter for sound sources far away, due to absorption.
- A fairly broad *band-reject* filter around 2-4 kHz combined with a mild boost (*band-pass* filter) in the 7 kHz region (Pope and Fahlén, 1993) to simulate elevation of the sound source.
- R/D-ratio factor (see section 2.5) to enhance the impression of distance.

Some synthetically produced reverb can also be added to increase room sensation and to enhance externalisation.

***Low-pass:*** *A filter that attenuates the high frequencies(a), i.e. passes the low frequencies. There are also high-pass filters(b), passing high frequencies and band-pass/reject filters (c, d) that pass and reject certain frequency ranges.*



*How are the Cues Implemented?*

First, an implementation of the simplest low- and high-pass filters will be shown, and subsequently a realisation of ITD, IID and some filters is given.

The output of a filter can depend on either some of the values put into the system ( $y = f(x)$ ), where the input can be delayed inside the system, or both the input and the output ( $y = f(x, y)$ ), where the output might be delayed and then fed back into the system. The order of the filter is a measure on how long the samples are stored in the filter. The filters showed here are both of the first order, since the samples are delayed by only one sample (time unit).

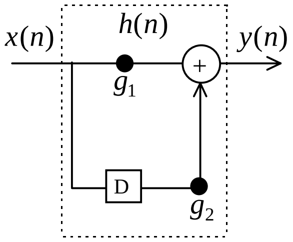


**Figure 3.7** Configuration for a FIR (feed forward) filter where $y(n) = g_1 x(n) + g_2 x(n-1)$.

The first system (Figure 3.7), where the output is dependent only of the current and/or old input values, is called a Finite Impulse Response (FIR) filter or a feed forward filter. The latter system (Figure 3.8) is called an Infinite Impulse Response (IIR) filter, or a feed back filter. It is quite obvious why they are called feed forward and feed back filters when one looks at the figures. The FIR (feed forward) filter does not have a loop back of output values, but can only delay its input a certain amount of time and then use it in a calculation. An IIR (feed back) filter on the other hand can use its calculated output values and feed them back into the system, which is also why it is called infinite; a value that has been input to the system affects the output of the system for the rest of its lifetime.

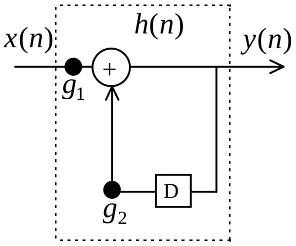


**Figure 3.8** Configuration for an IIR (feed back) filter where $y(n) = g_1 x(n) + g_2 y(n-1)$.

The FIR and IIR filters have different spectral characteristics depending on their different construction but also dependent of the different values on $g_1$ and $g_2$. Below in Figure 3.9 and 3.10 are the transfer functions shown. Note that if the coefficients in the IIR filter are too big the system will produce bigger and bigger values, thus eventually overflow.

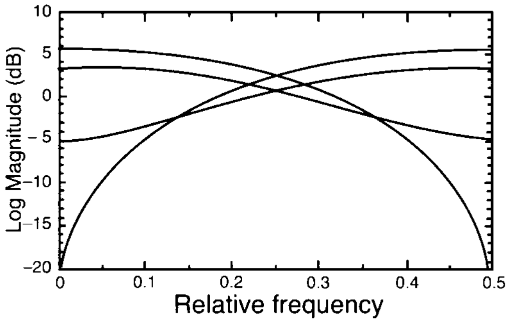


**Figure 3.9** Transfer functions of the FIR filter shown in Figure 3.7 with the gain coefficient g$_1$ set to 1, and g$_2$ set to 0.9 and 0.45 (the low-pass filters) and -0.45 and -0.9 (the high-pass filters). (From Begault, 1994).

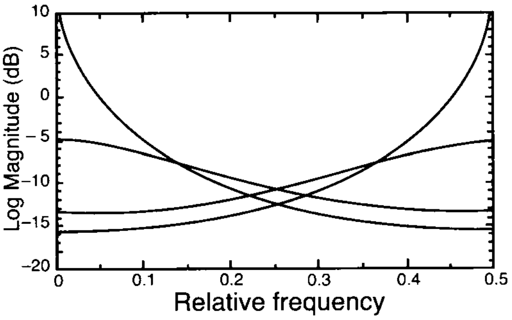


**Figure 3.10** Transfer functions of the IIR filter shown in Figure 3.8 with the gain coefficient g$_1$ set to 0.9 and g$_2$ set to 0.9 and 0.45 (the low-pass filters) and -0.45 and -0.9 (the high-pass filters). (From Begault, 1994).

By combining these kinds of simple elements, one can produce much more complicated transfer functions with arbitrarily placed peaks and notches through the use of so called biquad filters. Biquad filters are beyond the scope of this text and the reader is again referred to Proakis and Manolakis (1992). We shall later in section 3.6 see how an IIR filter with a longer delay can be used as a reverberation module, since the feed back operation is analogous to multiple reflections of sound in a room.

We shall now see how the simple DSP elements showed earlier can be coupled together to produce a schema that models both the ITD, IID and some of the filter cues. The values that can be used in the gains, delays and filters are not given here, but instead in chapter four where I present my model including the values I used. Note that filters can be coupled in series since they just multiply their transfer function onto the input signal and can thus also be connected in an arbitrary order. Below in Figure 3.11 the schema is shown.

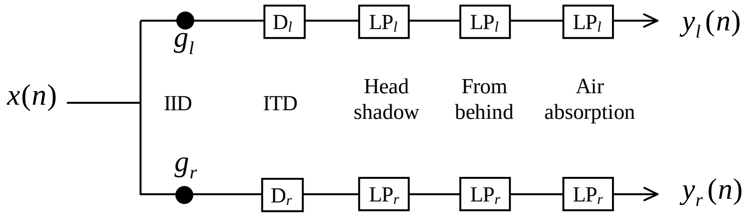


**Figure 3.11** Schema to model ITD, IID and some spectral cues.

## 3.3. HRTF MODELLING

Now that we have the means to build some simple DSP realisations, we will see that a DSP model that can handle an HRTF filter is very easily assembled. We will also see how the HRTF filters can be stored and what problems there might be with interpolation between them. Finally a discussion on the computational power needed to use HRTFs and how the needs might be reduced is given.

*Realising HRTF Cues with Digital Filters*

As discussed earlier in section 2.4, an HRTF filter is really just the impulse response from the pinnae, head, shoulders and torso, and since the impulse response can be realised with a FIR filter (try having $x(n)$ in Figure 3.7 to be the Dirac-pulse, and the output will become $y(n) = \{ g_1, g_2 \}$, i.e. the gain coefficients), there is no problem with implementing this. The two-channel FIR filter in Figure 3.12 below, represents the HRTF filter pair.

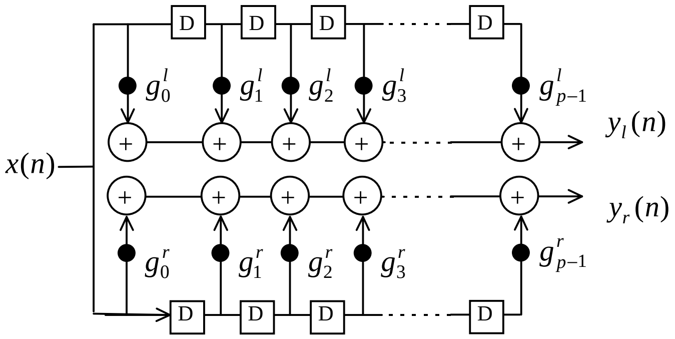


**Figure 3.12** Convolution using two separate FIR filters for binaural output. Two impulse responses, one each for left and right ear HRTFs, are applied to the single input, resulting in a two-channel output.

The schema above is actually a two-channel realisation of the convolution formula on page 20, with each channel's $g_n$s corresponding to the values of that channel's $h(n)$, which in turn is the HRTF filter pair. This form of FIR filter with delays and gains for convolution, is also known as a tap-delay filter, where the "tap" refers to each of the multiple, summed delay inputs that are placed at the output after scaling. One can also refer to an N-tap FIR filter, where N is the number of taps.

*Multiple Filters and Interpolation of HRTF Filters*

The collected HRTF filters are usually 512 samples long and sampled at 50 kHz. This corresponds to a 10.24 ms long filter, which is quite enough to hold both the temporal response of the pinna and the interaural delays (recall the maximum ITD being about 0.65 ms).

The discussion above has been about a single HRTF filter pair corresponding to the impulse response of one specific direction, but to simulate an arbitrary direction, one must naturally have an HRTF filter pair for all directions. However, it is not feasible to measure continuously many filter pairs since they would be infinitely many, but normally there are measured at an angular displacement of 15–20 degrees in the horizontal plane and 10–20 degrees of elevation displacement (see section 2.4 for the technicalities of measuring HRTFs). This procedure yields about 350 filter pairs to be stored in the computer memory.

When simulating a position that does not perfectly match with a specific filter pair, the HRTF for the desired position must be interpolated. The procedure is bound to the assumption that an in-between HRTF filter pair would have in-between spectral features. Some investigation (Begault, 1994) shows that this may be the case.

When interpolating an HRTF pair, the usual procedure is to take the filter pairs for the four measured positions closest to the desired one, and then produce a new filter pair by averaging the weighted values of the chosen filters. One problem is that the averaging is taking place in the time-domain instead of, as it should be, in the frequency domain. The resulting spectral response is not in between the spectral curves of the interpolated filters at all.

The problem also becomes apparent for the temporal features of the filters. For instance, consider $h_1(n) = \{5,1,1\}$ and $h_2(n) = \{1,1,5\}$, where the value five corresponds to the peaks of the two measured impulse responses, due to the interaural time difference. When interpolating these two filters, it is desired that the resulting filter becomes $h(n) = \{1,5,1\}$ to agree with the temporal interpolation, but instead the resulting filter turns out to be $h(n) = (h_1(n) + h_2(n))/2 = \{3,1,3\}$, which does not correspond at all to the time delay.

One solution to the problem of temporal interpolation is to separate the ITD from the impulse response pairs. Since the delay manifests itself as silence before the actual impulse response, it is quite easy to find the ITD and store it as a separate delay value realised in the DSP with a delay block, before the spectral shaping is taking place (see Figure 3.13). Lost in this process is the frequency dependent time-delay due to different wavelengths being diffracted differently around the head, but the artificially inserted time delay can be an averaged ITD value over the range of frequencies affected by the head shadow.

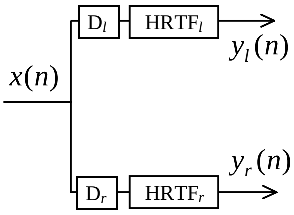


**Figure 3.13** A DSP system using separate delay and magnitude processing.

The problem with the erroneous interpolation of the spectral magnitudes has a slightly pragmatic solution: Ignore the problem! A study by Wenzel and Foster (1993) shows that the interpolation errors get drowned by the overall localisation errors made by the subjects.

Another solution to the problem might be to transform the filter pairs to be interpolated to the frequency domain using the *FFT* before interpolating them, and then transform the resulting filter pair back to the time domain.

***FFT:*** *Fast Fourier Transform. An algorithm to perform a discrete Fourier transform in one order of magnitude operations less than the brute force transform algorithm. See Proakis and Manolakis, 1992.*

### *Hardware and Reduction of HRTF Filters*

For hardware implementation purposes, it is often beneficial to create versions of the HRTFs with shorter impulse response lengths. The normal filter lengths when measuring HRTFs usually get 512–1024 coefficients long. A *Motorola 56001* running at 27 MHz can process a maximum of 153 filter coefficients in the impulse response when convolving a stereo signal at 44.1 kHz (Burgess, 1992).

***Motorola 56001:*** *One of the most commonly used DSPs today. Found in hardware ranging from SGI and NeXT stations to all kinds of commercial audio products.*

There are some methods to decrease the length of these filters. The first is to reduce the sample rate of the filters. For instance, a filter sampled at 50 kHz can accommodate frequencies up to 25 kHz (due to the Nyquist sampling theorem). But in the HRTF measurements, there is usually just a lot of noise above 15 kHz (Begault, 1994), yielding that the filter can be *downsampled* to 30 kHz, thus shortened by 40%. Another way to shorten the HRTF filters is to window them. This means that a narrower selection of the impulse response is picked out to be the HRTF filter. It can be done quite effectively and Begault (1994) shows that a 512 samples long impulse response can be shortened to about 64 samples with a close to preserved transfer function.

***Downsample:*** *Convert an amount of samples to a smaller amount of samples by interpolation, i.e. an array with 50 samples meant to be played in 50 kHz is interpolated to an array with 30 values meant to be played in 30kHz, yielding the same output.*

There are also entire hardware systems providing spatialisation of sounds. These systems are more usually found in research centres and movie post-production studios. One of them is the Convolvotron from Crystal River Engineering. The Convolvotron is a PC-based system where the DSP hardware consists of two or more computer boards put in the expansion slots of the PC. It has a capacity of convolving (spatialising) four to eight sound sources with both their direct sounds and their first order reflections, depending on how many of the computer boards you plug into the PC. For more information about the Convolvotron see CRE (1996), Begault (1994) and Kalawsky (1993).

## 3.4. PRESENTATION OVER LOUDSPEAKERS

Until now, the output of the spatialised sound has been meant to be presented over headphones. A further step is to present the 3D sound over loudspeakers. The problem is that it is difficult to control the spatial sound imagery when presented over loudspeakers in such a way that the imagery can be transported to several listeners in a predictable manner.

### *Troubles with Presentation Over Loudspeakers*

There are three main reasons why the presentation of 3D sound over loudspeakers is problematic, compared to headphone listening.

- The room where the listening is taking place superimposes its environmental characteristics over the signal from the loudspeakers. Especially the early reflections distort the direct signal, although these reflections can be damped in an acoustically treated room or in an anechoic chamber.

- The positions of the listener and the speakers in the room cannot be predicted. Most often there is not an audiophile listening actively, facing towards the loudspeakers, but there are multiple persons viewing a television monitor listening to the speakers as background sound.
- Unlike listening with headphones, the sound reaching each ear is a mix of the signals from two loudspeakers, as in Figure 3.14. Since 3D sound depends upon being able to control the spectral filtering of the signals occurring at each ear, it is quite unfortunate that both loudspeakers are heard by both ears. This cross-talk (Figure 3.14) interferes with the spectral balance and interaural differences significantly.

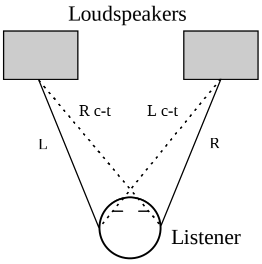


**Figure 3.14** Cross-talk.

*Cross-talk Cancellation*

The first problem could as mentioned be solved by acoustical treatment of the listening room. The second point is hardly possible to do anything about, but is unfortunately the basis of the solution of the third point, i.e. to solve the cross-talk problem one has to know the position of the listener.

The solution to cross-talk, known as cross-talk cancellation, can however be used when there is a listener sitting relatively still at one position, like when sitting in front of one’s computer. The cross-talk cancellation is based on inverting the cross-talked signal and play it at a lower volume in the other channel, thus cancelling the sound by interference. A small delay corresponding to the difference in distance must also be added. In effect, in the left speaker there is the left channel plus an attenuated, delayed, and inverted version of the right channel being played and vice versa.

There are however some problems with the cross-talk cancellation. First, the cross-talked channel is being filtered by the HRTF, meaning that the compensation signal has to be “inversely” HRTF-filtered to compensate for that as well. The trouble now is that we do not know the angle of incidence of the loudspeakers, and the listener might also move away.

The other problem is that the listener cannot move his head sideways for more than about 10 centimetres before a 2 kHz sound wave will be back in phase again, thereby amplifying the cross-talked signal. The moral is that normal head movement causes the cancellation to stop working and thus only low frequency sounds will be effectively cross-talk cancelled.

## 3.5. FOUR CHANNEL PRESENTATION

Since spatialising with HRTF filters is so computationally demanding, most of the research has been done in the last ten to fifteen years. One of the earlier proposals to spatial sound was presented by John Chowning in 1971. In his paper, he presented techniques to synthesise the position of a sound source, and in particular a moving sound source, using four-channel reproduction.

*The Spatialisation Model*

The presentation of the sound is intended to be over four speakers placed in the corners of a square surrounding the listener. A sound straight ahead of the listener would then be played at equal level in the two frontal speakers and a sound from the back in the two rear speakers. As the sound moves to the listener’s side, for instance to his left, the level falls at the right frontal speaker. When the sound passes 45 degrees to the left, the direct sound level at the right frontal speaker becomes zero and the level of the left rear speaker starts to increase.

The distance cue is simulated by manipulating both the intensity of the direct sound and the level of reverberant sound being distributed to all speakers (see R/D-ratio in section 2.5). The reverb consists of two parts to enhance room sensation. First the “global” reverberation that is evenly distributed to all speakers and is proportional to the distance, and second the “local” reverb, that is played in the speakers where the direct signal is played. This can to some extent be a fair approximation to the real acoustical situation Chowning states, because as the distance of the sound source increases, the relative distance to the reflecting surfaces decreases, thus giving the reverberation some directional emphasis.

The velocity cue is proposed to be achieved by simulating Doppler shifts. The radial velocity is calculated and the frequency of the sound is being altered correspondingly.

## 3.6. Modelling of the Environmental Context

In section 2.5, it was described how the reverberation in a room was perceived and how it could be theoretically described. The reverberation was divided into early echoes and late reverberation. Both give us cues to distance and the acoustical characteristics of the room. In this section will be discussed how reverberation can be modelled.

### *Early Echoes Using the Image Model*

The aim of the image model and the ray tracing algorithm is to find out from what directions the early echoes are coming. In the image model, this is done by making “virtual” mirrored copies of the room with the sound sources positioned at their new “virtual” locations representing the reflected sound path (see Figure 3.15).

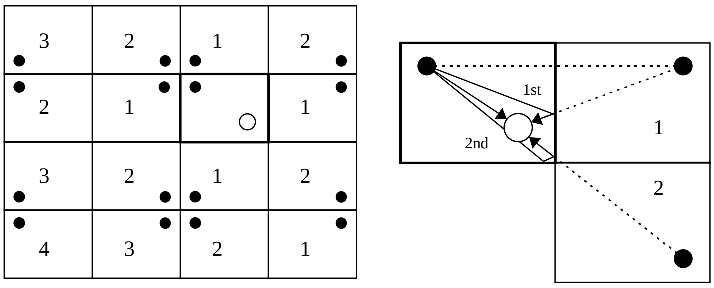


**Figure 3.15** Use of the image model to calculate early reflection patterns.

The outlined circle in the figure is the listener and the filled circles are the actual and virtual sound source. The sound source in the actual room (bold lines) is mirrored outward to create first order reflections, which are labelled ‘1’. The mirrors are mirrored to create second order reflections, labelled ‘2’, etc. In the magnification to the right, the direct sound path can be seen along with one of the first and one of the second order reflections.

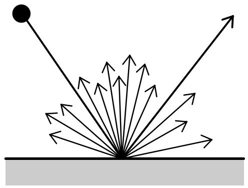
The positions of the new (virtual) sources can be added to the spatialisation algorithm (perhaps through HRTF filters) making the early reflections emanate from the right directions. Borish (1984) shows a way to extend the image model to an arbitrary polyhedron, making it possible to be used in any volume being modelled.

*Early Echoes Using Ray Tracing*

In the ray tracing method, “sound rays” are emitted from the sound source in all directions, possibly according to the sound source’s pattern of emission. As a ray encounters a surface, it can be appropriately filtered with respect to the angle of incidence and to the wall’s reflection qualities.

Unlike the image model, the effects of *diffusion* at surfaces can be more accurately modelled with the ray tracing method since one can simply initiate some more rays at the point of incidence. The ray tracing method is also much more computationally intensive than the image model since a large number of rays are needed for a somewhat accurate result.

**Diffusion:**



The need for a large number of rays is the ray tracing method’s large drawback. No matter how many rays that are emitted from the sound source, there is always a possibility that one ray is passing very close to the listener and not considered as a hit. Then another adjoining ray perhaps also misses the listener and that particular path of reflections is lost, since no ray found the listener. In the image model, all the reflections are by definition found.

*Late Reverberation*

Late reverberation is, as mentioned, used to simulate the reflections of the sound after about 80 ms. At this point the number of reflections has grown so big that a single reflection cannot be perceived, and the reflections can be modelled as a decaying “diffuse” sound.

When previously presenting the IIR filter, it was stated that it could be used as reverberation unit. That was only partly true. It clearly can be utilised, but the IIR filter in its simple form shown earlier imposes undesired spectral alteration. In Figure 3.16 the transfer function is shown for an IIR filter with the delay greater than one sample. As can be seen the transfer function is far from the desired flat one.

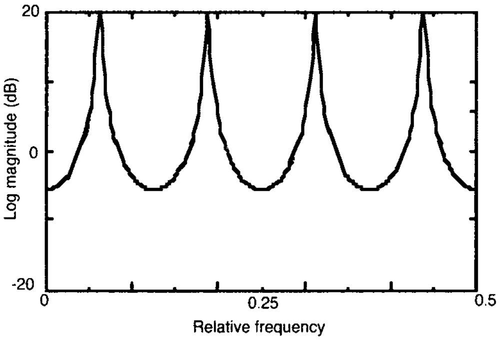


**Figure 3.16** IIR filter resulting in a “comb filter”. The number of peaks corresponds to the length of the delay

To avoid these spectral properties of the comb filter, giving a very characteristic unpleasant metallic sound, Schroeder (1962) presented the all-pass filter. The transfer function of the all-pass filter is completely flat, thus the name all-pass. Despite of its peculiar name (why would one want a filter that does not do any filtering?), the all-pass filter is very useful when a signal is to be delayed and fed back to the filters input again, consequently being perfect for use in reverberators. Besides avoiding the metallic sound, the flat transfer function gives the opportunity for colouring (filtering) the sound separately from the delaying. Figure 3.17 shows the schema of the all-pass filter along with the comb filter (IIR filter) for comparison.

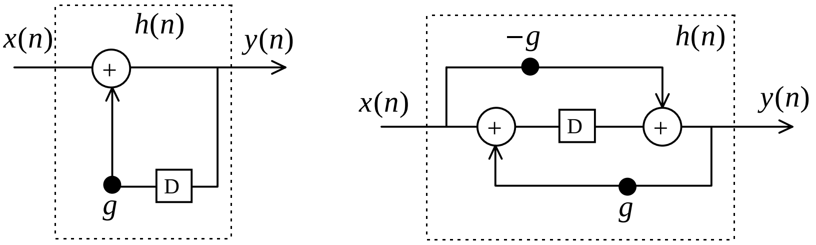


**Figure 3.17** To the right is the all-pass filter with the function $y(n) = -g \cdot x(n) + x(n-1) + g \cdot y(n-1)$ .

In Moorer (1977) the design issues of a reverberator are discussed. It is proposed that a reverberator could consist of multiple all-pass filters in cascade (connected in series). The delay values in each of the all-pass blocks are also important. Because the resulting reverberated signal should be dense and have no distinguishable repetitions, the delay lengths are chosen to be relatively prime. This is stated to produce the most dense reverberation.

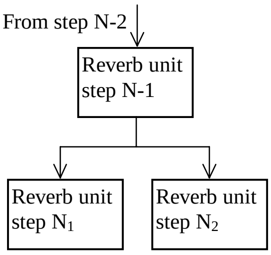


**Figure 3.18** Multi-channel reverberation output

When the output is presented over multiple channels (headphones for instance), the reverberant signal output must also be multiple. This is because if the same reverberation signal were presented in both the left and the right channel, the room feeling would be lost since the reverberation in both channels would be fully correlated, thus not corresponding to the real situation. Instead Moorer proposes the schema shown in Figure 3.18, where the reverberation signal is uncorrelated in the last step where the two all-pass filters in parallel have slightly different delays. This saves computational power, since all reverberation steps before the last step are shared between the two output channels. For four output channels, each of the two output channels are shared by yet another step, yielding four different reverberation signals.

The models shown above are quite old fashioned and today's reverberators are much more complex and intricate. However, they are still mainly based on the theory behind the reverberators shown above.

## 3.7. Integration of 3D audio

***Working environment:*** *The system that the audio system is going support with 3D sound. It might be a virtual reality system or perhaps an air traffic control system. The working environment is here also called the main system.*

When a satisfactory model has been developed, the next step is to integrate the 3D audio system in its *working environment*. With integration it is not meant integration *within* the working environment, but integration *with* the working environment. A 3D sound system might range from being an external hardware sub-system to being fully integrated within its working environment. In any case, for a 3D sound integration to be successful, there are some things that must be kept in mind.

### Demands on the Working Environment

The integration of 3D audio can be really easy. If the working environment is already a 3D system, then the integration is naturally very straightforward. But if it is not realised that 3D sound sources exist in a 3D environment and furthermore might coexist with visual and temporal events, and consequently are forced into a system not supporting these qualities, then there is a possibility of trouble.

The working environment must be able to continually feed the sound system with positional data of the sound sources. Furthermore it is the responsibility of the working environment to provide the timing of the aural events since the audio system is more or less slaving under the working environment. The 3D audio must be considered during the whole development process of the main system and cannot be added at the last stage, like a soundtrack at the end of a movie.

As discussed earlier, the audio sub-system can be an external hardware unit or it can be fully built-in inside the main system. An example of an external hardware unit is the Convolvotron, described in section 3.3, which communicates with the main system via a serial (RS232) line. As mentioned, the drawback of this is that the main system cannot be brought to another machine without bringing the spatialisation hardware there as well. Another approach is to integrate the audio sub-system within the main system, as we will see in the case study in the next chapter.

## 3.8. Applications of 3D audio

As a conclusion to the theoretical backgrounds to 3D audio, it could be interesting to learn about some applications for it. The examples below are from Begault (1994) and are only an extract of all the different applications that are being developed, not to mention those not yet invented.

### 3D Audio for Virtual Reality Systems

This is probably what most people think of first when they hear 3D audio or virtual audio; audio for virtual reality systems. In this case the audio is (should be) an obvious part of the VR system and produces in conjunction with a graphics display the illusion of being *immersed* in a virtual environment.

***Immersed:*** *Absorbed, involved.*

An efficient presentation unit is some type of gear with screens that cover the whole field of vision and headphones to play the 3D sound on. The gear’s movements are usually being followed and thus the VR system can respond accordingly. It is of highest importance that the lag (the time from user action to system response) is no more than some ten milliseconds (20–50 Hz), otherwise the user gets frustrated that his movements and actions are not responded upon, and true immersion can never be reached.

### Air Traffic Control Systems

For both air traffic controllers and pilots, an *auditory display* can be of big use. This is especially because their visual perception system is

***Auditory display:*** *A display for information, where the information is conveyed by auditory means, for instance over headphones.*

already overloaded and for them to be able to attain more information using an auditory display is a possibility.

In a cockpit, where the pilots are already using headphones, a 3D audio system can be incorporated along with other information systems. An example could be a collision warning system, that alerts the pilots with a spatialised voice, and they both know which one of them that must look outside the window and in what direction.

### *Audio User Interfaces for the Blind*

There is an example in Begault (1994) describing a user interface for the blind where common interactions like mouse movement, window pop-up and sizing, and button press have been mapped to auditory icons in 3D. The mouse movements are in this example mapped to steps walking, window sizing to elastic band, and button press to electric switches.

### *Games*

Last but not least, there is the 3D audio in computer games. It is expected that this industry, which has become leading in developing fast algorithms for 3D graphics will also be inventing new and computationally cheap algorithms for 3D audio. As the computer entertainment business grows this is also where the financial funding will be found.

# CHAPTER FOUR
# IMPLEMENTATION OF 3D AUDIO IN DIVE

In this chapter, I will present the implementation of a 3D audio model in the virtual reality system DIVE. The chapter starts with an introduction to the working environment DIVE. Next is a section where I discuss which of the HRTF and algorithmic models to use, considering the conditions and possibilities of DIVE. The following section is a thorough presentation of the details in the implementation I chose. A section on the different types of sounds in a virtual reality system is given before the chapter is concluded with a section on aspects of real-time transfer of sound on a network.

## 4.1. INTRODUCTION TO DIVE

This section is a short introduction to the working environment DIVE, which is being developed at SICS (the Swedish Institute of Computer Science). The SICS Distributed Interactive Virtual Environment (DIVE) system is a multi-user virtual reality system that is run within a heterogeneous multiprocessing environment (i.e. a network of different UNIX workstations). Independent DIVE applications run on nodes within a local- or wide-area network, and update a shared (node-wise replicated) object database. DIVE is mainly developed to be a virtual reality framework for building applications and user interfaces in 3D-space.

In the rest of this chapter I will explain the details of the DIVE system as they become crucial to understanding the audio implementation. For a further general presentation of DIVE, the reader is referred to the papers written by Hagsand (1996), Carlsson and Hagsand (1993) and the electronic information in DIVE (1996).

## 4.2. CHOOSING THE ALGORITHMIC OR THE HRTF MODEL

As stated earlier, the advantages of the HRTF model are that the spectral cues provided by the pinnae help in externalising the sound, can disambiguate front from rear, and also impart elevation information. The disadvantages are that the spatialisation using HRTF filters cannot be done in software, but one has to use some kind of additional hardware (if not running the spatialisation on a super-computer). The pros and cons of the algorithmic model are the opposite; spatialisation can be done in software, thus integrated with the working environment, but the quality of the spatialisation model lacks in comparison with the HRTF model.

*What Aspects Were Important to the DIVE System*

After several discussions with Olof Hagsand, Emmanuel Frécon and Lennart Fahlén, it was concluded that what was wanted in the DIVE system was audio that could work on all platforms, since DIVE is developed for a heterogeneous environment. A requirement was also that the audio should be integrated in the DIVE system since DIVE is under constant development and this tends to make external applications incompatible over time. Connections with external audio conferencing applications like VAT and NeVoT (1996) had been done, but had a history of problems. Furthermore, the integration within DIVE should be chosen to enable as many users as possible to take advantage of the 3D audio, which excluded the hardware option.

*What Choices Were There?*

The choices from the start was to implement the audio system according to the algorithmic model or to the HRTF model.

To run the HRTF model, there was a hardware system from Crystal River Engineering supporting the whole spatialisation process; the Acoustetron II (CRE, 1996). The system consists of a standalone PC with the DSP boards inserted in the expansion slots. Positional information is sent over a serial (RS232) line from the computer running the scenario control to the audio computer. Since the software that would control the audio computer could be integrated within DIVE, this solution was only inadequate due to the last requirement mentioned above, that as many users as possible should be able to enjoy the 3D audio.

There were some implementation problems as well. I tried to contact Crystal River Engineering by email to their information mailbox, but in spite of repeated mails I received no answer. Questions about the cost of the system and if a set of HRTFs would be included with the system remain unanswered. Therefore no conclusion could be drawn as to whether it would be feasible to acquire such a system or not. If no HRTFs were to be included, then the recording (there is however a set to be downloaded from MIT, 1996), data reduction and ITD processing would be almost a thesis by itself. Also, the cost per user of the Acoustetron II would be quite high since only one user can use the system at any one time. This means that some additional audio system had to be implemented as well to provide audio to users not having the audio hardware.

The fact that the algorithmic model fulfilled all the requirements on the audio system and that it would be possible to build the audio system from the ground made me choose to implement the algorithmic model.

## 4.3. USING THE ALGORITHMIC MODEL

The implementation of the algorithmic model has been a very interesting task, especially since theory and practice did not always go hand in hand, and empirical values in different parameters had to be chosen to make the sound "feel" good. I will in this section present the spatialisation algorithm and the cues implemented in it, but I will start with an introduction to how the audio system is organised and what the connections are to the rest of the DIVE system.

*The Architecture of the 3D Audio System*

From a user's perspective, sound can be either played or recorded. The first direction is sound being recorded via the computer's microphone or other sound input channels and sent in 1,000 bytes big blocks on the network. Included in each sound packet is, among other things, an object id to identify the sender of the sound. The network transport protocol used is UDP/IP (Stallings, 1994) where the additional multicast service is utilised. This means that all other DIVE-users will get the sound packets sent.

The object id uniquely identifies an object in a world. From the object id, the object's spatial position, among other things, can be derived. Another source of sounds to be sent out on the network are audio files. Before loaded and sent on the network, the user can mark an object to be the sound source, thus providing the required object id in the sound packets. In section 4.5 is a brief discussion on the transfer of audio streams over a network given.

The other direction is sound received from the network. The sound packet is received, processed, and then played. The processing between when the packet is received and played includes the following stages:

- **Decoding:** The arriving sound packet can be stored in a number of ways. It might be coded in some linear format using 8, 16, 24 or 32 bit integers or floating point numbers. Another possibility is the packet being encoded using some adaptive algorithm (one of the standards G721–G723) or perhaps μ-law or A-law coding (G711) which stores the sound logarithmically yielding a dynamic range of 12 bits inside 8 bits. The μ-law and A-law coding are used in telephone systems, and are the most usual encoding schemes on UNIX systems. In the decoding section the *byte order* (that can be different on different platforms) is also taken care of. To learn more about encoding schemes please see Kleijn and Paliwal (1995), and for byte orders, see Stevens (1991).
- **Ordering:** When sending packets over a packet-switched network, the packets may arrive unordered due to being sent via different paths in the network. To keep track of the order of the sound packets, a sequence number is included in each sound packet. As the packet arrives, the sound block contained in the packet can be sorted into the receive buffer before being played. A thorough discussion of problems with sending real-time data over a packet-switched network is given below in section 4.5.
- **Sound source processing:** Different types of sound sources with different emission patterns can be defined. Before the sound block is spatialised, it can be processed according to the sound source type, the position of the source and the position of the listener. Examples of sound source types are omnidirectional and megaphone type sources.

***Byte order:** When storing numbers in data types larger than a byte, the bytes can be ordered differently. For a 16-bit value, which consists of two bytes, the bytes can come in the order most significant or least significant first. For a 32-bit value it is even more complicated. For transferring data over a network, there is a standard ordering scheme and when receiving data from the network, the byte order is adjusted for the actual platform.*

- **Spatialisation:** The last step before the sound block is output to the user is the spatialisation. Before each block is spatialised, the position of the listener and the source is updated. Distance, interaural coefficients and filtering coefficients are calculated and eventually applied in the spatialisation. The resulting packet is added to the audio device queue.
- **Audio device:** The audio device is the unit providing the timing in the system. A so called double buffering scheme is used. It means that the device always has two audio blocks in its queue, one block being played and one block waiting to be played. When the first block is finished, the device continues by playing the next block at the same time as a message is sent to the spatialising module that a new sound block is needed. The spatialiser starts to produce the new sound block, and it is of course crucial that the time it takes to produce a new block is less than the time it takes to finish the block being played.

### *Schema of the Spatialisation and Its Components*

Below in Figure 4.1 the spatialisation schema is presented. Its components are presented in detail in the sections below. The schema is in the logical form, meaning that in the implementation some optimisations have been done. For instance, the distance factors are multiplied before being applied to every sample in the block. In other words, instead of multiplying one sample with three different direct sound factors, each sample in the sound block is multiplied with just one composite factor, thus removing two multiplication operations.

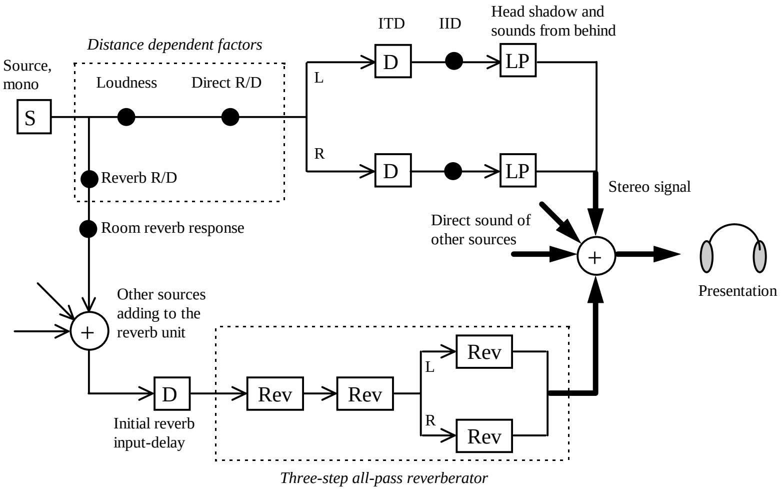

ITD
IID
Head shadow and sounds from behind
Distance dependent factors
Source, mono
Loudness
Direct R/D
S
L
R
D
LP
D
LP
Reverb R/D
Stereo signal
Room reverb response
Direct sound of other sources
+
Presentation
Other sources adding to the reverb unit
+
Rev
L
D
Rev
Rev
R
Rev
Initial reverb input-delay
Three-step all-pass reverberator

**Figure 4.1** The spatialisation module. The direct signal follows the upper path.

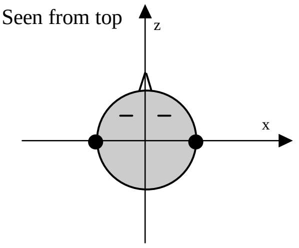


**Figure 4.2** Coordinate system used in geometric calculations.

### *Interaural Cues*

The interaural cues previously explained in section 2.2 are the interaural time difference, ITD, and the interaural intensity difference, IID. In the calculations below the listener is considered to be positioned at the origin facing down the positive z-axis, see Figure 4.2. The model implemented to calculate the distance difference between the right and the left ear is the simplest one, shown in section 2.2, Figure 2.3, where the path around the head is excluded and only the geometrical distance is measured. As will be seen below, only the IID calculation uses the sine function. The ITD uses the Pythagorean relation instead due to speed.

When calculating the IID, it is not the intensity difference that is calculated, but the gains which the sound samples are to be multiplied with. First, one of the channels' gains is calculated. The other channel's gain is then one minus the first gain.

In the implementation, the left ear's gain is calculated according to the formula $0.5 \cdot C \cdot \sin(\theta)$, where $C$ is the angular response factor and $\theta$ is the horizontal plane angle to the sound source. The value of $C$ is set to 0.25, which yields a range for the IID gains from 0.25 to 0.75. This results in the gains for a sound straight ahead of the listener being $\mathrm{IID}_L=\mathrm{IID}_R=0.5$, where $\mathrm{IID}_L$ and $\mathrm{IID}_R$ are the gains for the left and right ear respectively. A sound to the listener's absolute right will yield the gains $\mathrm{IID}_L=0.25$ and $\mathrm{IID}_R=1–\mathrm{IID}_L=0.75$. The gains correspond to a level difference ranging from 0 dB for a sound straight ahead, to 9.5 dB for a sound source to the listener's far side.

The calculation of the interaural time difference, ITD, is done in another fashion. It uses the Pythagorean relation and the formula looks as follows:

$$\Delta t = \frac{\Delta d}{c} = \frac{\sqrt{(d_x - R)^2 + d_z^2} - \sqrt{(d_x + R)^2 + d_z^2}}{c}$$

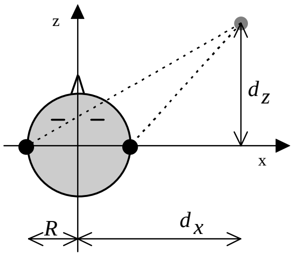


**Figure 4.3** Calculation of the distance difference.

In the formula the difference between the two hypotenuses is calculated, yielding the distance difference to the two ears (see Figure 4.3). The distance difference $\Delta d$ is divided by the speed of sound $c$, and the quotient $\Delta t$ is the desired time difference.

The head radius R is set to 0.12 m in this implementation which is quite large. A normal head is about 0.18 m in diameter. The value is empirically found to be realistic in hearing tests. The rather large value is maybe due to the model excluding the head (remember the erroneous value in section 2.2), and thus the size of the head is somewhat larger.

*Direct and Reverberant Sound Level Coefficients*

As can be seen in Figure 4.1, the incoming sound signal is divided into two paths. The upper path represents the direct sound and the lower the reverberant signal. The three intensity adjustments in the box labelled "Distance dependent factors" all depend on a distance factor. The distance factor is calculated according to the inverse square law, mentioned in section 2.5. However the formula $DF=1/d^2$ ($DF$ being the distance factor and $d$ the distance in meters) could not be used because it resulted in the signal's level being attenuated too fast, i.e. the sounds' volume became too low at too short a distance. Consequently, the inverse square formula had to be rewritten in some form and empirical studies resulted in the formula $DF=1/(d \cdot 0.05+0.9)^2$. The distance calculations pointed out in the schema in Figure 4.1 are all written on the form $C \cdot DF$, where $DF$ is the distance factor and $C$ the constant denoting how much effect the distance factor is to have in that particular multiplication step.

As can be seen, the direct signal is first attenuated due to distance (labelled "Loudness"). The signal is being multiplied with the product $C \cdot DF$, where C in this case has the value 1.3. It can be mentioned that all constants given on the signal coefficients below are empirically found and are no universal values, but more what according to the author sounded "right".

The "Direct R/D" and the "Reverb R/D" are the values constituting the quotient reverb-to-direct sound ratio also mentioned in section 2.5. Both of the constants are multiplied with the distance factor *DF* before multiplied with the sound signal. The direct R/D-ratio constant's value is set to 1.3, and the reverb R/D-ratio value is set to 0.5.

Additionally a room reverb response factor is multiplied with reverberant signal before added to the reverb unit. The room reverb response factor is used to set the reverberance in the space surrounding the sound source and the listener. A greater value gives a more reverberant volume and lower gives less reverberation. An anechoic chamber would have the room reverb response factor set to zero, thus eliminating the signal added to the reverb unit. A quite good value of the room reverb response factor is 0.3, which gives a garage kind of reverb, but on the other hand helps very much to externalise the sound. The garage reverb is unfortunately not very appropriate for sound in a free-field, which is sometimes the case in virtual environments. This makes it quite a balance act to set the room reverb response to acquire externalisation, since no HRTF-filters are implemented that can help us with that, but not having the reverb sounds too unnatural.

As mentioned before, the sound intensity factors are first multiplied due to optimisation before multiplied with the signal. Below are the full formulas given for the two intensity factor on the direct and reverberant path. $I_L$ and $I_R$ are the factors multiplied with the left and right channel respectively and $I_{Rvb}$ is the factor multiplied with the reverberant signal before added to the reverb unit.

$$I_L = Loudness \cdot DF \cdot DirectRD \cdot DF \cdot IID_L = 1.3 \cdot 1.3 \cdot DF^2 \cdot IID_L$$
$$I_R = Loudness \cdot DF \cdot DirectRD \cdot DF \cdot IID_R = 1.3 \cdot 1.3 \cdot DF^2 \cdot IID_R$$
$$I_{Rvb} = ReverbRD \cdot DF \cdot RoomReverbResponse = 0.5 \cdot 0.3 \cdot DF$$

Below in Figure 4.4 is a graph that shows the levels of $I$ and $I_{Rvb}$ respectively. The values used are the same as above and the sound source is straight ahead of the listener yielding $I=I_L=I_R$.

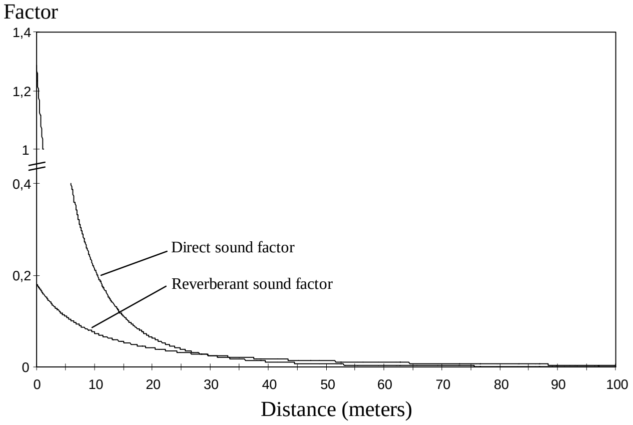


**Figure 4.4** Direct and reverberant sound intensity factors.

It can be seen that the direct signal level goes below the reverberant level at about 30 metres. This is due to the squared distance factor

(*DF*) in the direct sound formulas. Actually it gives the direct and reverberant sound a quite real relationship, since the direct sound decays faster than the reverberant and on longer distances, the intensity of the reverberant sound is, as mentioned earlier, higher than the direct sound intensity.

### *Low-Pass Filtering for the Head Shadow and Sounds from Behind*

In section 2.4 we saw, and maybe experienced, how noticeably the high frequencies were attenuated due to the head shadow and by the pinnae for sounds emanating from behind. This is a very stable cue that is also easily implemented.

The simple first order FIR filter shown in Figure 3.7 on page 21 is a cheap filter that can be used. Its only drawback is that the low-pass filtering produced by the system is very unrefined, but as the transfer function in Figure 3.9 shows, the system is quite adequate for our needs. As can be seen in Figure 3.9 the gain coefficient $g_1$ is set to 1, thus it is omitted in the implementation. The gain coefficient $g_2$ is the only value controlling the filter function and, as Figure 3.9 shows, results in a useful transfer function for values in the range 0 to 0.9. As the schema in Figure 4.1 shows, two different gain coefficients $g_2$ are needed, one for the left ear and for the right ear. Below I will call them $LP_L$ and $LP_R$. Notice that for an $LP_L$ or $LP_R$ having a value of 0, the direct sound path is unaffected by the low-pass filter.

The calculation of $LP_L$ and $LP_R$ is done similarly to the previous gain coefficients. The formulas can be written $LP_L = LP_B + LP_{HSL}$ and $LP_R = LP_B + LP_{HSR}$, where $LP_B$ is the low-pass coefficient for sounds from behind, and $LP_{HSL}$ and $LP_{HSR}$ are the low-pass coefficients for the head shadow on the left and right ear respectively. The $LP_B$ is operative for horizontal angles greater than 90° or less than –90° and has its maximum weight at ±180°. Thus the first step in the calculation of the low-pass filter coefficients is:

$$LP_B = \begin{cases} C_B \cdot -\cos\theta, & \text{abs}\theta > 90^\circ \\ 0, & \text{abs}\theta \le 90^\circ \end{cases}$$

The coefficient $C_B$ states to what extent the angular displacement will have effect. In the implementation in DIVE, its value is 0.5. This means that for an angle $\theta = 120^\circ$, the gain coefficient $LP_B$ will be $LP_B = 0.5 \cdot -\cos 120^\circ = 0.25$.

The effect of head shadowing is greatest when the sound source is at the considered ear's far opposite side. The formulas for the head shadow are the following:

$$LP_{HSL} = \begin{cases} C_{HS} \cdot -\sin\theta, & -180^\circ < \theta < 0^\circ \quad \text{(sounds to the right)} \\ 0, & 0^\circ \le \theta \le 180^\circ \quad \text{(sounds to the left)} \end{cases}$$

$$LP_{HSR} = \begin{cases} C_{HS} \cdot \sin\theta, & 0^\circ < \theta < 180^\circ \quad \text{(sounds to the left)} \\ 0, & -180^\circ \le \theta \le 0^\circ \quad \text{(sounds to the right)} \end{cases}$$

The constant $C_{HS}$ has in the implementation the value 0.2.

The low-pass filtering does not help the inexperienced listener since the low-pass filter is quite unnatural, but when the listener has learned what it sounds like, i.e. has learned to use that cue, the low-pass filtering can help him judge whether the sound source is to his right or left, or front or back.

### *Reverberation Gains and Delay Lengths*

The reverberation unit consists, as seen in Figure 4.1, of three steps of all-pass filters. The only values that have to be supplied to an all-pass filter (see Figure 3.17) are the gain coefficient and the length of the delay unit. In the paper by Moorer (1977), he presented gains and delay unit values that produce a dense reverb. The values were slightly adjusted to fit the three-step reverberator with a last parallel step. The gains used are $g_1 = 0.7,\ g_2 = 0.665,\ g_{3l} = 0.63175,$ and $g_{3r} = 0.6175$, and the delay times used in the implementation are in milliseconds $D_1 = 49.5,\ D_2 = 40.0,\ D_{3l} = 31.8,$ and $D_{3r} = 30.0$ .

Before use, the delay times must be transformed to a corresponding delay buffer length in samples. As stated earlier, the delay buffer lengths gave the best result when they were relatively prime. Therefore the buffer lengths are replaced by the value returned by the C-function below, which returns the nearest higher prime to the given number. The transformation to prime sample numbers is done only when the sound system is initialised, therefore any optimisation like, for instance, a lookup table is unnecessary.

```
static int audio_get_next_prime(int number)
{
  if (number<=1) number=1;
  if (number<=2) return 2;
  if (!(number % 2)) number++;
  if (number==1) number=3;
  while (! (audio_isprime(number) ))
    number+=2;
  return number;
}

static int audio_isprime(int thenumber)
{
  int isitprime=1, loop;
  for(loop=3; (isitprime) && (loop<(sqrt(thenumber)+1));
      loop+=2)
  {
    isitprime = (thenumber % loop);
  }
  return(isitprime);
}
```

Finally an explanation of the delay unit labelled ‘Initial reverb input-delay’ in Figure 4.1. Its delay time corresponds to the time it takes for the first reflection to reach the listener, and thus gives the listener a sense of the room’s size. The implementation does not regard any objects or volumes in the virtual environment but is simply a constant value of 25 ms.

### *Summary of the Implemented Cues*

The cues implemented are almost all the ones listed in section 3.2. The only cues not implemented are those dealing with elevation, since they are more expensive and not so stable. Hence, this implementation is actually a 2D and not a 3D spatialiser.

Still the 2D spatialising gives the listener a quite good ability to locate sounds in his horizontal plane. This is satisfactory to the extent that DIVE is mostly used to model human concepts, where the most action actually is taking place in the horizontal plane.

In section 5.3 I discuss what future work can be done on the sound in DIVE. There is among other things mentioned that elevation cues are left to be implemented but that the best solution would be to use HRTF cues.

*A Small Word about Audio Drivers*

The audio implementation for the DIVE system has so far involved writing interfaces to the audio drivers for the Silicon Graphics workstations running IRIX 5 and Sun SparcStations running Solaris 2. Each of the operating systems has its own interface to the computer's audio driver and they are completely different.

On the SparcStation it is supposed that the programmer should interact with the audio driver using the stream concept. The Silicon Graphics workstations on the other hand, have a very elaborate software interface including a number of function calls. The interested reader can look in IRIS (1995) and SunOS (1994).

It is easy to draw a parallel to graphics libraries. For graphics there are some two or three standard software graphics library interfaces. Implementations of these are provided by either the manufacturer or by a third-party manufacturer. This makes it easy (at least in theory) to write programs that utilise graphics, since the interface for a specific graphics library looks the same for every different platform. Perhaps it is too much to ask for, but it would be nice to have a standard interface to all the different audio drivers, so that one would not have to reinvent the wheel over and over again.

## 4.4. SOUNDS IN A VIRTUAL ENVIRONMENT

This section discusses some types of sounds that can exist in a virtual environment. The types explained are conference, environmental and common world sounds.

*Conference Sound*

Conference sound is sound that is used to transfer speech signals between human peers in a virtual environment. Demands are imposed on the delay of the speech signal to be as short as possible. A delay of more than one half second is perceived as very disturbing. Compare this with a transatlantic telephone call where the short delay of two or three tenths of a second can be enough to make it hard to converse.

Since the sound is recorded at one physical point (one workstation) and is sent to another, it puts requirements on the transfer of the data to the other workstation. More about real-time data transfer over a packet-switched network is found in the next section.

*Environmental Sound*

Environmental sounds emanate from objects in the virtual world as response to being interacted with. An example would be a chair creaking when turned or sat on. These sounds are important to make the virtual environment more persuasive.

The sounds can be either previously recorded samples, just played when needed, or algorithmically produced when the sounding object is being acted upon. The algorithmically produced sounds have the advantage of being able to respond in real-time to reflect different types of actions on an object in an ongoing sound stream. An example is a tree responding to a wind of increasing strength. The sound from the tree as a function of the wind strength can hardly be done with a discrete number of differently sounding samples without clicks being produced when changing samples.

A further presentation of algorithmic models for production of interactive sounds can be found in Gaver (1993).

### *The Voice of God and Ambient Sound*

The last type of sounds are the common world sounds. These sounds are heard by every user in every position of a world. One subtype is the "Voice of God", acting as an intercom system for broadcasting messages to all users in a world where the voice is positioned at one meter right in front of every user.

Another subtype is ambient sound, like thunder and rain. Ambient sound does not have a precise spatial position but is more emanating from all directions. For this subtype, the direct sound level is almost put at zero level and a large amount of the sound is lead through the reverb path.

## 4.5. TRANSFER OF SOUNDS ON A NETWORK

In the present implementation, all sounds that occur in a world are transferred over the network from the user causing the sound to the other users who locally decide whether the sound is audible or not. The traffic load is kept down due to the use of multicast, but at the end of this section it is discussed how the load can be further reduced through the use of adaptive techniques.

The section starts with an explanation on how the network flow control is presently implemented and a includes a brief discussion on some adaptive *flow control* techniques before closing with the subject mentioned above.

***Flow control:*** *The regulation of the pace with which packets are being sent on the network.*

### *Flow Control*

The current implementation of flow control is a very simple scheme almost only applicable to conference sound. It is quite effective since each user with the conference sound turned on continuously sends sound data, and furthermore allocates a sound channel in the spatialiser of each listening user.

The flow control consists primarily of a silence detector which simply makes the recording unit stop sending sound packets if their content does not meet a certain condition. The condition that the data of a packet must fulfil to be sent is that a given percentage of the samples' levels must exceed a given threshold value. The percentage value's default is 10 percent and the threshold value is 4000 where the maximum is 32767.

Additionally, there is a value that states how many silent blocks of sound are sent after a nonsilent block has been sent. This is to prevent silent fractions of speech from being left out, resulting in a very cut up sound stream where many of the consonants would be missed, since they contain quite small amounts of sound energy.

The only trouble is that this silence detection scheme tends to miss utterings that start with a low-energy consonant, like an 'H'. For instance, if the sending is turned off due to silence and the user wants to start by saying the word 'Hello', then the silence detection will not let the ‘H’ phoneme through, because it contains a too small amount of energy, but the sent audio packets will contain the sound formed by the following ‘ello’.

As a last parameter, which also affects the environmental sounds, the receiver can adjust the number of packets being *buffered*. To understand why it is important to restrict the number of packets being buffered at the receiver, imagine the following example: A group of sound packets are delayed in the network. This results in an audible gap at the receiver side. When the sound packets finally arrive, they arrive in a burst with a total length of perhaps three seconds. These sound packets cannot be played any faster than any other sound packets and results in a buffer build-up that is three seconds long. When new sound packets arrive, without any network delay, they are delayed for three seconds due to the buffer holding the previously delayed packets. What has to be done is to restrict the amount of packets being buffered and if that amount is exceeded then excessive packets are dropped. This results however in an audible gap, but guarantees that the shortest possible delay is maintained.

***Buffering:** The storing of packets that cannot be used for the moment, like a sound packet that comes prior to a packet that has been recorded earlier but has been delayed in the network.*

Normally the amount of packets being buffered at the receiver is dependent on the type of sound data sent. Conference sound, for instance, should not be buffered by more than a quarter of a second, due to the difficulties a lag in speech imposes on communication, but sound effects on the other hand should be buffered infinitely long (or at least as long as the buffer memory allows). The adjustable buffer length is necessary, because when communicating over a wide-area network, packets tend to be delayed more than in a local-area network. Thus, when conferencing over a wide area network (like the Internet) it is often necessary to crank up the default maximum buffer lengths to allow for this delaying of packets.

### *Adaptive Techniques*

There are more advanced techniques of controlling the flow of packets on a packet-switched network. In these techniques, models that adapt to the variance in arrival time are used.

The adaptation to the variances in arrival time (i.e. random delays caused by the network) makes it possible to estimate the minimum length of the receive buffer that is needed to prevent audible gaps depending on packets that arrive too late to be played.

In Jacobsen (1994), the above techniques are described along with techniques to also estimate the trend in the estimation of the variance in arrival time. The estimation of the trend is very useful in situations where the sender and receiver processing rates are mismatching. An example described is that the sample rate can differ up to ±10%. This means that an 8 kHz sound sample might be produced by the sender at 9,000 samples per second while the receiver consumes samples only at a rate of 7,000 samples per second. This would lead to the receiver buffer being filled up, resulting in delay being maximised and a very bursty packet drop behaviour. The estimation in the trend can prohibit this situation by for instance downsample (section 3.3) the incoming sound packets or drop packets in a smooth fashion.

None of the techniques described above is implemented in DIVE. In a future extension of the flow control, these and perhaps other techniques should be considered for implementation.

*Conference Sound, Environmental Sound and the Network*

Conference sound per definition has to be sent across the network. Environmental sound, on the other hand, is known in advance and can therefore be stored locally at every peer participating in a virtual world.

One problem is the distribution of all these previously known environmental sounds. Should they be distributed at the joining of a world, or later as they are needed. The latter case is similar to conference sound in the sense it should be sent in real-time, but it is possible to store it at the receiver for later use.

As mentioned, in the present implementation all sounds (both conference and environmental) are always sent from the origin of the sound when needed. This implies synchronisation problems when environmental sounds should match object movements. This is today unsolved.

Algorithmically produced sounds, on the contrary, solve this problem since they advantageously can be produced at the computer where the sound is played, thus being perfectly in sync with the movements of the object.

# CHAPTER FIVE
# RESULTS AND FUTURE WORK

In this chapter I will present some observations on users of the audio system, both when using an audio test program and when using audio inside DIVE. A conclusion on the qualities of the algorithmic model is presented followed by list on proposed future work on the sound in DIVE. The chapter is concluded with a section on the expectations of 3D sound in general.

## 5.1. USER TESTS

In this section I will present some observations done in informal tests and by inspecting how new users experience the audio facilities.

Before the audio was integrated in DIVE I had a test program where the user could position and drag around eight sound sources in a two-dimensional grid. A point marked the centre where the listener was positioned, and small boxes marked the positions of the sound sources. The user could also turn on and off each of the sound sources, but actually, all eight sounds could not be turned on simultaneously because it resulted in a big cacophony where the user had to spatially separate the sounds thoroughly to be able to perceive each individual source.

All together about twenty people have tested the audio system, including both the test program and inside the DIVE environment. In early tests, the subjects were presented to a spatialised sound and were asked to judge the position of the sound and the degree of reality of the sound image. In successive tests, the subjects could freely control the positions of the sound sources as well as their own position. The subjects were now asked to judge how well the audio image responded to their movements and again the degree of reality of the sound image.

### *The Audio Test Program*

When users were presented to a sound that they did not position by themselves, they had rather big difficulties to estimate the position of the sound source. If the listener had never before heard the sound within the sound system, the position was pointed out to be somewhere on the left or right side, and the listener could not tell front from back. On the other hand, if the listener was used to the sound, and knew what it sounded like when played in front of and behind him, he became more accurate in separating front from back.

When the users got to position and drag around the sounds by themselves, a majority experienced that the perceived sound position followed their movements quite well.

The tests where the users moved around the sounds on their own, guided me in fine-tuning the parameters of the spatial calculations. Since the algorithmic approach never can convey the exact positions of sound sources, the best parameter in creating a good spatialiser, was to check for how the users experienced the sound impressions, and how well the impressions matched what they thought it should sound like.

### *Audio Within DIVE*

When the audio system was integrated with DIVE, people started to try out the sound. A common opinion was that the reverberation sounded very artificial and did not fit in. The interesting part is that almost no one remarked about the reverb in the test program, but when listening in DIVE, almost everyone disapproved.

One possible reason to this phenomenon is that the visual impressions given when using a VR system like DIVE are enough to bring up expectations on what the sounds should sound like. For instance, when in an open field, the garage sounding reverb does not fit very well, neither does it fit very well when in a small room. In the two-dimensional test program on the other hand, the users had no clue of what kind of enclosure they were in, and thus the reverberant sound only helped to externalise the sound and did not annoy anyone.

Another test on the sound in DIVE was done in the form of a game. A 100×100 meter large grid was created in a virtual world. Somewhere on that grid an invisible sound source was placed. The sound source only became visible to a user when the user got closer than five meters from the source. Now both a single user could exercise to learn to locate sounds within DIVE, and multiple users could compete to be the first one to find the sound source.

The author can proudly state that he is still unbeaten in this game. Can this have something to do with him being the implementer of the spatialiser?

## 5.2. QUALITATIVE ASPECTS ON THE ALGORTHMIC MODEL

When using algorithmic cues, one realises that this cheap model is completely satisfactory for systems where a "pretty good" 3D audio is wanted. The cues are quite raw, but are possible to use to achieve the desired effects. One problem is that some cues are perceived as being unrealistic. Especially the low-pass filtering and reverberation are the cues noticed to be "wrong".

An interesting phenomenon with the algorithmic cues is that when a listener has learned what different cues are cues for, so to speak, it does not matter any longer that they are unrealistic, but they start to work quite adequately. Of course the listener still notices that the cue is unrealistic, but he can nevertheless use it in his location of sound sources.

To disambiguate the cone of confusion and furthermore add elevation cues, the process must be taken to its full extent by the use of HRTF filters together with special hardware. When the step has been taken to use specially designed hardware, the hardware can also be used to compute the early reflections, by using a simple model of the room, to produce a convincing aural impression.

When using HRTFs the implementer does not have to think about making a fake cue sound as natural as possible, that is taken care of by the filter. Instead, issues like filter collection, the use of individual or non-individualised filter sets and filter reduction needs attention.

The important question is to consider what demands are imposed on the 3D sound that is to be integrated with a main system and the capabilities of the main system. When that is done, then one can choose the best possible solution for these demands.

## 5.3. FUTURE WORK

In this section I will list some things that can be implemented in the future. Some are quite realisable and some are more of the kind of being the author's dreams.

### *A Refined Algorithmic Spatialiser*

A first step in evolving the sound capabilities on DIVE is to extend the existing spatialiser. For instance, elevation can be brought into the calculations if some cheap and steady cues are found. This would take today's 2D spatialiser into the third dimension.

Another extension of the current spatialiser is to give priority to sounds. Since the limit currently is eight output channels, set by both computational and *perceptional limits*, there might be situations where giving priority to sound sources has to be done. The easiest way is perhaps to let the distance from the listener to the source determine, and play the eight closest sound sources, but refined algorithms can be imagined.

***Perceptional limits:*** *in a discussion in Schneider (1996), it is stated that the human hearing cannot perceive more than six to ten individual simultaneous sounds, but ignores the unperceived ones.*

The last extension, that might be somewhat harder to implement is to add Doppler effects to moving sounds. For sampled sounds it would mean that some more advanced digital sound processing is needed and those calculations might be too heavy for an 'any-computer' to perform in real time. For sounds produced algorithmically based on FM synthesis (section 4.4), it is very easy to add Doppler effects since the sounds' base is exactly a carrier frequency which can be smoothly adjusted to produce the desired frequency shift.

### *Virtual Microphones and Loudspeakers*

A virtual microphone is a *sound sink* placed somewhere in a world picking up sound and sending it further to a virtual loudspeaker that functions as a plain sound source, emitting sound. The microphone performs a very simple spatialisation by just regarding the distance to the sound source and attenuating the sound accordingly.

***Sound sink:*** *An object receiving and processing incoming sounds, much like a user peer.*

At a first glance it seems quite easy to implement virtual microphones and loudspeakers, but when considering the general case, it becomes quite cumbersome. For instance, a virtual mono microphone playing the picked up signal at one or more virtual loudspeakers is no problem. The troubles start when multiple virtual microphones should be connected to multiple virtual speakers, for instance a stereo microphone producing a two-channel direction dependent signal that is to be played over four loudspeakers, and moreover in synchronisation with each other. The interface for this becomes quite complex, especially the network routines where the loudspeakers positions have to be paired up with multiple channel sound streams contained within a sound packet needed to maintain synchronisation.

The applications for virtual microphones and speakers might be a virtual rock concert, telephones or other communication devices, or perhaps a virtual movie (which actually is a quite bad example; why not make the virtual world the movie?).

*Using HRTFs and the Convolvotron*

The most comprehensive implementation of 3D audio possible today is, as mentioned above, by the use of HRTF filters. In sections 3.3 and 3.7 the Convolvotron is presented and as one of the only products providing a complete system for spatialisation with HRTF cues and furthermore being used in a couple of virtual reality systems it is not an extraordinary choice to make.

Choosing HRTF filters to spatialise sound must however be an explicit strategy, since the cost per possible user is quite high.

*Room Simulation*

A last wish would be to implement some room simulation. This is however probably a task for a dedicated workstation, but depends on the level of complexity chosen.

What has to be done is to calculate the early reflections heard from the point where the listener is positioned and regarding the positions of the sound sources. The problem is to do this in real time where the listener as well as the sound sources and the environment change their positions and appearance continuously.

## 5.4. Expectations

The need for conveying sounds that have a spatial position have grown with visual interfaces becoming increasingly overloaded and media producers wanting us to be more immersed in the TV shows and video games we consume.

In the recent years the technology has become powerful enough to enable the exchange of the one dimensional stereo audio with the three-dimensional audio for a reasonable cost. This is what interests the mass-market producers and is also what is going to take the technology from the labs into our homes.

# REFERENCES

BEGAULT, D.R. (1991). “Challenges to the Successful Implementation of 3-D Sound”, *Journal of the Audio Engineering Society*, Vol. 39, No. 11, pp. 864–870..

BEGAULT, D.R. (1992). “Perceptual Effects of Synthetic Reverberation in Three-Dimensional Audio Systems”, *Journal of the Audio Engineering Society*, Vol. 40, No. 1, pp. 895–904.

BEGAULT, D.R. (1994). *3-D sound for virtual reality and multimedia*, Academic Press, Cambridge.

BLAUERT, J. (1983). *Spatial Hearing. The Psychophysics of Human Sound Localization*. MIT Press, Cambridge.

BORISH, J. (1984). “Extension of the Image Model to Arbitrary Polyhedra”, *Journal of the Acoustical Society of America*, Vol. 75, No. 6, pp. 1827–1836.

BURGESS, D.A. (1992). *Real-Time Audio Spatialization with Inexpensive Hardware*, Report No. GIT-GVU-92-20, GVU center, Georgia Institute of Technology, http://www.cc.gatech.edu/gvu/people/Faculty/David.Burgess.html, May 21, 1996.

CARLSSON, C., HAGSAND, O. (1993). “DIVE – a Multi-user Virtual Reality System”, In *Proceedings of the IEEE Virtual Reality Annual International Symposium 1993*, Seattle, pp. 394–400.

CHOWNING, J.M. (1971). “The Simulation of Moving Sound Sources”, *Journal of the Audio Engineering Society*, Vol. 19, No. 1, pp. 2–6.

CRYSTAL RIVER ENGINEERING (1996). *Audioreality True 3D Sound Home Page*, http://www.cre.com, May 21, 1996.

DIVE (1996). *The DIVE Home Page*, http://www.sics.se/dive/, May 28, 1996.

GARDNER, M. B. (1969). “Distance estimation of 0 degree or apparent 0 degree oriented speech signals in anechoic space”, *Journal of the Acoustical Society of America*, Vol. 45, pp. 47–53.

GAVER, W.W. (1993). “Synthesizing auditory icons”, In *Proceedings of InterCHI'93, ACM conference on Computer-Human Interaction 1993*, Amsterdam, pp. 228ñ235.

HAGSAND, O. (1996). “Interactive Multi-user VEs in the DIVE system”, *IEEE Multimedia*, Spring 1996, pp. 341–350.

HALL, D.E. (1990). *Musical Acoustics*, 2nd edition, Brooks/Cole Publishing Company, Pacific Grove, California.

IRIS (1995). *Iris Digital Media*, Online Programming Guide, Silicon Graphics Inc.

JACOBSEN, V. (1994). “Meeting Real-time Delivery Constraints”, In *Tutorial notes of SIGCOMM 1994, ACM Conference*, Boston.

JACOBSON, L., ed. (1992). *Cyberarts*, Miller Freeman Inc., San Francisco, California.

KALAWSKY, R.S. (1993). *The Science of Virtual Reality and Virtual Environments*. Addison-Wesley Publishing Company Inc., pp. 68–74, 183–187.

KLEIJN, W.B., PALIWAL, K.K., ed. (1995). *Speech Coding and Synthesis*, Elsevier Science.

LINDERHED, A. (1991). *Algoritmer för tredimensionellt ljud,*. LITH-ISY-EX-0954, Linköpings tekniska högskola, Linköping.

MIT (1996). *The Machine Listening Group Home Page*, http://sound.media.mit.edu/, May 26, 1996.

MOORER, J.A. (1977). "Signal Processing Aspects of Computer Music: A Survey", *In Proceedings of the IEEE*, Vol. 65, No. 8, August, pp. 1108–1137.

MØLLER, H. (1992). "Fundamentals of Binaural Technology", *Applied Acoustics*, 36, pp. 171–218.

NEVOT (1996). *Guide to NeVoT 3.34*, http://www.fokus.gmd.de/step/nevot/, May 27, 1996.

DE POLI, G. (1983). "A Tutorial on Digital Synthesis Techniques", *Computer Music Journal*, Vol. 7, No. 4, pp. 429–447.

POPE, S.T., FAHLÉN, L.E. (1993). "The Use of 3-D Audio in a Synthetic Environment: An Aural Renderer for a Distributed Virtual Reality System", In *Proceedings of the IEEE Virtual Reality Annual International Symposium 1993*, Seattle, pp. 176–182.

PROAKIS, J.G., MANOLAKIS, D.G. (1992). *Digital Signal Processing*, 2nd edition, Macmillan Publishing Company, New York, New York.

REICHERT, H. (1992). *Introduction to Neurobiology*, Georg Thieme Verlag, Stuttgart.

SCHNEIDER, T. (1996). "Virtual Audio", *VR News*, Vol. 5, April, pp. 38–41.

SCHROEDER, M.R. (1962). "Natural Sounding Artificial Reverberation*", Journal of the Audio Engineering Society*, Vol. 10, No. 3, pp. 219–223.

STALLINGS, W. (1994). *Data and Computer Communications*, 4th edition, Macmillan Publishing Company, New York, New York.

STEVENS, W.R. (1991). *UNIX Network Programming*, Prentice-Hall Inc. International.

SULLIVAN, R.F. (1996). *Audiology Forum: Video Otoscopy*, http://www.li.net/~sullivan/ears.htm, May 14, 1996.

SUNOS (1994). *Man pages*, "audio(7)", 14 Apr 1994 and "audiocs(7)", 31 Jan 1994, SunOS 5.4, Sun Inc.

WENZEL, E.M., FOSTER, S.H. (1993). "Perceptual consequences of interpolating head-related transfer functions during spatial synthesis", In *Proceedings of the ASSP (IEEE) Workshop on Applications of Signal Processing to Audio and Acoustics.* New York.

WENZEL, E.M., WIGHTMAN, F.L., KISTLER, D.J., FOSTER, S.H. (1988). "Acoustic origins of individual differences in sound localization behavior", *Journal of the Acoustical Society of America*, Vol. 84.

WENZEL, E.M., ARRUDA, M., KISTLER, D.J., WIGHTMAN, F.L. (1993). "Localization using nonindividualized head-related transfer functions", *Journal of the Acoustical Society of America*, Vol. 94, No. 1, pp. 111–123.

WIGHTMAN, F.L., KISTLER, D.J. (1989a). "Headphone simulation of free-field listening. I: Stimulus synthesis", *Journal of the Acoustical Society of America*, Vol. 85, No. 2, pp. 858–867.

WIGHTMAN, F.L., KISTLER, D.J. (1989b). "Headphone simulation of free-field listening. II: Psychophysical validation", *Journal of the Acoustical Society of America*, Vol. 85, No. 2, pp. 868–878.