Diploma Thesis

Telecommunications & Media
University of Applied Sciences St. Pölten

3D Audio Systems through Stereo Loudspeakers

Completed under supervision of

Hannes Raffaseder

Completed by

Miguel David Botía Fernandez
Tm060001

St. Pölten on ……

Signature ……

Abstract

In this dissertation it has been deepened on the 3D sound reproduction trough stereo loudspeakers obtaining from it conclusions about it performance and reliability. To make this possible a wide study has been done about the 3D sound systems, cues used necessary modules, implementation as well as a documentation of the different available technologies to make possible the reproduction in loudspeakers like crosstalk filters and head tracking systems. A mathematical analysis of the crosstalk filters has been done as well as also the implementation of a crosstalk filter in matlab. To obtain reliable results the crosstalk filter was tested on subjects obtaining this way data about its efficiency and performance.

Index

1- Introduction

Since some years ago some multimedia applications, videogames, communications and devices like the auditory displays are using 3D sound technology.

In the most cases to achieve the real 3D audio illusion these applications are forced to use headphones. If we use a pair of conventional loudspeakers with this technology the effect of 3D sound is lost because the sound of the right speaker interferes on the sound of the left speaker and vice versa cancelling thus the 3D perception. This effect is called crosstalk effect.
Even in the best case without crosstalk effect the point where the 3D illusion is produced (Sweet Spot) is very tiny and with only a little movement of our head the 3D illusion is lost.

But in the recent years some researchers had developed crosstalk cancellers, that avoid the crosstalk effect, and head-tracked 3D audio systems that optimize the acoustical presentation and thus produce a much more realistic illusion over a large area.

Along this dissertation we will see the history of 3D sound and how to produce it, how many parameters and processes are involved in this creation, the theory and implementation of crosstalk canceller and head-track systems and finally we will discuss the implementation and testing of a 3D audio system with crosstalk canceller.

For this purpose some localization tests will be made on different human subjects with headphones and with loudspeakers with crosstalk canceller.
After that we will discuss the results and conclusions.

1.1 - What is 3D audio?

A 3D audio system is able to position a sound in an arbitrary point of the space around a listener. The sound came from headphones or from two loudspeakers in front of the listener but the listener's perception is that the sounds come from different points in the space around him.

This is similar to the first stereo sound systems when they start to use the panning between the two speakers to create phantom images of the sound where there isn't any speaker. However this system can't position the sound at the sides or at the rear of the listener.

To accomplish these 3D systems uses parameters based on psychoacoustics and parameters like the response of the external ear or the shadowing of the head and torso and how these parameters affects the audio signal. These indicators are called Cues and we are going to see them widely later.

To obtain a 3D illusion the audio signal is processed with a procedure known as binaural synthesis and the signal is inverted using a crosstalk canceller to make possible the reproduction over loudspeakers.

1.2 - History of 3D audio

Our ability to localize sounds in a 3D environment has been studied by researchers for almost 100 years since Lord Raleigh's Duplex theory was published in 1907 [RA07]. Since then many studies have appeared trying to understand and explain the human perception of 3D audio.

Although the concept of 3D audio existed since some years, only recently computing technology is able to perform a real-time processing of the signal.
The first real time 3D audio system was developed for NASA and its first use was an astronaut communication system. In an Space-walking the astronaut must be in constant communication with the mission control and he can receive instructions from different persons and he will hear each voice in different places making easier for the astronaut to recognise the person who is talking to him and enabling less confusing in the simultaneous communications.

In 1991 two engineers (Palmer and Degani) designed a new touch screen computer panel for commercial airline pilots. The screen was tested in a flight simulator at the NASA. After the test the pilots said that they were confused because they were uncertain if they had positively engaged the virtual switch. The solution comes from the engineers of NASA that developed a MIDI control for this screen. This system creates representational auditory icons for each button and the samples were spatialized using 3D audio techniques to avoid confusion between the pilots. The experience was a complete success and after that NASA starts to develop auditory displays for their space shuttle [BEG00].

At the same time many different sound-processing technologies start to emerge due to the acceptance and standardization of digital sound in the recording industry and the diminishing price of digital technology in the 80's.
The game industry also began to improve the audio in their computer games.
The sound in games advanced from the annoying beep from a little speaker inside the computer in the early 80's to a very realistic 3D illusion nowadays.

At the moment this technology is being applied in military devices like auditory displays, but this is still in research because they are very expensive, difficult to use and voluminous.

Also some advantages in high-end are starting to solve the main problems of this systems like versatility, cost and processing capabilities necessary to obtain a suitable perform.

1.3 - Applications of 3D Audio

There are many uses for 3D audio that are known already in the entertainment, broadcasting and virtual simulation application areas. We are going to see some examples where the 3D technology is used.

As we told before in the space each astronaut has an auditory display and a 3D environment where each sound or communication is spatialized along it to avoid confusion. Commercial airlines are starting to develop auditory displays for their pilots where each notification signal is represented by an auditory icon, also called earcon. To avoid confusion all the signals are spatialized with 3D audio techniques. These systems are very suitable for high stress environments like the plane cockpit.

There are 3D audio environments specially designed for blind people where each interaction in a computer is substituted by a sound.

Within the field of entertainment we can find nowadays videogames with really good 3D environments were each sound is spatialized to make more realistic the gaming experience.

Furthermore there are some projects in music with 3D audio like AUDIOEAROTICA by Gordon Mumma, a sound sketch that probably represents one of the first combinations of 3D audio and auralization techniques for headphone music composition.

Most of these systems actually exist and are operable however some of these systems are still in development.

1.4 - Problems of 3D audio using loudspeakers

There are many problems when broadcasting a 3D sound signal through loudspeakers instead of using headphones. When using headphones the audio signal goes straight ahead to the listener's ears without any interference from the environment around the listener and each channel reach its corresponding ear without any trouble. But when using loudspeakers the sound signal has to cross an unknown environment before it reaches the listener's ears. In other words the room, the environment and the loudspeakers will affect in a non linear way to the sound signal. The environment where the signal will travel will affect to the frequency, amplitude, phase and reverberation of the signal. These factors will change more or less depending on the environment.

Another problem when using loudspeakers is that the sound reproduced in one loudspeaker can interfere with the signal coming from the another loudspeaker obtaining this way, in one ear, the corresponding signal from its speaker more the sum of the interference from the remaining loudspeaker instead of hearing only the signal from its related loudspeaker. This effect is called crosstalk and it's one of the main problems of the 3D systems using loudspeakers [WG97a]. This crosstalk will affect the spectral balance and interaural differences significantly. This problem can be solved using one technique called crosstalk cancellation. This technique is going to be explained afterwards.

In the case of the headphones playback, the distance between each speaker to the listener's ears is always the same, thus the virtual image is formed always in the same point inside the head. It's impossible to predict the position of the listener or of the speakers in any given situation and impossible to compensate for multiple listeners. With loudspeakers the problem is that the head of the listener must remain in the same place along the entire playback because the point where the 3D image is constructed, known as the sweet spot, is a very small area and only with a little movement of the listener's head the 3D illusion will be lost.

All these problems can be solved using some techniques that are going to be described later on.

1.5 - Surround versus 3D audio

One of the problems with 3D sound using loudspeakers is that control over perceived spatial imagery is greatly sacrificed, since the sound will be reproduced in an unknown environment. In other words, the room and the loudspeakers will impose unknown nonlinear transformations that usually cannot be compensated for by the designer or controller of the 3D audio system. Headphone listening conditions can be roughly approximated from stereo loudspeakers using a technique called crosstalk cancellation described afterwards.

In the commercial world, there is frequent confusion between devices that attempt to produce 3D virtual images and those that are in actuality surround sound or spatial enhancement systems. Intended for music, these systems claim to improve a normal stereo mix, via process as simple as including an “add-on” box to the home stereo system, or even as a switch on some “boom boxes”. Simply the intent of a surround sound system or spatial enhancer is to make normal stereo sound more spacious than stereo.

A further distinction needs to be made between surround sound and theatre surround sound systems. The setup used by Dolby laboratories for a 35mm optical stereo soundtrack is shown in the figure below.

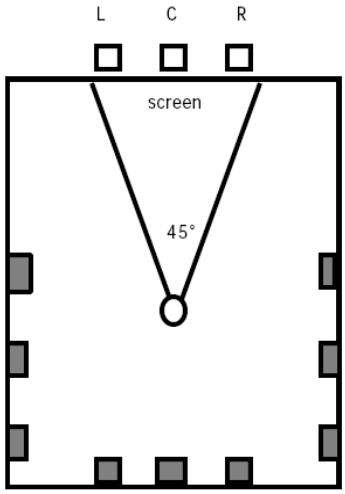


Figure 1.1 Dolby speaker arrange for stereo optical 35 mm release [BEG00]

By means of Dolby’s system it’s possible to derive four channels, center, right and left channel and multiple surround channels for the rear. In this system the speaker labelled as C is used for dialogues, L and R are used to route effects and music, and the shaded speakers are for surround sound effects. The surround channels, which are only one channel, use ambiance material like background sounds that helps to establish the environmental context. None of the systems described before are 3D sound systems. [BEG00]

1.6 - Loudspeaker displays

This thesis is based on the study of 3D sound trough stereo speakers. But what means stereo? Stereo is the reproduction of sound using two or more independent audio channels. Strictly speaking, stereo refers not to the use of two channels, but the ability of the sound system to reproduce three dimensional sounds. However we will use the terms stereo to denote two-channel reproduction.

Stereo systems have been in use for decades, and have been extensively studied. Essentially, the stereo technique relies on the ability to position a sound between the two loudspeakers by adjusting the amplitude and delay of the sound at each speaker. These techniques are called intensity panning and time panning, respectively [BEG00].

Time panned stereo is problematic. Using equal amplitude broadband signals, delaying one channel by less than a millisecond is sufficient to move the auditory event to the opposite speaker.

Intensity panning is far more effective and robust than time panning. About 25 dB of level difference is sufficient to move the auditory event completely to the stronger speaker. Intensity panning works fairly consistently with different signal types, even with narrowband signals, although high-frequency sinusoids give degenerate results. The success of intensity panning has led to a number of coincident microphone techniques for stereo.

Stereo techniques can be analyzed using phasor methods, which assume the signals are steady state sinusoids, and thus are completely specified by a complex value. The signal at one ear is the sum of the same-side speaker phasor and the opposite-side speaker phasor, which is delayed to account for the interaural time delay. Phasor analysis demonstrates that intensity differences between the two speakers result in ear signals which have the same intensity but different phase.

Stereo techniques may also be explained as a consequence of summing localization, whereby a single auditory event is perceived in response to two sources radiating coherent signals. When either one of the sources radiates a locatable signal, the auditory event appears at the location of the source. When both sources radiate the same signal in some amplitude proportion, a single auditory event is perceived at a location between the two sources, even though the actual ear signals are nor entirely consistent with this perception.

2 – Background

To understand better how a 3D audio system works we must know how the humans can localize sounds.
A sound generated in space creates a sound wave that propagates to the ears of the listener. When the sound is to the left of the listener, the sound reaches the left ear before it reaches the right ear and thus the right ear signal is delayed with respect to the left ear signal. In addition the right ear signal will be attenuated due to the shadowing of the head. Both signals are also subject to a filtering process caused by acoustical interaction with the torso, head and the pinna modifying the original frequency content of the signal. Depending on the direction of the sound it will have different pitch for each angle.

We unconsciously use the time delay, amplitude difference, and tonal at each ear to determine the location of the sound. These indicators are called sound localization “Cues”. They are split in two categories, interaural and spectral cues.

Since many years is known that the principal cues for sound localization, particularly localization to the left or right, are the time and level differences at the ears of the listener. Rayleigh [RA07] in 1907 in his Duplex theory states that the low frequencies are localized using time (Phase) differences, and high frequencies using intensity cues. Then he called ITD to the interaural time differences and IID to the interaural intensity differences. Both of them are frequency dependent, for instance the ITD at frequencies below 1500Hz is about 3/2 larger than the ITD above this frequency.

The spectral modification of a sound is due to the interaction of the sound with the external ear. Batteau in 1968 [BA68] proposed that directional cues are encoded by multipath reflections off the pinna that sum at the ear canal; the pattern of the delayed reflections depends on the direction of the source. The pinna contributes significantly to the elevation localization and also front back discrimination. Also the outer ear filters the spectral content of the audio signal.

There are also cues with the motion of the head, and with the HRTFs of the listener. The use of HRTF helps the listener to distinguish sources for instance from front and back.

After this brief description we are going to talk more deeply about these cues.

2.1 - Sound Perception

Sound is the sensation produce in the ear due to de vibration of the particles that moves through an elastic medium like the air. The sound is a wave and it travels through a medium such as air or water.

The sound source consists of undulations within the elastic medium of air resulting from the vibration of a physical object, such as vocal cords within a larynx, or of a loudspeaker cone. If the sound propagates in a manner irrespective of direction, it can be said to be omnidirectional forming thus a spherical field of emission. If the sound source is within an environmental context without reflections, for instance an anechoic chamber, then beyond a certain distance the sound waves arrive at the front of a listener as a plane field meaning that the sound pressure will be constant in any place perpendicular to the direction of propagation.

In a nonanechoical environment the sound arrives to the listener by both direct and indirect paths creating a field called diffuse field due to the effect o the environmental context. In this case the environmental context is the main component of the medium in the case of spatial hearing.

The transmission path from the source to the listener can be described with the model source, medium and listener.

Figure 2.1 Source-Medium-Listener model

The sound reaches the listener and it's interpreted by the ear. The ear is divided in three parts the outer, the middle and the inner ear.

In the next figure is showed an overview of the human auditory system.

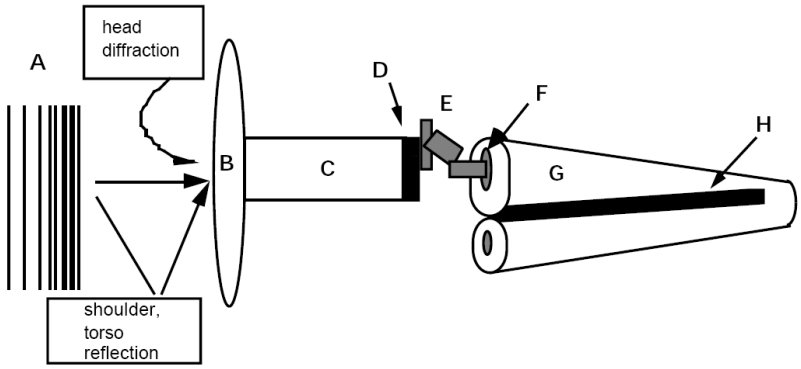


Figure 2.2 Diagram of the ear [BEG00]

The outer ear (A) is formed by the pinna, the visible part of the ear, and proximate parts of the body such as the head and the shoulders.

The purpose of the pinna (B) is to collect sound. It does so by acting as a funnel, amplifying the sound and directing it to the ear canal. While reflecting from the pinna, sound also goes through a filtering process which adds directional information to the sound. The filtering effect of the human pinna preferentially selects sounds in the frequency range of human speech.

The pinna works differently for low and high frequency sound. For low frequency it directs sounds toward the ear canal. For high frequencies some sounds gets directly to the ear canal and others reflect off the contour of the pinna first and they reach the canal with a very slight decay. That delay can be translated into phase cancellation where the frequency component whose wave period is twice the delay period is virtually erased. This is known as the pinna notch where the pinna creates a notch filtering effect.

Following this are the effects of the meatus (C) or ear canal that leads to the middle ear. The ear canal protects the eardrum (D) and acts as a resonator, providing about 10 dB of gain to the eardrum at 3300Hz.

The middle ear consist of the eardrum (D) and ossicles (E).The tympanic membrane, colloquially known as the eardrum, is a thin membrane and it's function is to transmit sound from the air to the ossicles.

The ossicles are the three smallest bones in the human body. The ossicles are in order the hammer, anvil and stirrup. The hammer articulates with the anvil and is attached to the eardrum, from which vibrational energy is passed. The anvil is connected to both the other bones. The stirrup articulates with the anvil and is attached to the membrane of the oval window between the middle and the inner ear. The ossicles are also known by the Latin terms Malleus, Incus and Stapes.

Sound is transformed at the middle ear from acoustical energy at the eardrum to mechanical energy at the ossicles, then the ossicles transforms the mechanical energy into fluid pressure within the inner ear, the cochlea (G), via motion the oval window (F).

The fluid pressure causes frequency dependent vibration patterns of the basilar membrane (H) within the inner ear; which causes numerous fibres protruding from auditory hair cells to bend. These in turn activate electrical action potentials within the neurons of the auditory system, which are combined at higher levels with information from the opposite ear. These neurological processes are eventually transformed into aural perception and cognition including the perception of spatial attributes of a sound resulting from both monaural and binaural listening.

2.2- Spatial Hearing

Spatial hearing is the group of techniques used by the humans to localize a sound in the space around them. These techniques are called cues and they are hereafter explained.

A distinction must be made between the two kinds of spatial hearing. Natural spatial hearing refers to how we hear sounds spatially in everyday hearing, with our ears uncovered, our head moving and in interaction with other sensory input. This is not only confined to two ears or binaural hearing, one-ear hearing can provide enough spatial information. A special case of binaural hearing is virtual spatial hearing, this refers to the formation of synthetic spatial acoustic imagery using a 3D sound system and stereo headphones or speakers with the respective crosstalk filter.

In a 3D audio system the operator by changing some parameters has a complete control of the spatial auditory perception of someone. A 3D sound system uses processes that either complement or replace spatial attributes that existed originally in association with a given sound source. This is referred to as spatial manipulation and this process involves not only engineering based on physical parameters, but also psychoacoustics considerations.

Only the imagination and the tools at hand pose limits to the spatial manipulations. With the virtual acoustic simulations we can replicate an existing spatial auditory condition like a violinist playing in a big hall or we can create a completely unknown spatial auditory condition like hearing a sound source with the size of a fly under the water.

Fundamentally spatial perception involves an egocentric frame of reference, measurements and orientation of sound images are given from the listener's position. In Psychophysical studies the reference point for describing distances and angles of sound images is located at an origin point between the ears, approximately at eye level in the centre of the head.

For describe the angular perception of a virtual sound source we are going to use the azimuth and the elevation.

Azimuth perception is the more robust of the two because the human ears are almost opposite positioned on either side of the head. That layout makes easy to the humans to localize sound sources in a plane parallel to the surface of the ground. Normally azimuth is described in terms of degrees, where 0 degrees elevation and azimuth are at a point directly ahead of the listener along a line bisecting the head outward from the origin point. In some systems azimuth is described as increasing counter clockwise from 0-360 degrees along the azimuthal cycle.

For descriptive purposes is more convenient to split the azimuthal plane in two halves, right and left, and then increase clockwise from 0 to 180 degrees.

Elevation increases upward from 0 to 90 up where is positioned in a point above the listener, and directly below the listener at 90º down. In software implementation the elevation is measured with a 360º polar model.

Azimuth and elevation can indicate the position of a sound source only within the sphere that surrounds the listener's head, but if we want a complete localization of the sound we will need the distance. This is because the sound perception is multidimensional.

Finally there is the environment context. This refers to the effects of reverberation caused by repeated reflections of a sound from the surfaces of an enclosure. Reverberation can potentially cause a significant effect on how a sound source is perceived. The effect of the environmental context is manifested by a set of secondary sound sources dependent on and made active by the primary localized sound source.

There are two categories of perceptual effects caused by this, an effect on a given virtual sound source's location and the formation by the listener of an image of the space occupied by the sound source.

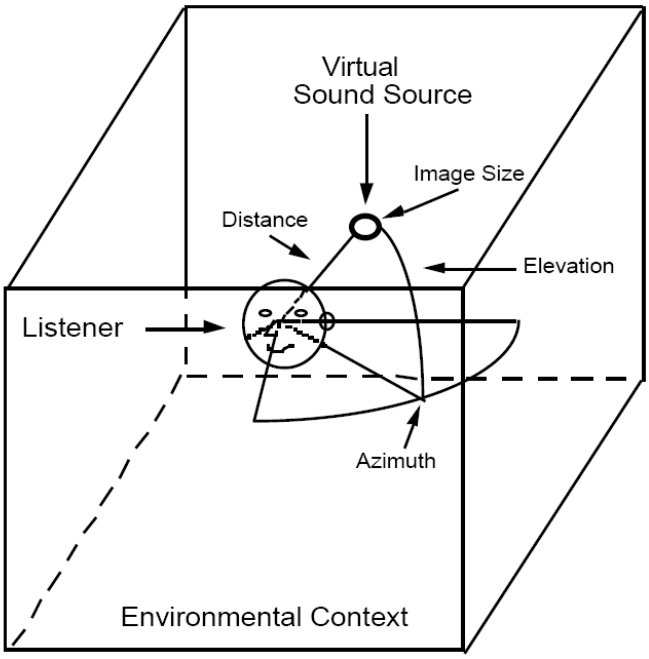


Figure 2.3 Taxonomy of Spatial Hearing [BEG00]

2.3 - Interaural cues IID and ITD

The most important cues for localizing a sound source's angular position involve the relative difference of the wavefront at the two ears on the horizontal plane. The horizontal placement of the ears maximizes differences for sound events around the listener, rather from below or above.

In this part we will see the frequency dependent cues of interaural time differences (ITD) and Interaural intensity differences (IID) [RA07]. In order to describe these differences cues, psychoacoustic experiments are constrained according to a lateralization paradigm. These experiments involve the manipulation of ITD and IID in order to determine the relative sensitivity of physiological mechanism to these cues. The word lateralized has therefore come to indicate a special case of location where the spatial percept is heard inside the head and the means of producing the percept involves manipulation of interaural time or intensity differences.

The experimental designs have focused generally in three areas.
The first area are the measurements of "Just noticeable differences" (JNDS) for different interaural conditions and stimuli and the "Minimum audible angles (MAA). The second are studies about interaction between the cues specially the way in which each cue cancels the effect of the other. The third area involves assessments of the overall efficacy of interaural cues. [MID91]

Lateralization mimics the interaural differences present in natural spatial hearing .To see this better consider a listener in an anechoic chamber with a perfect round head and no outer ears with a source oriented directly ahead on the median plane (A) and a source displaced to 60 degrees azimuth (B) at the eye level.

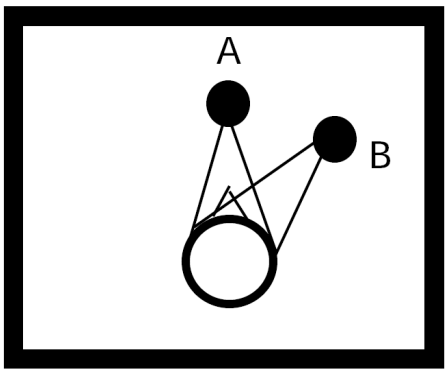


Figure 2.4 Diagram of the situation [BEG00]

Modelling this situation involves to calculate two paths representing the sound source waveform from its centre to two points representing the entrance to the ear canal. With the source at position A at 0 degrees azimuth the path lengths are equal, causing the wavefront to arrive at the eardrums at the same time and with equal intensity. At position B the sound source is at 60º azimuth to the right of the listener, and the paths are now different and the sound will arrive first to the right ear than to the left ear.

The sound path length difference described is the basis of the interaural time difference (ITD) cue, and it relates to the hearing system's ability to detect interaural phase differences below approximately 1 kHz.

The sound source at position B will also yield a significant interaural intensity difference (IID) cue, but only for waveforms with a wavelength smaller than the diameter of the head, for instance frequencies greater than 1.5 kHz.
In this case higher frequencies will be attenuated at the left ear because the head acts as an obstacle creating a shadow effect.

Independent of frequency content, variations in the overall difference between left and right intensity levels at the eardrum are interpreted as changes on the sound source position from the perspective of the listener.

At the beginning the earlier research into spatial hearing found results that suggested that IID and ITD cues operates in exclusive frequency domains. One neurological component for spatial hearing responded to ITD cues only for frequencies below 1.5 kHz and the other component operates in high frequencies above 1.5 kHz. But later research has revealed that timing information can be used for higher frequencies because the timing differences are detected in amplitude envelopes.

For frequencies with half periods larger than the size of the head, it's possible for the auditory system to detect the phase of these waveforms, but for frequencies above 1.5 kHz the period is smaller than the size of the head creating an ambiguous situation. But if the sine waves are increased in amplitude, via amplitude modulation, then amplitude envelope is imposed on the sine wave. The auditory system somehow extracts the overall amplitude envelope for higher-frequency components at both ears and measures the difference in time of arrival of the two envelopes.

Although the literature claims that lateralized sound sources are heard only along the interaural axis, there is also a vertical component experienced by some listeners, even lateralized sound is heard sometimes behind, and sometimes in front of the interaural axis by listener. [VB60]

We must talk also about the precedence effect. The precedence effect, also called de Haas effect due to the first author that writes about it, explains an important inhibitory mechanism of the auditory system that allows one to localize sounds in the presence of reverberation. [HH72], [BL83]

If we take stereo headphones or loudspeakers and we process the left channel using a variable delay device we will obtain a series of perceptual events.

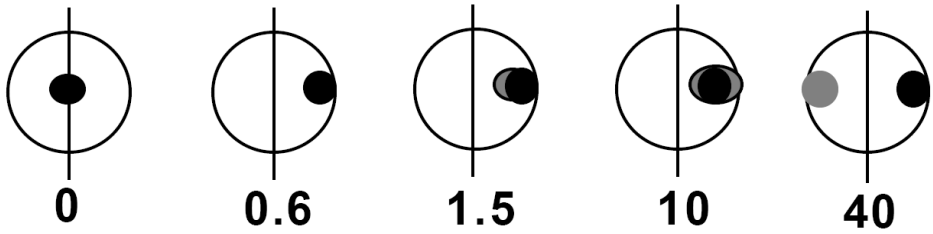


Figure 2.5 perceptual effect of increasing ITD from 0 to 40 msec in the left ear. [BEG00]

If we delay the signal of the left ear between 0-60 the virtual image shifts along the lateral axis, in the range of 7-25 msec the lateralization of the source is still to a position associated with the undelayed channel. At 1.5 msec, a delay greater than the maximal ITD value for lateralization, the width will continue to increase over a certain range. Other characteristics also change when the delay is increased like tone colour and the loudness. At 10 msec the center of gravity of the source can also move back toward the center. At a somewhat ambiguous point, marked as 40 msec, the broadened source will split into two separate images where the latter image is called echo. Other percepts that change as the time delay is increased can include tone colour, loudness, and a slight shift position toward the delayed channel. [BL83]

2.4 - Head motion cues

In every day perception we use head motion with our aural and visual senses in an attempt to acquire sound sources visually. When we hear a sound we wish to localize, we move our head in order to minimize the interaural differences, using our head as a sort pointer.

Several studies have shown that allowing a listener to move the head can improve localization [WA40] [TH67]. Listeners apparently integrate some combination of the changes in ITD, IID and the movement of spectral notches and peaks that occur with head movements and use this information to disambiguate, for example, front imagery from rear imagery.

Consider a sound at right 30 degrees azimuth with another source at 150º, the listener would attempt to center this image by moving his head to the right, the sound source becomes increasingly centered because interaural differences are minimized and the sound source must be in the front, but if instead the differences become greater as the head is turned then the source must be rear.

Unlike natural spatial hearing the integration of cues derived from head movement with stereo speakers will provide false information to the listener for localizing a virtual sound. With headphones the head movement has no effect.

2.5 - Spectral cues of the pinna

If we look at the figure 1.6 we will notice that the sound source at position A will have identical ITD and IID as the source at position B, and the same for the position X and Y. Along the positions of this circle the IID and the ITD would not change. This is only theoretical because in the figure we assumed that the head is completely spherical. With a real person this would never be possible. But when ITD and IID cues are maximally similar between two locations a potential for confusion between the positions exist in the absence of a spatial cue other than ITD and IID.

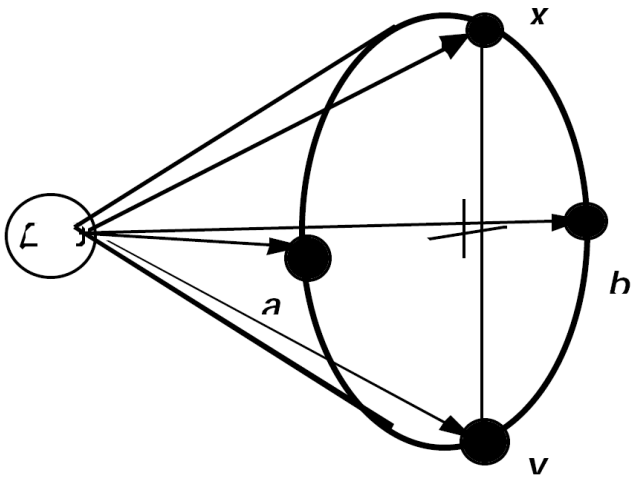


Figure 2.6 The cone of confusion. [BEG00]

In fact identical values for ITD and IID can be calculated for a sound everywhere on a conical surface extending out from the ear. This is called the cone of confusion in the literature [MI72]. The cone of confusion has influenced the analysis and design of many location studies. The ability to disambiguate sources from front to back or from above and below in cases where the ITD and IID would not supply this information has brought about hypotheses regarding the role of spectral cues and localization. The most significant locationally dependent effect on the spectrum of a sound source as it reaches the eardrums can be traced to the outer ears or pinna.

2.6 – Head Related Transfer Function

The spectral filtering of a sound source before it reaches the eardrum that is caused primarily by the outer ear is termed the head-related transfer function (HRTF). The binaural HRTF can be though as frequency-dependent amplitude and time-delay differences that result primarily from the complex shaping of the pinna.

The folds of the pinna cause several time delays within a range of 0-300μsec that cause that the spectral content at the eardrum differ from a sound measured with an omnidirectional microphone [BA68]. The shape of the pinna causes this spectral modification and the complex construction of the outer ears causes a unique set of time delays, resonances and diffractions, this means that there is a unique HRTF for each sound source position [BA68] [BL83]. Like a sound thumb print the HRTF alters the spectrum and timing of the input signal in a location-dependent way that is recognized by a listener as a spatial cue.

HRTF is considered the key component for a 3D sound system. This is based on the theory that the most accurate means to produce a spatial sound cue is to transform the spectrum of a sound source at the eardrum as closely possible to the way it would be transformed under normal spatial hearing conditions.

Depending on the criteria for each application we will use a different outer ear model for the measurements of the HRTF. Some measurements are only taken for the outer ear and the head, others incorporate features of the body, such as ear canal, shoulder and torso.

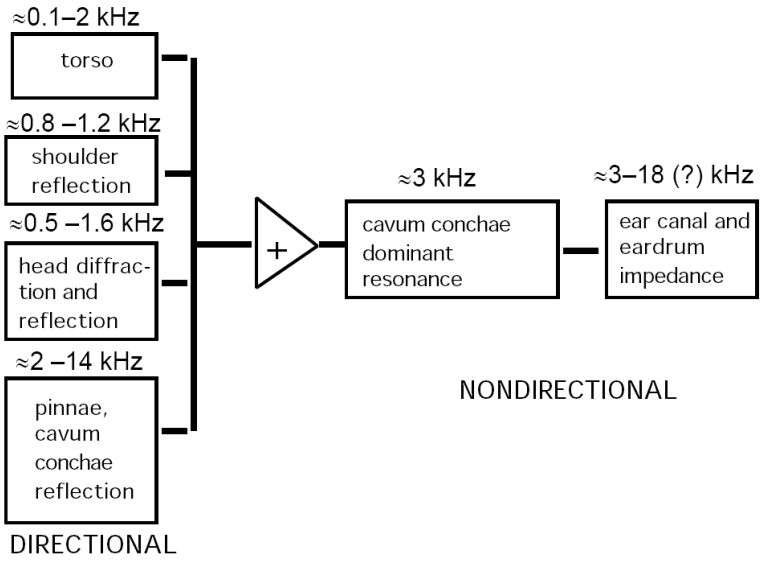


Figure 2.7 Diagram of HRTF components [GIE92]

In the figure 2.7 is showed a diagram of various HRTF components in terms of directionally dependent and independent cues [GIE92]. The relative importance of the cues is arranged bottom, most important, to top. In each feature is indicated the range of frequencies where that part has an effect.
For instance the shoulder reflection describes a directionally-dependent influence of the upper body and shoulders between 100 Hz-2 kHz in the media plane.

To all the influence of the outer ear we must add the effect of the cavum conchae and the ear canal. The ear canals are a resonance tube and any resonance tube will effect a spectral modification to an incoming sound.

2.6.1 – How to obtain HRTF data

The HRTF can be obtained from an individual or from a dummy head by playing an analytic signal at a desired position, at least a meter distance, and measuring the impulse response with probe microphones placed inside the ear canals. These measurements can be formatted to use directly within spatial filtering DSPs.

To obtain the HRTF data a sound source generates an acoustical signal whose time and frequency characteristics are known beforehand. The optimal place to take the measures is an anechoic chamber because it minimizes the influence of reflected energy.

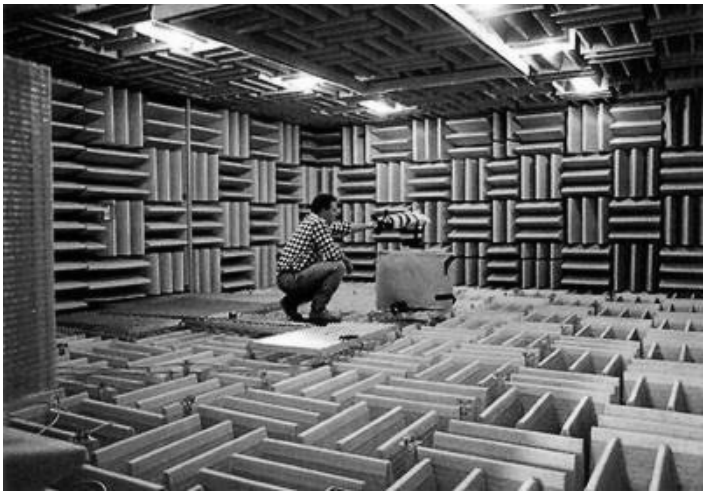


Figure 2.8 Example of an anechoic chamber

The binaural impulse response of the outer ear is obtained from the pressure responses of the microphones and digitally recorded in stereo as a series of sampled waveform points. Both ears are measured at the same time. The data obtained is reformatted and loaded into a DSP filter for left ear spectral modification and similarly for the right ear. The sound source to be spatialized is filtered by these DSPs during simulation.

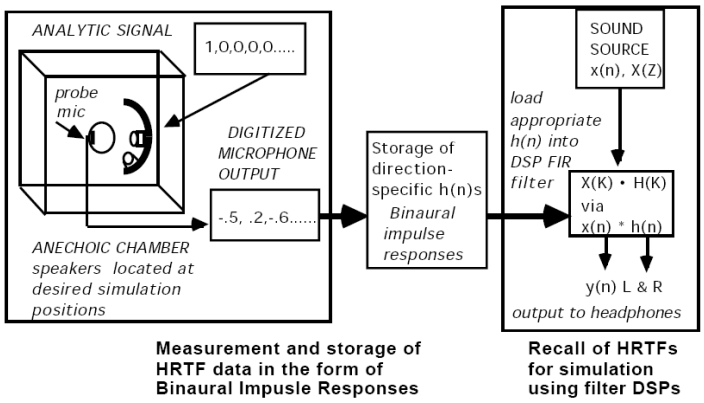


Figure 2.9 Overall plan of the HRTF measurement-storage-simulation technique [BEG00]

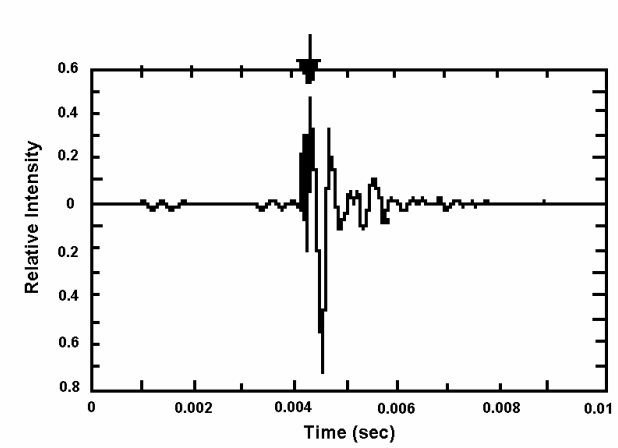


Figure 2.10 Time domain impulse response measured in the ear canal in anechoic chamber. The arrow indicates the peak of the response [BEG00]

There is a used method to obtain the HRTF impulse known as maximum-length sequence (MLS) [RV89] where one can extract the impulse response of the system after using a MLS technique by cross-correlating the sequence with the resulting output, thereby deconvolving the pseudo-random sequence with the digital sequence recorded at the microphone, in other words, you get the same result as the impulse response, but with a better signal-to noise ratio.

Another robust technique is to use a swept sine type of analytic signal; this technique is referred to as time delay spectrometry (TDS). The analytic signal in this case can be thought of as a sine wave that is swept quickly across the entire frequency range in a short elapse of time

2.6.2 – Collecting HRTF Measurements

Nowadays no single scientifically accurate method has been established for measuring HRTFs. Constantly researchers are exploring new techniques and equipment to improve signal to noise ratio of the measurement hardware and to determine the optimal place for the measurement microphone.
We can find a lot of literature from different authors about their method to obtain the HRTF measurements.

In most of the cases the designer of a 3D audio system will want to obtain a library of HRTF measurements. Already some HRTFs are available on computer networks and there is a paper with the HRTF magnitude data in numerical form for 43 frequencies, however no phase data is included in this table. [SH74]

In 3D sound applications intended for many users, we want to use HRTFs that represent the common features of a great number of individuals.
For this we can use the HRTFs measures of a person with a very good accuracy for localization, for example an experimented musician, and use for the rest of the people. After some test was showed that its better when any nonindividualized HRTF set is psychoacoustically validated using a statistical sample of the intended user population otherwise the system will be purely subjective.

There are many techniques for measure HRTF; between them we can find the techniques from authors like Blauert [BL83], Shaw [SH74], Griesinguer [GRI90] and many others. Here we are going to describe the Wightman and Kistler technique because it exemplifies a laboratory-based HRTF measurement procedure. [WK89a]

In this technique, following Wightman and Kistler, a probe microphone is placed in the ear canal of the listener at a distance of approximately 1 to 2 millimetres from the eardrum. An extremely thin custom-model shell made of Lucite was used to mount the microphone at the entrance of the ear in such a way as to occlude the ear canal as less as possible.

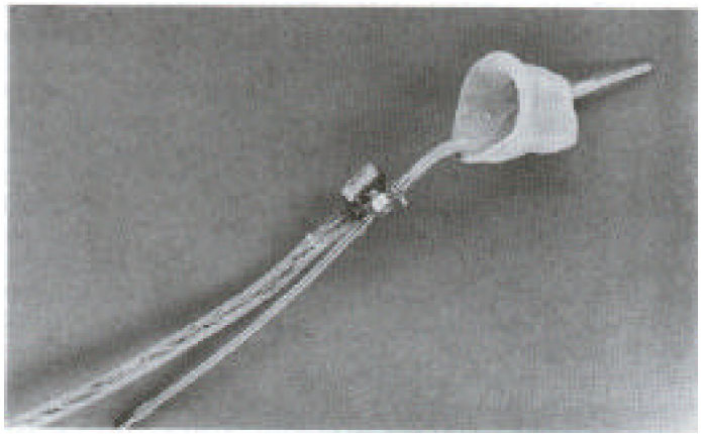


Figure 2.11 Probe Microphone used for HRTF measurements [WK89a]

Figure 2.12 Placement of the microphone [WK89a]

The measures are taken in an anechoic chamber. The listener is surrounded by eight loudspeakers mounted at ±36, ±18, ±0, ±54, ±72 and ±90 degrees elevation on a movable arc at a distance of 1.4 meters from the head. This kind of layout allows all desired elevations to be obtained at a particular azimuth.

Figure 2.13 A subject in an anechoic chamber with the movable arc [WK89a]

To measure HRTFs from different directions, one can move the analytic signal source around the listener. To avoid this and to take better HRTF samples in the Armstrong laboratory of Wright-Patterson Air force base they built a 4.5 meters diameter geodesic sphere with 272 speakers mounted on it. The whole cage structure was built in absorbent material and each speaker is at every 15 degrees azimuth and elevation. The location of each speaker represents a measured position. The listener and the speakers do not need to be moved, facilitating the collection of HRTFs measurements.

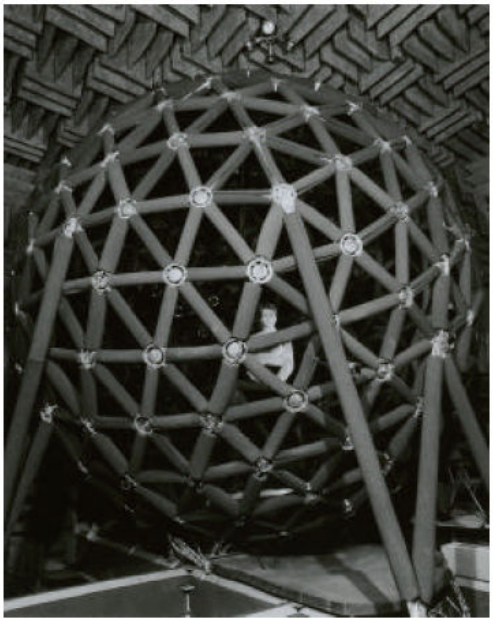


Figure 2.14 The "Speaker Cage" of Armstrong laboratory [BEG00]

The optimal 3D audio system is that that uses the HRTFs of the listener that is using it, but in many applications the taking of individual HRTF impulse response measurements is impractical. It would be better to use the measurements that represent the best compromise between the HRTFs features of a given population.

To obtain these Nonindividualized HRTFs we should do an average of the HRTF of a group of listeners. This average can be originated from the analysis of either physical or spectral features. In the analysis of the physical features we can measure for instance the size of the cavum conchae and the length of the ear canal of a large number of people and then average. Another method is to examine the spectra of many binaural impulse responses via Fourier transform and perform a spectral averaging [SH74]. It's possible than this type of averaging will diminish spectral features of the HRTF relative to a listener.

Another method is to use a technique called PCA (Principal Components Analysis) [WK92]. In this technique all the HRTFs of different positions are evaluated simultaneously. Using statistical techniques one can isolate spectral features that change as a function of direction and those that remain more constant. The result of this technique is a set of curves that fall out of correlation analysis and that collectively explain all the variations in the analyzed data.

PCA HRTFs and Transfer functions are as robust as the actual measurements. Some experiments in subjects showed that there is a high correlation between responses to the synthesized and measured conditions.

Different type of average binaural impulse responses can be collected using a dummy head instead of a real listener. There are many different dummy heads and each one represents a particular design approach to create a standardized head with an average pinna.

The measurement with these dummy heads is very easy because most of them have a built in microphone inside the head. The result will be more replicable since the microphone and the head remain in the same position along the measurement process. Another advantage is that 3D sound systems based on dummy head HRTFs will be closely matched to actual recordings made by the same binaural head allowing compatibility between the two different types of processing.

When using a dummy head one must assume that the manufacturer has constructed features according with the manufacturer specifications. One head might sound more natural to a particular set of users than another. This could be caused by the microphones used, the manner used for simulating the ear canal or the dimensions and shape of the dummy head. The size of the head is probably the major component in the suitability of one dummy head versus another due to the importance of the interaural delay present in HRTFs for spatial hearing, for instance Asian people on average have a smaller head than Europeans and correspondingly a smaller pinna.

There is one device also used for measurements called KEMAR, a standard audiological research mannequin manufactured by Knowles Electronics. KEMAR is a complete head and torso simulator, its design evolved through a careful averaging of anthropomorphic data. It allows several varieties of pinna to be mounted on the head; it also includes an ear canal simulator. The coupling of the ear canal simulator to internal microphones is designed to meet international standards, facilitating audiological research for hearing aid devices.

Figure 2.15 KEMAR mannequin head and torso [BEG00]

There are other devices similar to the KEMAR like the HATS (Head and Torso Simulator) developed by Brüel and Kjaer for acoustical research evaluation of headphones and other types of binaural measurement, it also includes a mouth simulator. The geometry of the head is symmetrical and based on adult anthropomorphic data.

Figure 2.16 HATS for acoustical research [BEG00]

2.6.3 - Equalization of HRTFs

The raw impulse measurement must undergo both time and frequency domain modifications before being formatted for use in a DSP. There will also be a need for numerical formatting and conversion inherent to the particular architecture of the DSP, for instance from 16-bit floating point to 24-bit signed integers. Normally there are some software routines that perform these conversions within integrated measurement and synthesis systems or signal-processing packages.

The first time domain operation performed on a set of raw impulses is to discard the blank portion at the beginning of the impulse from the time it takes the impulse to travel from the speaker to the microphone. After this a normalization procedure is applied to make the best use of the available digital quantization range. To apply this first we take the loudest sample in a sequence and then multiplying all samples of that sequence making the loudest point of the sample the maximum quantization value. Note that if we normalize all impulse responses individually the IID cues could be compromised.

A final time domain take place because the importance of overall ITD [WK92]. One can customize the delay inherent to each binaural HRTF impulse. The process consists in insert or subtracts blank samples at the beginning of each impulse. The overall ITD cue can be customized to a particular head size.

The postequalization of HRTFs is applied in order to eliminate the potentially degradative influences of each element in the measuring and playback chain. Most of the spectral nonlinearities are originated by the elements in this chain like the loudspeaker used to play the analytic signal or the measuring microphone. To repair this nonlinearities the postequalization a gain with the same value as the attenuation should be applied on the affected frequencies.

It is possible to describe this process theoretically using a mathematical representation where the spectra of each element are either multiplied with or divided from each other. The frequency domain transfer functions for one ear can be represented as follows. [BEG00]

A(Z) = Analytic signal;

M(Z) = Probe microphone;

C(Z) = Ear canal;

L(Z) = Loudspeaker;

HP(Z) = Headphone;

H(Z) = Naturally occurring HRTF in a free field;

RAW(Z) = Uncorrected HRTF for virtual simulation;

COR(Z) = Corrected HRTF for virtual simulation;

INV(Z) = Inverse filter for correcting RAW(Z);

YE(Z) = The signal arriving at the eardrum;

YM(Z) = The signal arriving at the probe mic;

X(Z) = An input signal to be spatialized.

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

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

In a virtual spatial hearing simulation we want that

$$YE(Z)_{virtual} = YE(Z)_{natural}$$

First the uncorrected HRTF is measured for a particular direction by playing the analytic signal through the loudspeaker.

$$RAW(Z) = YM(Z) = A(Z)M(Z)C(Z)H(Z)L(Z)$$

After this the headphone and the ear canal are measured by playing the analytic signal through headphones.

$$YM(Z) = A(Z)M(Z)C(Z)HP(Z)$$

To obtain *COR(Z)* is necessary to find the inverse filter *INV(Z)*

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

Then the uncorrected HRTF is corrected as follows:

$$COR(Z) = RAW(Z)INV(Z)$$

$$= \frac{A(Z)M(Z)C(Z)H(Z)L(Z)}{A(Z)M(Z)C(Z)HP(Z)}$$

$$= \frac{H(Z)L(Z)}{HP(Z)}$$

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

$$YE(Z)_{virtual} = [X(Z)COR(Z)][HP(Z)C(Z)]$$

$$=[X(Z)\frac{H(Z)L(Z)}{HP(Z)}][\ HP(Z)C(Z)]$$

$$=X(Z)H(Z)C(Z)L(Z)$$

And finally natural spatial hearing and the virtual simulation are equivalent:

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

The equalization of HRTF impulses is also used to boost the lack of bass frequencies in the impulses measured with probe microphones.

2.6.4 - HRTF Magnitude characteristics

In the next figures there are shown examples of the HRTF from three persons measured in the same laboratory. The sources were positioned in 90º degrees azimuth at each side of the listener.

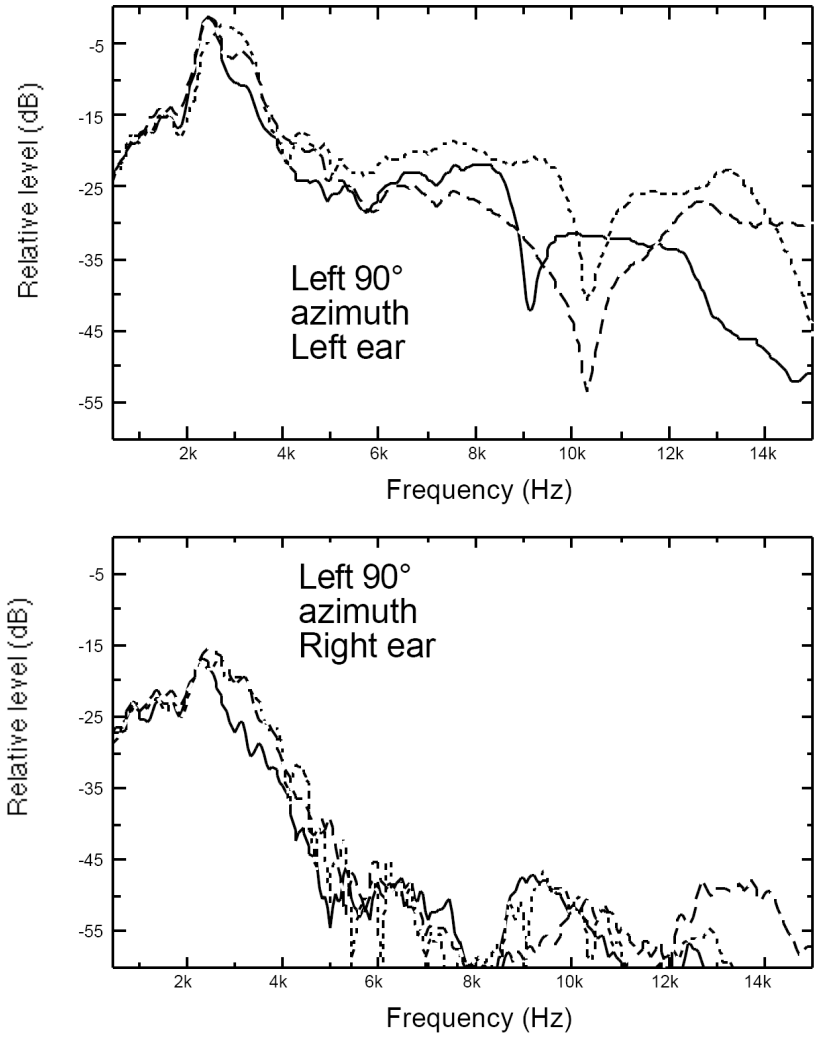


Figure 2.17 HRTFs left and right ear, 90 degree of three different persons [WW93]

Note not only de spectral changes as a function of position but also the differences between the three subjects. These differences are caused by the fact that the size and shape of the pinna is different on every listener. It is important to know that in a 3D audio system is incapable of using the HRTF of the listener, in most cases the system uses nonindividualized HRTFs. This kind of HRTFs can be derived from individualized HRTF averages, although this can diminished the minimum and maximum of the spectra. [BEG00]

2.6.5 - HRTF phase characteristics

There are also changes in interaural phase as a function of frequency that are caused by the HRTF. When broadband signals, with all the audible frequencies, reach the pinna, some frequencies will arrive later than others to the eardrum. This delay can be measured in degrees by realizing that a phase delay of 360 degrees equals a delay of one cycle of the wavefront. For instance a wave with a frequency of 1000 Hz delayed 360º equals to 0.001 sec of time delay. The figure below shows, from 500 to 4000Hz, the unwrapped phase differences for a person in increments of 30 degrees azimuth, from 0 to 150 degrees, at 0 degrees elevation. The bold lines show the conversion from phase delay to interaural time delay.

The phase response at a single pinna is less critical for determination of localization than the interaural time difference [BEG00].

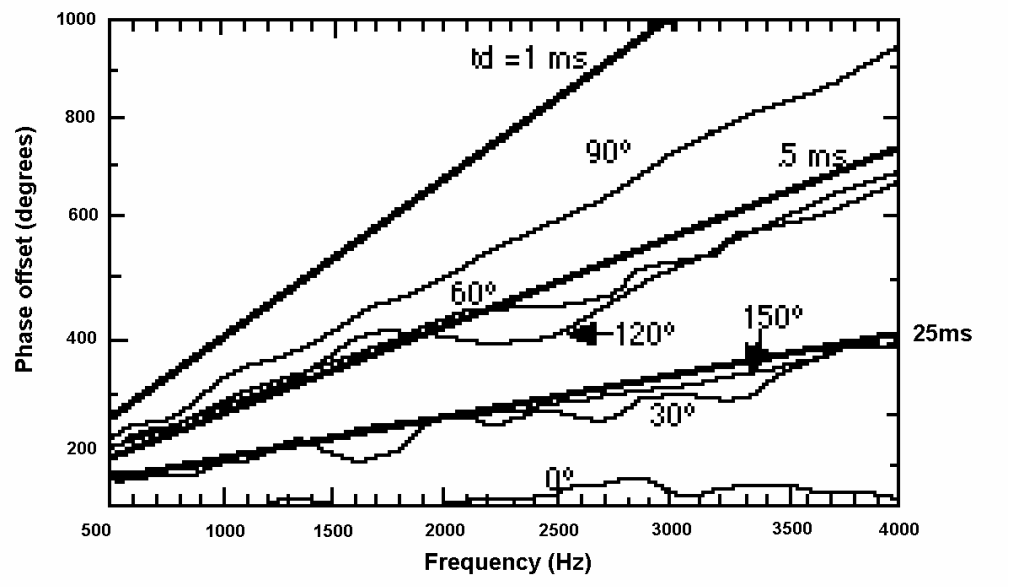


Figure 2.18 Phase difference for 0, 30, 60,90, 120 and 150 degrees azimuth. Bold lines show interaural time differences [BEG00]

2.6.6 - Localization with HRTF cues

In a 3D sound system the inclusion of ITD and IID based on the spectral alteration of the HRTF is more realistic and accurate than the IID and ITD described previously in the lateralization. But this is only a qualitative sense not in terms of azimuth localization accuracy. The main role of HRTF cues from psychoacoustic standpoint is though to be the disambiguation of front from back and from up and down for sources situated on the cone of confusion.
Figure 2.19 shows the spectral difference of the HRTF of one listener at one ear between two positions in the cone of confusion at 60 and 120 degrees azimuth. The main differences are in the upper frequencies, especially at 5 and 9 KHz.

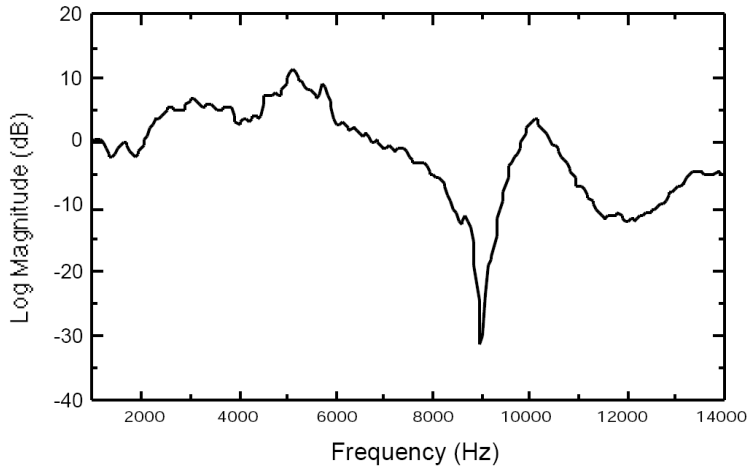


Figure 2.19 Spectral differences between front and back source location at 60 and 120 degrees. [BEG00]

Theories surrounding the role of HRTF spectral cues involve boosted bands [BL69], covert peaks [BU87] and spectral troughs or notches [ST68], all suggest that a major cue for elevation involves movement of spectral through and peaks. Changes in the HRTF spectrum are effective as spatial cues, for example the movement in the center frequency of two primary spectral troughs could contribute to the disambiguation of front-back source positions on the cone of confusion.

Many experiments with listeners suggested that spectral cues can influence the perception of direction independent of the location of a sound source. For example Roffler and Butler and their studies determine that vertical localization requires significant energy above 7 kHz. Blauert used third octave filters to determine the directional bias associated with the spectral band in terms of three categories, above, in front or behind. The table below shows the approximate regions where the directional bands existed. The relationship between these bands and the spectral peaks of the HRTF led to the formation of a theory of boosted bands as a spatial cue.

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

Figure 2.20 Bandwidth and frequencies for directional cues. [BL83]

In the other hand Middlebrooks [MID92] used narrowband noise at 6, 8, 10 and 12 KHz. The results of Middlebrooks's study were the biases of judgments for elevation up and down and for front back, as a function of the center frequency.
The study was tested on real listeners and divided in three categories, overall results, tallest subject and shortest subject.

	Overall results		Tallest subject		Shortest subject	
Fc	u/d	f/b	u/d	f/b	u/d	f/b
6 kHz	U	F	U	F	U	F
8 kHz	D	F & B	D	F	U	B
10 kHz	D	B	level-D	F	D	B
12 kHz	level-D	F & B	level-D	F	D	B

Figure 2.21 Middlebrooks results. U = Up, D =Down, F = Front, B = Back. [MID92]

These localization biases were independent of the target direction. There were significant individual differences, due to the height and the size of the pinna of each one.

If we compare the two studies from Blauert [BL69] and Middlebrooks [MID92] show that directional biases can result from stimuli that mimic particular spectral modifications of the pinna. However the effect of these spectral modifications will depend on each listener.

A significant problem for the implementation of 3D sound systems is the fact that spectral features of HRTFs differ among individuals. The localization of virtual sources can be degraded when listening through another set of pinna. In these cases there's a significant decrease in azimuth localization accuracy using an artificial pinna instead the listener's own pinna.

The idea of using the HRTFs of the pinna of a good localizer was taken into a study of localization accuracy where several subjects gave judgments of localization through pinna that were not their own. The results are shown in the table below.

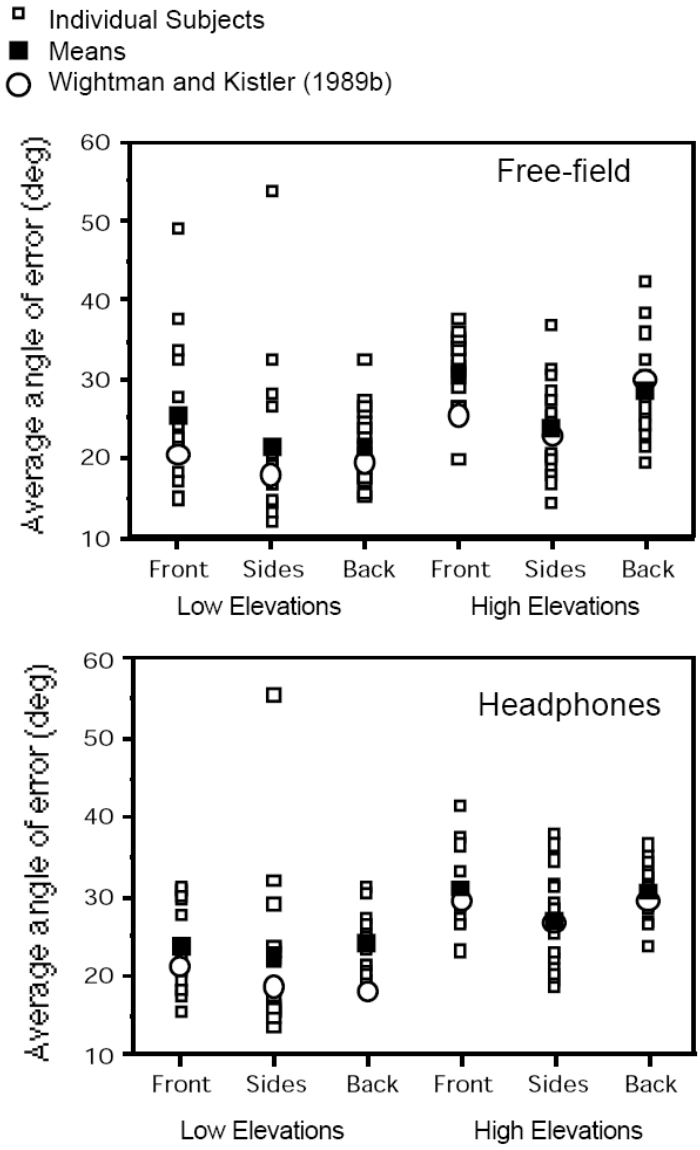


Figure 2.22 results of average angles of localization errors for various spatial regions [BEG00]

The results were divided in five areas, front (45º Right to 45º left); sides (60º to 120º left and 60º to 120 right), back (135º left to 135º right), up (18º up to 54º up) and down (36º down to 0º). The test was performed on 16 subjects. First the test was performed in free field conditions and then they test again with headphones on each listener using non individualized HRTFs. In the graphics also appears the data from the study of Wightman and Kistler for comparison.[WK89b]

Overall the error is similar between the virtual and headphone condition especially the overall trend of the measures. There is degradation in elevation localization on some listeners and an increase in the number of front-back, up-down reversals compared with the free field conditions where the listener's own pinna was used.

Also in localization when using non individualized HRTFs there is a variation of performance on each subject. In 1993 Begault and Wenzel [BEW93] make a study about the performance. They used speech stimuli for the measures. The result were that from all the subjects involved in the study six made a relatively accurate judgment, four of them pull toward their judgements to left and right (90 º) and only one pulled toward his judgement through the median plane (0º and 180º).

Martens in 1991 [MA91] also examined the apparent location of HRTF filtered speech stimuli by using short segments of speech like /i/, /ae/, /u/. The subjects in this study listened to this stimuli processed through nonindividualized HRTF at various elevation positions. The elevation judgments were very accurate in this experiment but hi notice that the front-back position was influenced by the brightness of the stimuli. Stimuli with energy concentrated at higher spectral frequency tended to be heard to the front and stimuli with energy concentrated at lower frequency were heard to the rear. Surprisingly when the stimuli were low pass filtered at 5 kHz the judgments of elevation remained accurate.

This contrasts the notion that elevation perception is based on higher-frequency spectral features of HRTF.

It's important to realize that the previous localization judgments have been reversal corrected. If the target was right 30 degrees azimuth and the judgment was 130 degrees the judgment is recorded at 50º. Reversals are very common in both free field and virtual acoustic simulations of sound. The listeners normally turn their heads to the source to disambiguate front from back.

2.7 - Distance cues

The inclusion of distance and environmental context effects within a 3D audio system is almost imperative for maintaining a sense or realism among virtual acoustic objects. In fact the reverberation is the key for an effective simulation of distance.

Further investigations of distance and environmental context effects have been facilitated by the development of more computationally powerful 3D sound systems. With the advent of computer assisted room modelling algorithms and binaural measurements of actual halls the spatial aspects of reverberation are now receiving more attention from researchers. This area is known as auralization and it can create either truly veridical room simulations and reveal for instance how the material of a wall absorbs the noise of the street.

Distance and environmental context perception involve a process of integrating multiple cues including loudness, spectral content, reverberation content and cognitive familiarity. But to obtain consistent results psychoacousticians must minimize the number of variables. For example the reverberation is usually not included in distance perception studies.

There are two kind of experimental method when studying auditory distance studies [BEG00]. The first methodological difference is whether an absolute or relative sense of distance is being evaluated.

Absolute distance perception refers to a listener's ability to estimate the distance of a sound source without the benefit of the cognitive familiarity. Relative distance perception includes the benefit from listening to the sound source at different distances over time.

The second methodological difference has to do with whether a listener is asked to estimate the apparent distance of the virtual sound image, for example the sound seems to be a meter away from my head, or to estimate the distance of the actual sound source, for example when the listener is asked to choose from which speaker the sound is coming within a group of speakers. In this case the ventriloquism effect comes into play, whereby the apparent position of a virtual sound image is influenced by the presence of a correlated visual object.

For 3D systems the most relevant studies are those where subjects give distance estimates of the virtual sound image without visual cues. What is relevant to virtual reality applications are the studies of interaction between visual objects and aural distance cues in a virtual world.

2.7.1 - Intensity, Loudness Cues

In the absence of other acoustic cues, the intensity of a sound is the primary distance cue used by a listener [CO63]. Auditory distance is learned from a lifetime of visual-aural observations, correlating the physical displacement of sound sources with corresponding increases or reductions in intensity. This is one of the means used by the humans for many survival tasks like when knowing when to step out of the way when an automobile is coming from behind. In the world of audio a sound engineer adjust the volume of a track in a multitrack recording to establish a difference between the background and the foreground of the recording.

As a cue to distance, loudness or intensity probably plays a more important role with unfamiliar sound than with familiar sounds. For instance when listening to sounds just before going to bed in a familiar or in an unfamiliar place. In the familiar place although the sound of a car in the street is louder than the clock of the kitchen we know that distance of the car is larger than the distance to the clock, familiarity allows distance estimations that can be reversed if we only use the cue of intensity. But if we camp outdoor in an unfamiliar environment, the distance precepts of different animal noises would probably follow an intensity scale.

To establish the relationship between sound source distance and intensity in anechoic conditions one can use the inverse square law to predict sound intensity reduction with increasing distance.

Given a reference intensity and distance, the intensity of a sound suffers an attenuation of 6 dB each time we double the distance from an omnidirectional source in an anechoic chamber. If the source instead of omnidirectional is a line source then the intensity reduction is commonly adjusted to 3 dB for doubling distance. This adjustment is used in noise-control applications.

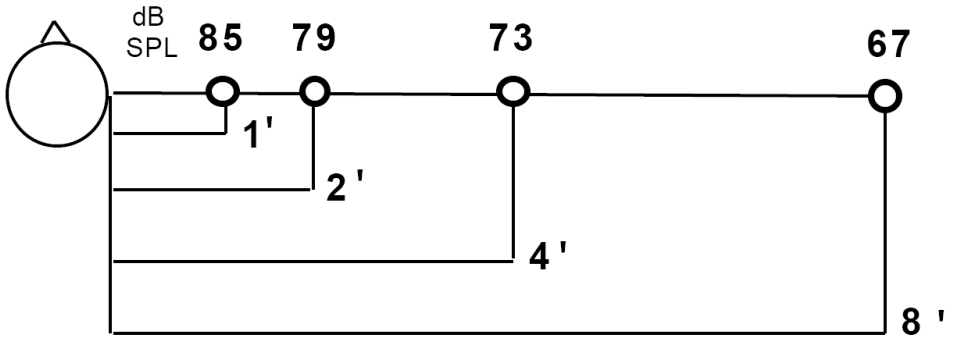


Figure 2.23 Reduction of intensity in dB for virtual sources using the inverse square law at 1, 2, 4 and 8 feet from the listener. [BEG00]

In a 3D sound system, using this law, one can specify several sound sources at progressively doubled distances.

A theoretical problem with the inverse square law is that intensity, measured in dB, expresses the level of a sound source's intensity to a reference level, but loudness is the perceive magnitude of intensity. In a 3D sound system is desired that loudness is the only available cue for distance where the relative estimation of doubled distance follows half-loudness instead of half-intensity.

In a 3D sound system a loudness scale can be used to adjust intensity for scaling the apparent distance of virtual sound sources images. The judgments of half or doubled loudness are related to the sone scale more than to the inverse square law. Sones are a unit that relate equal loudness contours (Isophones) for a given frequency to loudness scales. If we double the number of sones we are doubling the loudness. For instance an orchestra can range loudness from 40 to 100 dB SPL, this ranges from 1 to 50 sones. Each time the distance between the source and the listener is doubled there is a decrease of 10dB in the loudness scale of sones.

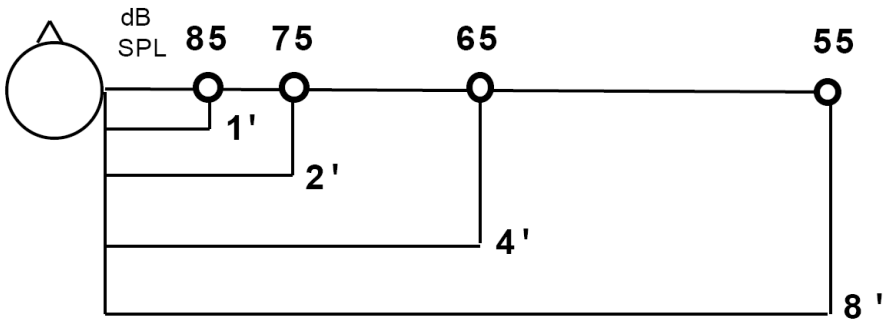


Figure 2.24 Reduction of intensity using a loudness scale based on sones [BEG00]

Now the question is if in a 3D sound system is better to use a loudness scale instead of the inverse square law to adjust relative virtual auditory distances. If judgments of loudness are the primary cues to judgments of distance, then 10 dB would be more appropriate for a general scaling factor in a 3D sound system. But loudness increments can only properly operate as a distance cue under conditions where other cues such as reverberation are not present. The operation to calculate the loudness of one source requires a complicated analysis, the role of the energy on each critical band will affect judgments of absolute loudness.

Remember also that the designer of a 3D audio system can't predict the sound intensity at the eardrum of the listener at the end of the communication chain since the end user of an audio system will always have final control over the overall sound pressure level.

In the case of a 3D audio system trough speakers, if the sounds are not heard in an anechoic environment the conjunction with reverberation and other sound sources can preclude the use of either the inverse square law or a loudness scale. The total sound intensity reaching a listener by direct and indirect paths can be calculated by using a room simulation program.

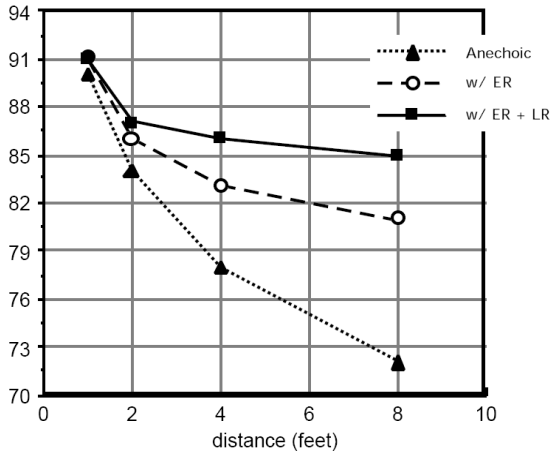


Figure 2.25 Reduction in intensity under anechoic and reverberant conditions. [BEG00]

In the figure above is showed the reduction of intensity under anechoic and reverberant conditions using an omnidirectional source at 1, 2, 4 and 8 feet from the listener. ER means early reflections and LR means late reflections. The triangles show the 6 dB reduction of the inverse square law under anechoic conditions. The circles show the reduction in intensity when only early reflections and the squares represent the intensity when all the reverberant energy is included. Note how the overall SPL does not change more than 3 dB between the closest and the farthest listening position.

What is the influence of the HRTF to the absolute estimates of distance? To answer this, a study used a speech stimuli recorded under anechoic conditions and it was played back at a nominal level of 70 dB, this is the level for a normal speech at 1 meter from the listener. 11 subjects were told to report a distance of the sound image within 4 distances, at the center of the head, inside the head but not centered, at the edge of the head and outside of the head. The means of the results are gathered in the figure below.

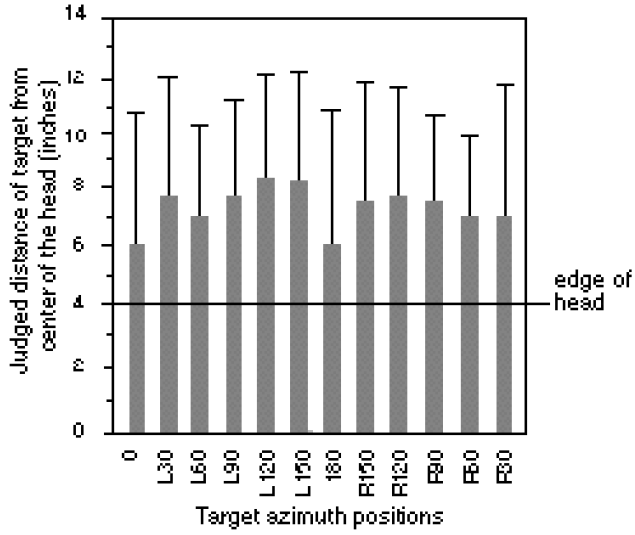


Figure 2.26 Means and deviations for distance judgments of anechoic speech stimuli. [WW93]

The results show an overall underestimation of the apparent distance if the normal speech distance is located 1 meter away from the listener. One of the reasons for this could be the absence of reverberation in the sample stimulus. Note that the standard deviation bars indicate a high amount of variability among subjects with familiar stimuli like the speech. This not implies that subjects are more accurate in estimating distance than another, but it suggests that different spectral weightings from the various HRTFs used could have influenced judgments.

2.7.2 - Influence of expectation and familiarity

Distance cues can be modified as a function of expectation or familiarity with the sound source. Sheeline [SE82] said that the listener's familiarity with both the source signals and the acoustic environment is clearly a key component in any model for auditory distance perception. If the sound is completely unknown for the listener then the listener may need more time to familiarize with the parametric changes in loudness and other cues that occur for different simulated distances. If the sound source is associated with a a particular location from repeated listening experiences, the simulation of that distance will be easier than the simulation of a distance that is unexpected or unfamiliar.

Gardner [GA69] conducted several studies for speech stimuli that illustrate the role of familiarity and expectation. In one experiment distance estimations were given by subjects of a sound source at 0º from numbered locations at 3, 10, 20 and 30 feet in an anechoic chamber. The perceived distance was a function of the sound pressure level at the listener instead of the actual location of the loudspeaker. But with a live person speaking inside the anechoic chamber subjects based their estimates of distance on the manner of speaking rather than on the actual distance.

Listeners overestimated the distance of shouting in reference to normal speech and underestimated the distance of whispering, although the opposite should have been true if intensity were the relevant cue.

2.7.3 - Spectral cues to distance

The spectral content of a sound source relative to a receiver position can vary as a function of its distance. This effect includes the influence of the atmospheric conditions, molecular absorption of the air and the curvature of the wavefront. All of these factors modify the spectral content of the signal along with the spectral cues of the HRTF. These cues are relative weak compare to loudness, reverberation and familiarity cues. One can might possibly effect a change in the perceived distance by using all these cues directly in a 3D sound system, but the experience with several algorithmic implementations suggest that this is not the case. Nevertheless this information may be useful in developing accurate auralization models.

In a 3D sound system the sound sources are constantly changing their location, this means that their spectra are constantly changing as well , making it difficult to establish any type of spectral reference for perceived distance. If one eliminates the cues arising from changes in loudness and in the reverberant structure the binaural system is very poor at determining the distance of a sound source.

The wavefront from a sound source a distance away from the listener will be planar when it reaches the listener's ears, but the wavefront from a closer source will be curved, this add an emphasis to lower versus higher frequencies of the sound source. The sound suffers a darkening effect when the sound source is approaching to us [CO63]. This effect is related to the equal loudness contour, which show that sensitivity to low frequencies increases with increasing sound pressure level. This tone darkening cue is not as usual than the experience of diminished high frequency content that occurs when a sound source increase the distance. This absorption is due mainly by the air absorption.

The air absorption depends on the humidity and the temperature of the air, in the literature the constants for calculating absorption coefficients are functions of air temperature and humidity. An air absorption coefficient can be calculated that represents the attenuation of sound as a result of viscosity and heat during a single pressure cycle. For example in the data given by Harris [HA96] show that for a distance of 100 meters, an air temperature of 68º Fahrenheit and 20% humidity the attenuation at 4 kHz will be of 7.4 dB.

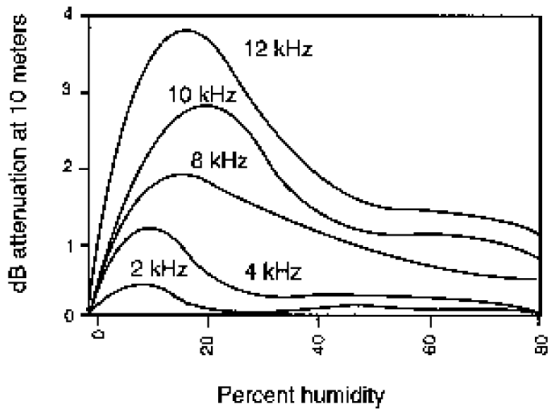


Figure 2.27 Curves of Harris for the absorption of sound in the air as a function of humidity. [HA96]

An interesting aspect of the air absorption coefficients is that they are a time varying phenomenon especially in indoor environments with HVAC systems (Heating-ventilation-cooling) because the humidity and the temperature will constantly be in flux.

For the auditory simulation the intensity of high frequency energy is scaled in relation to the intensity of low frequency energy with both nearby sound sources, closer than 2 meters, and distant sound sources, farther than 15 meters.

In the case of distance simulation using headphones the stimuli is processed with low-pass, high-pass and cut-off depending on the case. The low pass stimuli were perceived farther away than the high pass stimuli. Perhaps in the absence of other cues a low frequency emphasis applied to a stimulus would be interpreted as more distant compared to an untreated stimulus.

There are other factors that can contribute to the frequency dependent attenuation of a sound source with distance along humidity. There are also factors like the ground cover, sound barriers or the wind profile that can modify this effect.

The relevant data was gathered by noise control specialist who determines what the effects of existing or modified environmental features will have on the dispersion of noise sources. All this information has a considerable potential to be included within the modelling of a virtual environment.

For instance Gill in 1984 [GI84] calculated the attenuation for a sound source 100 meters distant at 500 Hz and 1 kHz as a function of wind profile and ground cover.

Type of attenuation		500 Hz	1 kHz
wind profile:	(downwind)	minimal	minimal
	(upwind)	up to 30 dB	up to 30 dB
ground cover:	short grass	3 dB	3 dB
	thick grass	5.4 dB	7.4 dB
	trees	8 dB	10 dB

Figure 2.28 Environmental effects on the spectra of a 100 meter distant sound source [GI84]

If we look at this data, to create a realistic acoustic mapping of an outdoor virtual environment, the creator of the virtual world have to known when the lawn was last mown and what the current wind direction was at a given moment.

It's assumed that beyond a distance of about one meter, the spectral changes caused by the HRTF are constant as a function of distance. But when one ear is turned at 90º azimuth towards a relatively close sound source, the ITD is greater at lower frequencies than for a sound at a greater distance, as a function of the wavefront no longer being planar but instead spherical as a function of frequency. This has been termed auditory parallax and has been interpreted by some to mean that the accuracy of estimation of a sound from the side should be improved when compared to distance perception on the median plane, but other authors found no significant differences in distance localization. In fact, these binaural cues are probably overwhelmed by other distance cues like loudness and reverberation.

Lateralization studies conducted with headphones result in inside-the-head-localization (IHL). IHL also occurs with 3D sound techniques, especially without reverberation. This is very important when using headphones instead of loudspeakers, because loudspeakers usually are heard outside the head. However, Hanson and Kock [HK57] obtained in 1957 IHL with two loudspeakers in an anechoic chamber, each playing the same signal but 180º out of phase. IHL must be avoided to eliminate the Necker cube illusion where the listener doesn't know if the sound is outside or inside his head.

Researchers after many experiments and studies have found that reverberation, either natural or artificial, enhances the externalization of 3D headphone sound. The inclusion of not only direct paths but indirect paths of reflected sound is necessary for a realistic simulation.

2.8 - Reverberation cues

A sound source's direct sound is defined as the wavefront that reaches the ears first by a linear path, and reflected sound, reverberation, refers to the energy of a sound source that reaches the listener indirectly, by reflecting from surfaces within the surrounding space occupied by the sound source and the listener. The surrounding space is a type of environmental context. If there are no obstructions between the sound and the listener the direct sound can be considered to be equivalent to a sound within an anechoic chamber.

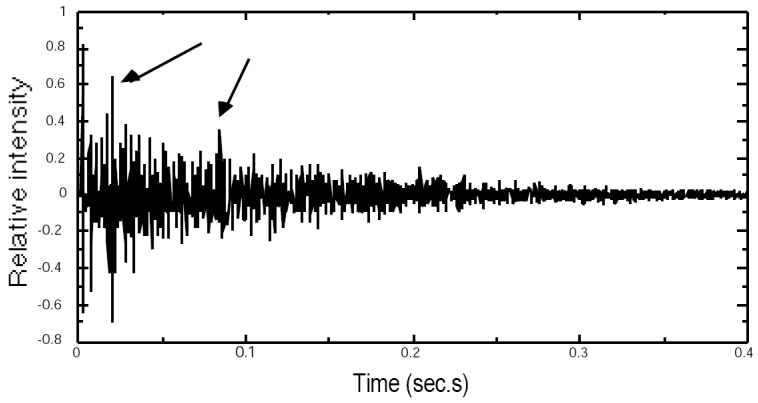


Figure 2.29 Impulse measured in a room with an omnidirectional microphone. Arrows indicate early reflections. [BEG00]

The impulse of the figure above was obtained by recording a loud impulsive noise such as a starter pistol being fired with an omnidirectional microphone. An impulse response is a useful time domain representation of an acoustical system, including rooms and HRTFs.

In the figure below are showed two possibly significant early reflections. A particular reflection within a reverberant field is usually categorized as an early reflection or as a late reverberation, depending on the time arrival to the listener. Significant early reflections, those with significant amplitude above the noise, reach the receiver within a period around 1 to 80 msec, depending on the proximity of reflecting surfaces to the measurement point.

The early reflections of a direct sound are followed by a more spatially diffuse reverberation termed late reverberation or late reflections. These later delays are the result of many subsequent reflections from surface to surface of the environmental context. This temporal portion of the impulse response contains less energy overall.

In the figure 2.30 D is the direct sound, the direct sound I followed by the early reflections (ER) that are within a period from 1 to 80 msec, after all comes the late reflections (LR). The late reverberation time is usually measured to the point when it drops 60 dB below the direct sound level.

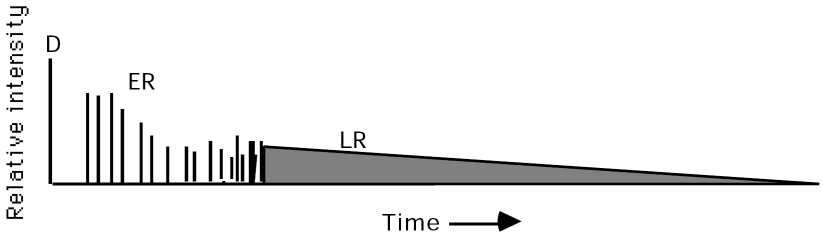


Figure 2.30 Reflectogram of an impulse response [BEG00]

In a theoretical room the build up of successive orders of reflections begins to resemble an exponentially decaying noise function during the late reverberation period causing individual reflections to be lost in the overall energy field. Individual early reflections are less likely to be masked than later reflections. The psychoacoustic evaluation of reverberation places most of its emphasis on the first 80 msec of the response, although this cut-off point is largely arbitrary.

In reverberation models the early reflections are modelled directly by tracing sound wave paths in a way like tracing light beams. This is a valid procedure for plane wavefront smaller than the incident surface. Due to the increase in density, due to the reduction in intensity of late reflections over time and the finite limits of computational power it's not common to model each individual reflection in the late reverberation period, but instead to model them using a Gaussian distribution with decaying envelope.

Figure below show how the calculation of the time of arrival and angle of incidence of a single reflected sound is taken into account.

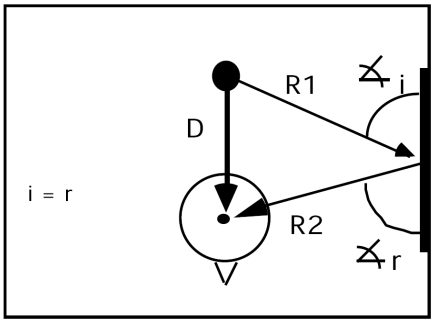


Figure 2.31 Sound reflections off of a reflective surface in an anechoic chamber. [BEG00]

The angle of incidence to the wall equals the angle of reflection to the listener, and the angle of incidence from the wall to the listener relative to the listener's orientation determines the spatial component of the reflection. Multiple reflections will arrive from many directions, but in many circumstances the significant early reflections come from a limited set of directions. The temporal aspect of the reflection is derived from the distance it must travel, divided by the speed of sound in the enclosure. Finally, an overall attenuation is the reflection will occur as a function of the inverse square law.

Depending on the number of times the sound bounce on the surface there is an increase in the reflection order. First order reflections have bounced from one surface and then to another surface before reaching the receiver, second order reflections bounce in one surface and then to another surface before reaching the listener and so on. The intensity of the reflection is progressively reduced by each successive bounce since the material of the wall will absorb a portion of the total energy in a frequency dependent way. However the number of low amplitude reflections is quite large and gets larger with each successive order, causing the combined effect of late reverberation.

Early and late reflections are described by three physical parameters, duration, usually evaluated as reverberation time or t60, the ratio of reverberant to direct sound, the R/D ratio; and in terms of several criteria related to the arrival rime and spatial incidence of early reflections.

The first parameter t60 or reverberation time is defined as the duration necessary for the energy of reflected sound to drop 60 dB below the level of the direct sound. The value of t60 is somewhat arbitrary; its original meaning was to simply indicate a value for relative silence. If the loudest sound in an environment was 90 dB and the ambient noise was 40 dB the t60, as opposed to t90, would be adequate measure, because the reverberation below 40 dB would be masked by the noise floor. In terms of the listening experience, t60 is related to the perception of the size of the environmental context; a larger value for t60 is usually associated with a larger enclosure.

The second parameter, the R/D ratio is a measurement of the proportion of reflected to direct sound energy at a particular receiver location. As one moves away from a source the level of the direct sound will decrease, while the reverberation level will remain constant. This is interpreted primarily as a cue to distance. The point where the reflected energy is equivalent to the direct energy in amplitude is called the critical distance.

The third parameter is related to temporal and spatial patterns of early reflections. The spatial distribution of early reflections over time in a given situation exists as a function of the configuration between sound source, listener and physical features of the environmental context. Different spatial temporal patterns can affect distance perception and image broadening, as well as perceptual criteria that are related to the study of concert halls. By measuring the similarity of the reverberation over a specific time window at the two ears, we can obtain a value for interaural cross-correlation.[BEG87]

2.8.1 - Perceptual aspects of reverberation

Relevant physical parameters of reverberation that affect the perception of an environmental context include the volume or size of the environmental context, the absorptiveness of the reflective surfaces, and the complexity of the shape of the enclosure. The size of the environmental context is cued by the reverberation time and level. The absorptiveness of the reflective surfaces will be frequency dependent allowing for cognitive categorization and comparison on the basis of timbral modification and on the basis of speech intelligibility. Finally the complexity of the enclosure will shape the spatial distribution of the reflections to the listener, especially the early reflections.

Late reverberation is informative perceptually as to the volume of a particular space occupied by a sound source. This is noticeable when a sound source stops vibrating and one can hear the time it takes for the later reflections to decay into relative silence and the shape of the amplitude decay envelope over time. The long term frequency response of a room, reverberation time for each third octave band, can also be significant in the formulation of a percept of the environmental context.

The first studies in concerts halls emphasized reverberation time as a strong physical factor affecting subjective preference. However t60 gives a very limited description of the reverberation amplitude envelope over the time some authors have pointed the usefulness of measuring the reverberant level over time. It's better to analyze the late reverberation level as a function of time, within each critical band, or in third of octave bands, across the entire audible range. The distinguishing characteristic between concert halls and another reverberant environment is the decay pattern of each of these frequency regions. Many of these regions do not decay exponentially although this is a standard approximation used to describe what happens between time 0 and t60.

The reverberant to direct sound (R/D) ratio has been cited in many studies as a cue for distance. Sheeline [SE82] found that reverberation was an important adjunct to intensity in the formation of distance perceptions. Reverberation provides the spatiality that allows the listener to move from the domain of loudness interferences to the domain of distance interferences.

Von Békésy [VB60] noticed that when he changed the R/D ratio the loudness of the sound remained constant, but a sensation of changing distance occurred. Through this alteration in the ratio between direct and reverberant sound can indeed be used to produce the perception of a moving sound image. He also observed that the sound image's width increased with increasing reverberation, along with this increase in distance there was an apparent increase in the size of the sound source. While direct and indirect sounds reach a listener in different proportions as a function of their distance, it's true that in some context the possible variation in the R/D ratio can be limited by the size of the particular environment. For instance, in a small room the ratio would vary between smaller limits than in a large room.

Sheeline [SE82] and Békésy [VB60] found that reverberation defines a boundary for distance judgments. There's an upper limit to how much reverberation can be mixed with a sound before it reaches an auditory horizon.

The reverberation of an environmental context can serve as an aid in identifying a likely range of sound source distances. Sheeline [SE82] comments on the contribution of reverberant energy as allowing the listener to set the appropriate boundaries for a source or a group of sources. If reverberation serves as a cue to the extent of the environmental context's size or characteristics, particularly the distance to the boundaries of surfaces, then the cognitive process, as informed by audition and any other available senses, could cause perceptual limits to the possible extent of a sound source distances.

An example of this is the sense of intimacy in small enclosures like an elevator, in this case if someone speaks to you in this enclosed space seems closer than when in a larger space although the physical distance is the same. With loudness equalized speech sources processed only spatialized early reflections subjects generally found that a virtual sound source at two meters distance seemed closer in a larger modelled enclosure than a small modelled enclosure. The physical difference between these two situations was that the early reflections were spread out over a greater time period and were weaker relative to the direct sound than in the small enclosure.

While the presence of reverberation can improve externalization of virtual acoustic images, it can also decrease localization accuracy under real simulated conditions. This happens in situations where the strength and timing of early reflections is such that the attack portion of the amplitude envelope of a sound is smeared in time and space.

In a localization study using the variable reverberation acoustics of an adjustable room, it was found that lateral early reflections degraded localization accuracy. The expanded apparent source width that can occur with reverberation could also make the estimation of a location more difficult since there is a larger area to place the azimuth and elevation judgment point.
The figure below shows the difference in localization accuracy for HRTF processed speech between anechoic stimuli and stimuli processed with synthetic spatialized reverberation. For most target azimuths, localization is worse with the addition of reverberation. This demonstrates how the precedence effect only partially suppresses the effects of reflected energy. [BEG92]

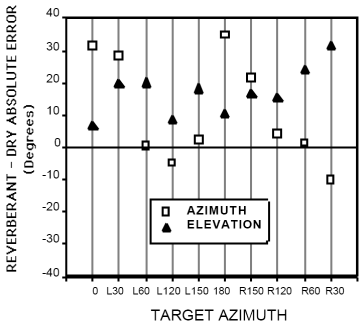


Figure 2.32 Azimuth Difference between reverberant and anechoic stimuli. [BEG92]

2.8.2 - Specific perceptual effects on early reflections

Echolocation is the effect of an initial reflection on perceived pitch or timbre. This is a feature used by the blind people to determine the distance to a surface. They use to whistle or to use clicking sound to estimate the distance width even material of the objects placed in the ± 90 º frontal arcs. Some people are better at echolocation than others, but most people tend to improve their skills. There's no reason to believe that such ability could not be eventually be made relevant to an everyday user of a 3D sound system. [KE62]

The timing and intensity of early reflections can potentially inform a listener about the dimensional aspects of the environment immediately surrounding the sound source. The reverberation gives clues for categorize the location of the sound sources according to a previous knowledge of the environment. The information of a particular spatial percept is a result of not only the direct sound, but its fusion with the early reflections, as exemplified by the effects of precedence, masking, and temporal integration.

Early reflections are not usually perceived as separable sound sources, but they are indirectly perceived in terms of the overall effect on a sound source. The relative intensity and temporal pattern of early reflections can influence the timbre of speech and even the intelligibility level. The presence of echoes can cause a disastrous situation in the design of conference rooms and spaces with large sound system installations. This has resulted in a number of room modelling approaches that are implemented within computer CAD programs.

The temporal resolution for integration of speech or music is around 20msec. Nevertheless, the perceived timbre and attack qualities of sound sources, especially of musical instrument sounds, can be changed substantially by early reflections without affecting intelligibility [SC73]. Besides altering the timbre of the instrument, early reflections can also modify the rise time of the attack portion of the amplitude envelope. The influence of early reflections has also been explored in connection with concert hall acoustic, as a mean of analyzing what physical parameters are responsible for making a good-sounding hall. In this aspect the spatial incidence of early reflections becomes important.

Surprisingly, consideration of the environmental context by using reflected sound as a variable parameter has only been recently explored in psychoacoustic studies of distance and azimuth perception [BEG92] [SE82]. In the concert hall reverberation studies they usually only study the perceptual effects of reverberation, particularly of early reflections. These studies are investigations of a very specific type of environmental context where the main data comes from the distance between the source and the listener and this data is only evaluated in a few positions. The character of the reverberation is also restricted in these studies since it must favour the all the types of music from the 19Th century to present.

As a result the criteria used for evaluating reverberation and the timing and the amplitude of early reflections according to changes in the modelling of the enclosure are toward a rather specific goal.

The temporal and spatial aspects of early reflections are especially significant in forming a preference criterion in a hall. The overall finding is that lateral reflections, located at ± 90º, should dominate over the reflections that come from the front or the rear. This dominance can be regulated through the arrival time, direction and level of the reflections. Lateral reflections should have little time differences between them to obtain a stereo effect on the hall. To obtain this effect the distance from the listener to the side walls should be shorter than the distance to the front or rear walls or to the ceiling. Otherwise multipurpose modern halls are wider than longer and the sound is worse than a shoe box design.

The apparent source width of a sound source is defined as its perceived extent or size of the sound source's image, and is related to the concept of auditory spaciousness. Blauert [BL83] defines the auditory spaciousness as the perceived spatial extension of the sound source, extension of the region that the sound seems to come from and spatial extent of the auditory event. Figure 2.33 shows levels of apparent source width in two dimensions.

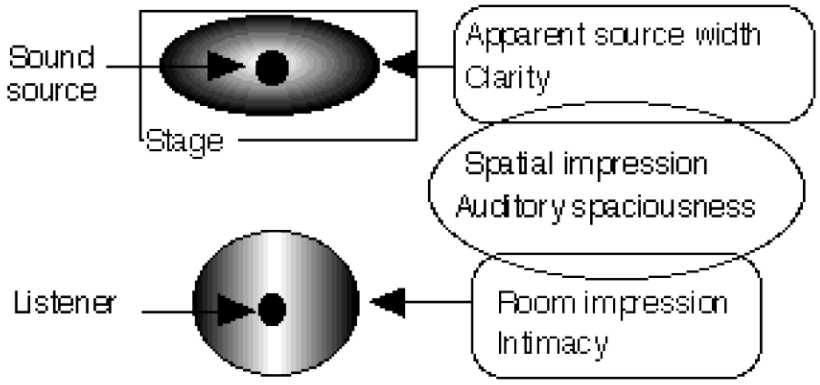


Figure 2.33 Perceived auditory spaciousness and apparent source width. [BEG00]

Due to the common dependence of early reflection temporal pattern, the percepts of spaciousness and apparent source width are often mixed with physical measurements of spatial impression. This is a physical measure of the relative interaural similarity of a room during the early reflection period from 0-80 msec. Spatial impression provides an impression of distance from the source [BL83]. Some authors refer to this as room impression [GRI93] or intimacy. Establish distinctive perceptual criteria in a concert hall study can be quite difficult. Sometimes match between physical and psychophysical parameter is not easy. Some have tried to simplify this limiting the numbers of possible perceptual criteria. Another example is the concept of clarity or c80. This is quantified by dividing the energy of the first 80 msec, the early reflections, by the reverb after this point, the late reflections. [GRI93]

3 - Approach
3.1 - Implementation of a 3D audio system

Begault [BEG00] claims that to obtain a successful implementation of a 3D audio system we have to get a balance between resources such as time and money and the ideal design of the system. The designer has to face these three issues before start.

What is the proposed use of sound within the application? How the sound will motivate or alert the user and if the sounds are representational of the visual events in virtual reality applications.

The use of sound within the application translates directly into the perceptual simulation requirements of the system. It would be very difficult to enable a person to localize elevation of a virtual source better than she is able to do in normal spatial hearing conditions, and it's known that the HRTFs of the user will yield better results than with non-individualized HRTFs.

Finally perceptual requirements must be balanced against available resources. A designer must envision the nature of the sound system integration requirements in terms of hardware and software

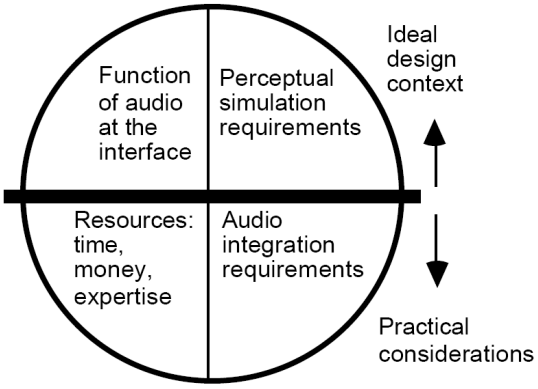


Figure 3.1 Requirements versus resource allocation chart for designing the system [BEG00]

The figure above summarizes the design considerations. At the top are the two areas that represent an ideal design context, at the bottom are the practical considerations that are certain to influence the outcome of any 3D sound implementation within a larger system. The challenge is to adapt to each of these areas into the design process, determine their priority and predict how big each slice of the pie chart should be.

The role of computing devices in this system can be segregated between the audio subsystem and the operating system. The operating system includes the computer's central processor which contains the virtual scenario generated by the reality engine software. The audio subsystem can be divided into hardware for signal sources and signal processors that are dedicated to 3D sound.

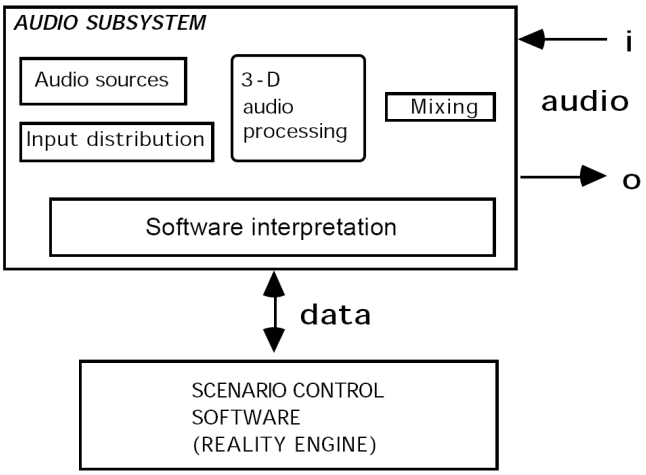


Figure 3.2 General system configurations in an integrated system [BEG00]

The overall function of the audio subsystem is to manage internally the input and the output of audio and control data to other audio components, and other tasks like to trigger signal sources from signal generators, to apply signal processing to various audio sources, to distribute and route internally or externally received audio signals and to combine and mix signals for multiple outputs to one set of left right outputs for each listener.

The audio subsystem responds to software calls from the scenario control system software. This system manages the input and output of multiple function effectors and sensors that affect both the visual and audio simulations. The scenario control system can include data from a head position tracker or the localization of a mouse. Data flows in the form of calls to and from the scenario control software to the audio subsystem software interpretation component.

For example, consider a scenario involving a sink in the kitchen of a virtual house. The scenario control software, contained within the main function may use a call to a separate function called play_wate_sound (x, y, z) which is activated when the user turns the virtual sink handle. The relative location of the sink to the listener is passed to the audio subsystem software interpreter with the variables (x, y, z). Then the software interpreter sends lower level commands within the audio subsystem to switch on the recording of the water sound, send the location of the sink to the 3D sound component, link the audio output of the signal source to the correct 3D sound signal processor, and mix the audio outputs with any other sounds that occur simultaneously.

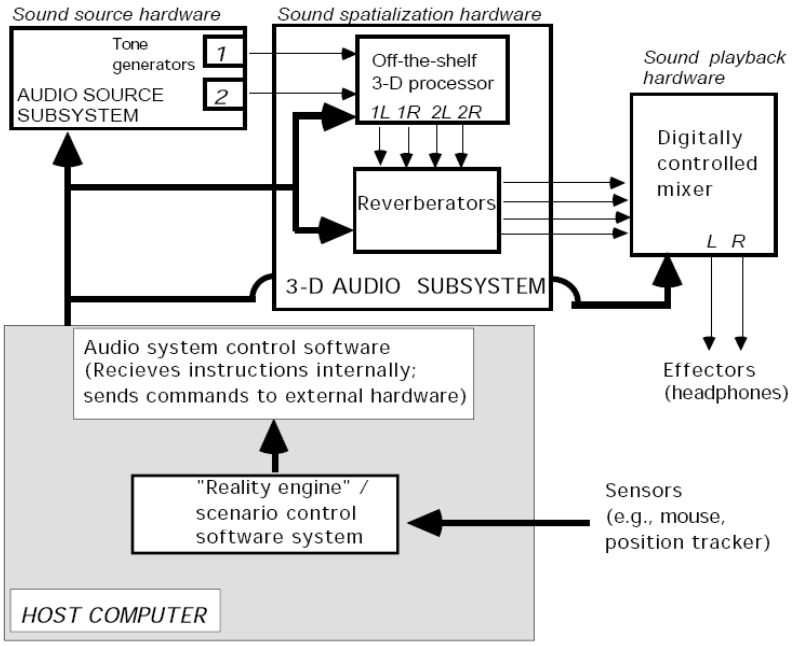


Figure 3.3 Distribute system using external devices for tone generation and spatialization. [BEG00]

In the figure above the data control signals are represented with thick lines and the audio signal is represented with thin lines. In this example the audio subsystem is separated from the computer. The scenario control system sends program calls to internal audio system software which in turn sends control signals to the external audio hardware. The multiple left and right channels are combined for output to the headphones.

The routing of all this information can be accomplished on any number of levels of technical sophistication. The one entire system could be integrated in a single chip or can be segregated into many devices.

There are two approaches to implement a 3D audio subsystem. The first involves assembling distributed, off-the-shelf components for spatial manipulation and components such as synthesizers and effect processors.
The second approach is to design an integrated system consisting o hardware and software contained within the same computing platform as the scenario system itself. This integration can mean involvement of additional personnel like specialist in digital processing and firmware hardware design that can make very expensive this type of systems. However the system cold be a hybrid approach between integrated and distributed elements, this layout is frequently used for practical reasons.

Once the system requirements are known, the perceptual requirements of the 3D sound system can be assigned. Any assessment of perceptual requirements will need to consider the issues listed below. [BEG00]

- Two important cues for manipulating a sound source's lateral position are interaural intensity differences (IID) and interaural time differences (ITD).

- The spectral changes caused by the HRTF are the cues that help externalize sound images, disambiguate front from rear and impart elevation information. The inclusion of the HRTF is fundamental for a 3D sound system.

- Changing HRTF spectra as a function of the head movement is generally considered to improve overall localization performance.

- Nonindividualized HRTFs result in a poorer localization than individualized HRTFs, but the latter are usually impractical for use in most applications. There are also notable localization performance differences between individuals

- The perception of virtual sound sources is significantly affected by the inclusion of reverberation. Reverberation enables externalization of images outside the head, allows a sense of the surrounding environmental context and is a cue to distance when used in terms of the R/D ratio. Reverberant sound also can cause deterioration in localization accuracy of azimuth and elevation.

- Reversals of virtual images, particularly front-back, occur for a significant proportion of the population. The problem appears more when using individualized HRTFs than when using nonindividualized HRTFs, and more with speech than with broadband noise. However this problem can be offset by previous expectation or associated visual cues.

- Distance perception is not particularly good even under spatial hearing conditions. A decision must be taken as to what cues to implement, loudness or intensity, R/D ratio and/or spectral changes caused by the air absorption. Intensity is the strongest cue; a decision needs to be taken as to whether it is implemented in terms of the inverse square law or a perceptual scale. In addition, distance can be implemented as an absolute cue or as a relative cue.

3.1.1 - DSPs for 3D simulation

Within a virtual reality or a multimedia application, 3D sound processing is almost always best accomplished digitally using audio DSP chips to manipulate signal sources. Audio DSP chips are designed to perform a limited of digital DSP instructions in real time. The number of instructions that can be performed is a function of the processor's clock rate. One can select one DSP that is optimized to perform a specific operation or digital filtering or there are some DSP that can be programmed to determine their operation. The 3D audio component of an integrated audio subsystem can contain anywhere from one to as many DSPs as can be accommodated. Each is usually assigned to a specific function. In the case where a 3D sound system is used to simulate the azimuth and elevation of two sounds sources within a reverberant room a single DSP can perform the filtering for both the right and the left pinna, for simulating elevation and azimuth. Two DSP are used to spatialize two sources to two different positions. Because both sound sources are in one room, the outputs of both positional DSPs can be fed to a common DSP that simulates stereo reverberation in a simple way. This is an example of a functional division between position simulation DSPs and environmental context situation DSPs.
Alternatively a hybrid system approach can be used, the position simulation DSPs are part of the integrated system and the environmental context simulation DSP is within an off-the-shelf reverberator.

Now we are going to see a rudimentary overview of the theory behind DSP algorithms used for implementation of 3D sound, particularly for implementation within an integrated system. [BEG00] [MO90]

A continuous analog waveform can be expressed as a function of continuous time *x(t)* and could be described with the parameters frequency, time and phase offset. In analog sound systems *x(t)* is transmitted via continuously varying voltage, a digital system samples sound to obtain a digital signal x[n] with discrete intervals. The DSP effects a modification on the input digital data and the operation performed by the DSP can be analyzed in terms of its impulse response symbolized by h[n]. To examine the digital output signal is used a specialized signal in the input called analytic impulse, which is a signal that doesn't vary along the time. This reveals the operation of the DSP in the time domain, but for filtering is more desirable to see the effect on the frequency domain. For that purpose we apply an FFT on the signal.

DSP algorithms are usually complex combinations of simple elements. The most basic operations are the multiplication of the signal by a value, which will increase the amplitude of the signal, a delay, which stores samples of the signal in a buffer to release them after an elapse of time in addition with the undelayed signal, or the addition of the signal with another signal.

The gain and/or delay systems shown can be used to impose simple lateralization cues on an input source. Varying the gain of a sound in a system we can manipulate and adjust the Interaural Intensity difference (IID) and we can use a delay for manipulating the Interaural time difference (ITD).

The basis of 3D sound processing lies in the imitation of the spatial cues present in natural spatial hearing. In this case the binaural HRTF imposes spatially dependent spectral modifications and time delays on an incoming wavefront to the hearing system. This is possible to imitate using a DSP as a digital filter. What filtering does is multiply the spectrum of two waveforms, which is equivalent to the convolution of the time domain representations of the waveforms. This is exactly what the HRTF in natural spatial hearing does, the spectra of the pinna is multiplied with the spectra of the sound source. Convolution can be thought as a time-indexed multiplication and summation operation performed on two numerical arrays. The process of the convolution can be directly implemented into a DSP algorithm for filtering in real-time signals by using techniques such as overlap-add and overlap-save. In the operation of convolution a finite impulse response (FIR) filter is the manager of the operation of outputting a sum of present and past inputs. If instead of summing the current input with a past in put value, we were to use one feedback loop and sum the input with a delayed version of the output we are using filter called infinite impulse response (IIR) filter. IIR filters are useful in the simulation of reverberation, because the feedback operation is analogous to the reflections in a room, but it's impossible to implement a binaural impulse response directly in an IIR filter. To imitate an HRTF in the frequency domain we will need combinations of FIR and IIR techniques.

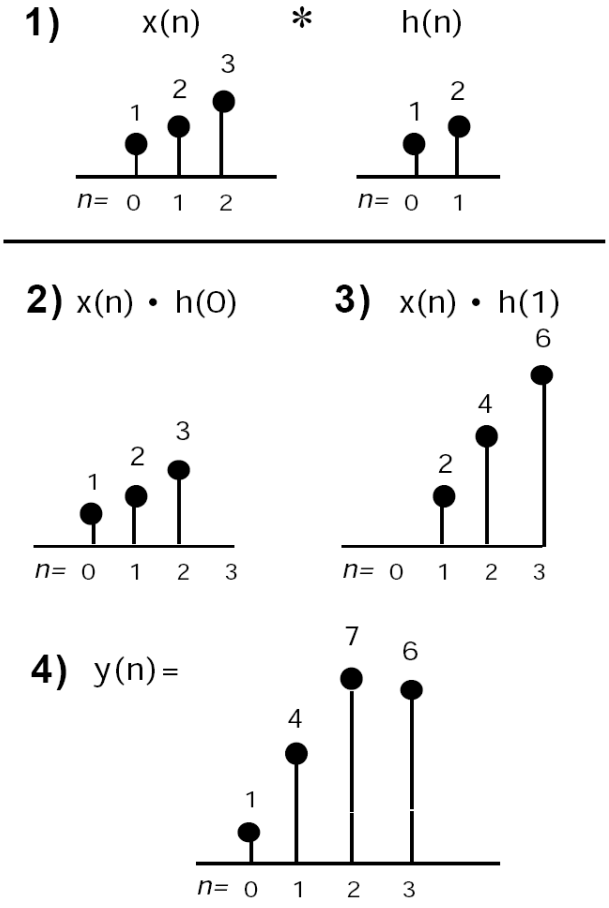


Figure 3.4 Convolution of two sequences in time domain [BEG00]

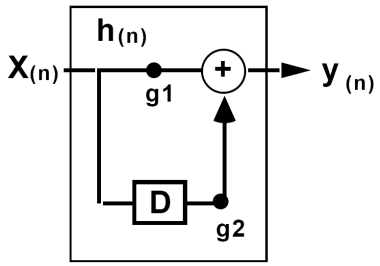


Figure 3.5 Configuration for a FIR filter, *Y(n)=g1*x(n)g2*x(n-1)* [BEG00]

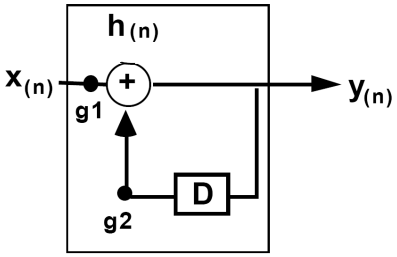


Figure 3.6 Configuration for an IIR filter, *Y(n)=g1*x(n)+g2*y(n-1)* [BEG00]

3.1.2 - Implementing HRTF

For hardware implementation purposes is often desirable to create versions of HRTF with shorter impulse response lengths. To perform this info reduction there are three basic methods.

The first is to simply reduce the sampling rate of the measured impulse to a lower value. The problem is that the most of the audio world is working with 44.1 kHz sampling rate.

The second is to perform on the signal a rectangular windowing. The simplest way to reduce HRTF data is to eliminate lower-amplitude portions at the start and end of the impulse responses keeping only the higher amplitude portions. A rectangular window, smaller than the original measurement, is applied to capture the essential features of the HRTF. The window must be applied in the same point for all measured HRTF to preserve the ITD information. The problem of this technique is the loss of low-frequency information. One commercial available 3D system will use this technique to allocate filtering resources between one and four sound sources. One source will be filtered by 521 taps, two by 256 and four by 128 points each.

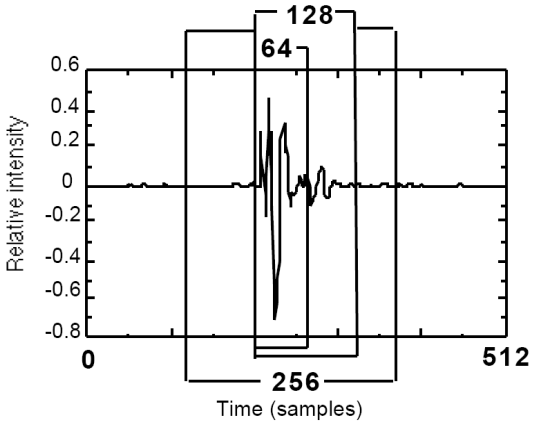


Figure 3.7 windowing of the HRTF impulse with 64, 128 and 156 samples [BEG00]

The third method is to use statistical or other analytic means to examine the measured HRTF magnitude responses and then create a shorter impulse response length filters that approximate the original HRTFs. This technique results in synthetic HRTFs and synthetic binaural impulse responses. The synthetic HRTF method involves taking a measured HRTF and producing data that meets both perceptual and engineering criteria. This technique has been studied by authors like Kistler and Wightman [WK92], Begault [BEG92] and Asano in [AS90].

Once the data is reduced is possible to implement this filters in a DSP.

One of the design considerations for any 3D sound system is to decide whether or not the simulation interface is constrained to a limited set of measured HRTF positions. The alternative is to allow arbitrary specification of azimuth and elevation via continuous controller or software interface.

Sound sources are not typically spatially static; there is usually either movement of the head, the sound source or both. Any system that includes head or sound source movement needs a finer resolution of spatial position that represented by the original measurements.

Because the filter coefficients require intermediate values in order to transition from one stored HRTF to another, the usual procedure to obtain a virtual moving sound source is to implement linear interpolation between measured values. This procedure is bound to the assumption than between HRTFs between measured positions would have spectral features that can be averaged via interpolation. However the result of an interpolation is not always satisfactory.[MO90]

3.1.3 - Implementing distance model

For distance an intensity scaling scheme usually is adopted where sounds are emitted at distances relative to the egocentric center, situated inside the head at the center point between the ears. Absolute perceived distances on the other hand are more difficult to predict in an application when the sound pressure level at the ears is an unknown quantity, the only option available is the digital storage and control of playback level. In absence of interaural attributes the system places the sound at the egocentric position by sending the same signal to both left and right channels.

In the system a single digital gain value applied to each output channel from a binaural HRTF filter pair can be used to create a simple virtual audio distance controller, according to the inverse square law and perceptual scales showed in figures 2.23 and 2.24.

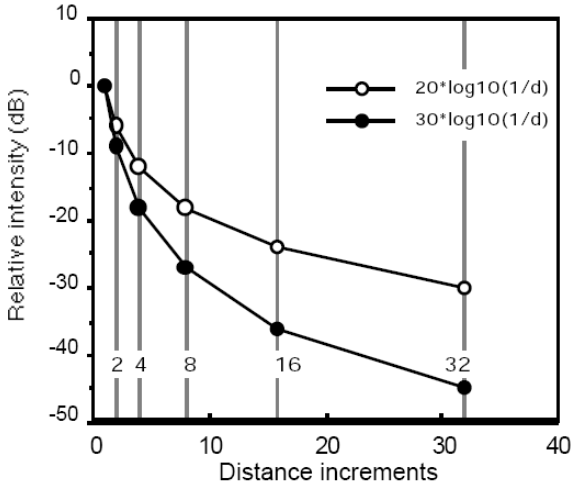


Figure 3.8 Intensity increment schemes for relative distance; the inverse square law (Open circles) and a 30*log10 is similar to the sone scale.
[BEG00]

Figure 3.8 shows a graph of db values that can be used for either perceptual or inverse square law scale for relative distance increments. The desired distance increment can be calibrated to a relative dB scale such that it is either equivalent to the inverse square law or to a perceptual scale approximating the reduction in sones.

As an example a distance increment of 0 would be reserved for the loudest two-channel monaural signal at egocentric center. An increment of 1 would indicate the minimum absolute distance from egocentric center caused by the use of the distance increment scheme which could be four inches from the egocentric center in the vicinity if the edge of the head. A distance increment of 2 would double this distance and so on out to the auditory horizon. These are not the only two distance increment schemes that could be used. In the ultimate 3D audio system different perceptual scale could be used according to the type of sound source input and adjusted to each individual's preference.

There is another important cue to distance, the relative ratio of reverberant to direct sound. The R/D ratio can be implemented by a simple mixing procedure, the reverberant field can be originated either from an integrated signal processing or in a distributed system from an off-the-shelf reverberator. It will be seen that the relationship between distance and environmental context cues based on intensity and reverberation, can all be addressed within a single paradigm. This paradigm involves the techniques of room modelling and auralization.

3.1.4 - Implementing reverberation model

The implementation of reverberant sound within a 3D sound system is desirable for many reasons besides distance simulation. A 3D sound system without reverberation only can reproduce simulations of sources heard in an anechoic environment. Natural spatial hearing normally occurs within a reverberant environment, the reproduction of a sound without reverberation can seem unrealistic.

As the digital processing is cheaper the number of methods and the complexity level for implementing synthetic reverberation increases. Some of these methods apply the HRTF not only to the direct sound but to the indirect field as well resulting in a synthetic spatial reverberation.

The impulse response of a measured room can be convolved with the sound to obtain the character and the directionally of reflected energy of the room where the impulse was measured. One can switch the virtual acoustic environmental context between a small hall, an auditorium or a cathedral. To perform this, the room impulse must be a good measurement of the room and there are several challenges to obtain a good impulse. To create spatial reverberation over a 3D sound, system, one would need to synthesize the spatial position of reflections if the room impulse response were measured with one or more microphones.

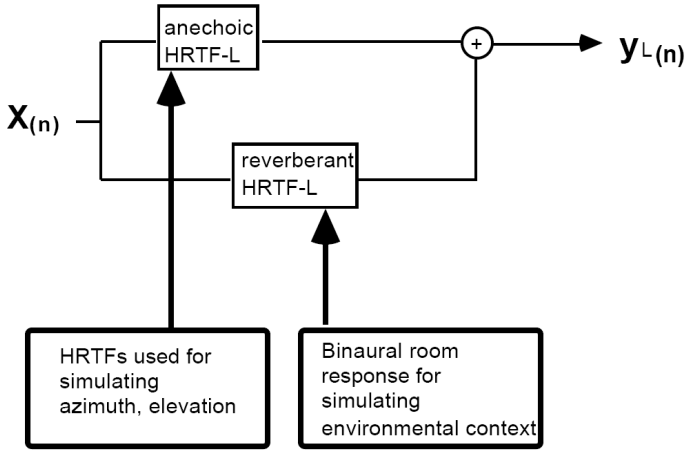


Figure 3.9 Spatial reverberation process applied to the left channel.
[BEG00]

In the figure above is showed a way of designing a 3D audio system with spatial reverberation. Anechoic HRTFs are taken from a library of measurements and are used to position the virtual sound source in elevation and azimuth. The binaural room response is used to stimulate the indirect sound field. At the end the direct and indirect fields are added obtaining this way the left channel. In this kind of design problems may occur if two different HRTF sets are used for the direct and indirect sound filtering and if the positional information imparted by the reverberation is held constant and the direct sound changes position.

The most accurate means to measure the impulse response of a room is binaurally using a recording dummy head. The impulses measured can be modified to create different effects. For instance the impulse response of a bright room can be low-pass filtered to create a synthetic warm room response. But in addition to the length of the FIR filter needed, other potential problems remain when using an actual binaural response for simulation, it is uncertain whether a match is necessary between the anechoic HRTFs and the pinna used in gathering the binaural room impulse response, and the position of the direct sound can change but the reverberant remains constant unless it is updated too. For this reasons a 3D sound system is better designed using one of the synthetic reverberation schemes discussed later.

For years, there were three methods to produce synthetic reverberation. One was placing a loudspeaker and a microphone in a reverberant room known as an echo chamber, the spring and plate reverberation methods that consist on sending the signal down one side of a spring or plate and then recording the signal at the other side and finally the use of tape head echo devices.

A plate reverberator consists of a rectangular sheet of steel, to which a moving-coil driver and one or two contact microphones are attached. Waves transduced through the plate by the driver are reflected back at its edges toward the contact microphones, the decay time is a function of the distance of the multiple reflections of the waves across the plate.

The spring reverb used in organs and guitar amplifiers uses magnetic coils instead of a driver and a contact mic and several springs of different length, density and number of turns. The reverberant effect is created when part of the audio signal bounce back and forth within the spring before dissipating into heat. In the case of the reverberant room the sound is played with a speaker at the end of a reverberant chamber and the sound is recorded at the other extreme of the chamber by a microphone.

One of the first electronic methods for producing a quasi-reverberant effect was to use tape head feedback with an analog tape system. On a professional analog tape recorder the erase, record and playback heads are separate, a small time delay results as a function of the tape's movement. By returning an attenuated version of the signal received at the playback head to the record head is possible to get repeated echo delays that suggest the delays heard in the real reverberation.

In the figure below is showed how the process is performed. The erase head (E) prepare the tape for the recording, the record head (R) record the signal and the delayed signal that comes from the playback head (H). The time delay corresponds to the time it takes the tape to travel from R to P.

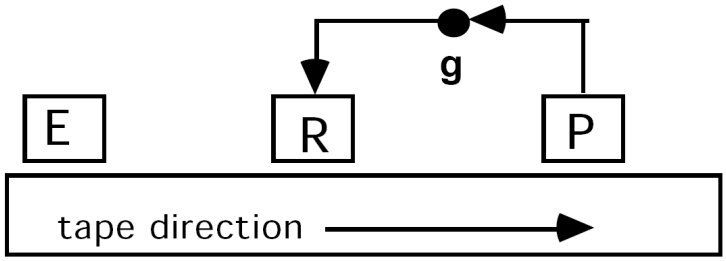


Figure 3.10 Tape head reverberator technique [BEG00]

This system sounds highly artificial because of the regularity of the delays, the lack of echo density and because the frequency response consist of a series of alternating peaks and valleys.

One of the first alternatives to this system was shown by Schroeder and Logan in 1961 [SL61]. Their solution was to use what is termed an allpass filter to design their reverberator. With this kind of filter it's possible to the filter to affect only the phase of the output signal, not its magnitude.

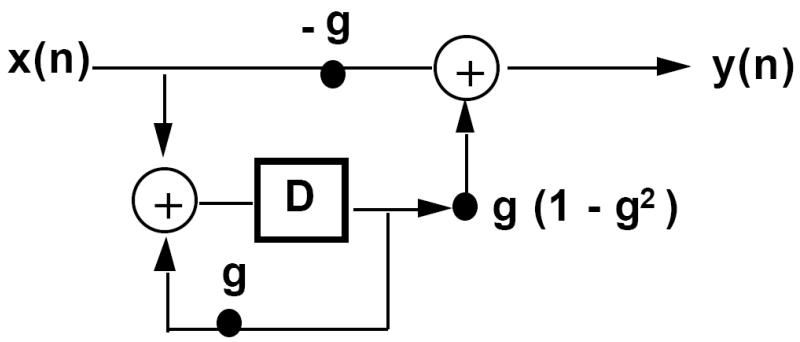


Figure 3.11 Allpass reverberator designed by Schroeder and Logan [SL61]

This means that X(n) will have the same magnitude than Y(n) but the phase response will differ. In effect, the group delay of various harmonics in a complex sound will be skewed in time, in a manner analogous to what happens to the phases of the waveforms bouncing back and forth between the surfaces of a room. As a result the echo response would become denser and less periodic. This technique is still imitated in many commercially available reverberators.

The most used reverberator at Stanford's CCRMA has been described by Sheeline [SE82] and Moore [MO90]. It consisted on algorithms that can be found in one form or another in commercial digital reverberators.

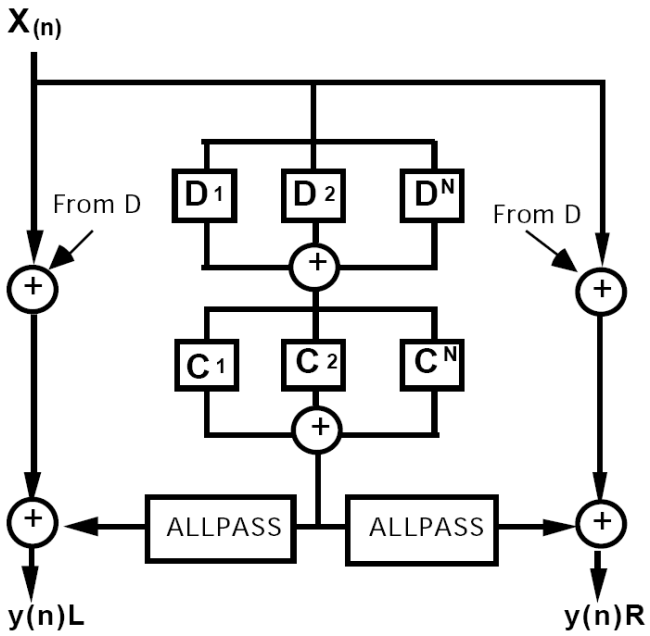


Figure 3.12 The Stanford CCRMA reverberator design [BEG00]

A copy of the direct signal is scaled at *g*, and then fed to the reverberation network. A delay line is used to simulate early reflections as a single delay; each value of *D* could be set to a particular time delay based on a model of early reflections. These tapped delays are fed directly to the output. The rest of the reverberator's design focuses on attaining a dense stereo reverberation in an inexpensive manner. The elements labelled *C1-C4* are low pass comb filters, the output is assigned to two or more separate allpass filters. A stereo effect is attained by setting delay and gain parameters for each channel slightly differently.

Convolving an input with noise that has been shaped with an exponentially-decaying amplitude envelope gives a very real reverberation effect if the noise decay matches that of an actual late reverberation response. This method is as computationally expensive as convolving with a real room impulse response but the technique has excellent sonic characteristics.

3.2 - Auralization

In natural spatial hearing, both direct and reverberant sound is altered by the listener's HRTF. In a virtual acoustic simulation when HRTF filtering is applied not only to the direct sound but also to the indirect sound the result is spatial reverberation. Although the binaural room response can capture this information it's more desirable to obtain a synthetic binaural impulse response characteristic of a real or simulated environmental context, this way that information can be processed in real time and enable a real time 3D sound simulation.

Auralization is the process of rendering audible, by physical or mathematical modelling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in a modelled space. Auralization involves the combination of room modelling programs and 3D sound processing methods to simulate the reverberant characteristics of a real modelled room acoustically. Auralization software/hardware packages will significantly advance the use of acoustical computer-aided design programs (CAD). The designer of the room will be able to listen to the effect the room on a sound in a determined position of the virtual room. This tool is very useful for the room designers when this room is going to be used for audio applications. [KL93]

There are other uses for auralization like the advanced reverberation for professional audio applications, interactive virtual acoustic environments and improvement of localization for virtual acoustic simulations.[BEG92]

The potential acceptance of auralization in the field of acoustical consulting is great because it represents a form of acoustic virtual reality that can be attained relatively inexpensively compared to visual worlds.

To obtain the auralization of one room first is necessary to represent a particular space through a room model modelled with a CAD program. A series of planar regions must be entered to the software that represents walls, doors, ceilings and other features of the environmental context. By selecting from a menu it's possible to specify one of several common architectural materials for each plane such as plaster, concrete wood…etc. Calculation of the frequency-dependent magnitude and phase transfer function of a surface made of a given material will vary according to the size of the surface and the angle of incidence of the waveform.

Once the software has been used to specify the details of a modelled room, sound sources may be placed in the model. After the room and the speaker parameters have been joined within a modelled environment, details about the listener can then be indicated, such as their number and position.

Prepared with a complete source-environmental context-listener model, a synthetic room impulse response can be obtained from the acoustical CAD program. A specific timing, amplitude, and spatial location for the direct sound and early reflections is obtained based on the ray tracing or image model techniques.

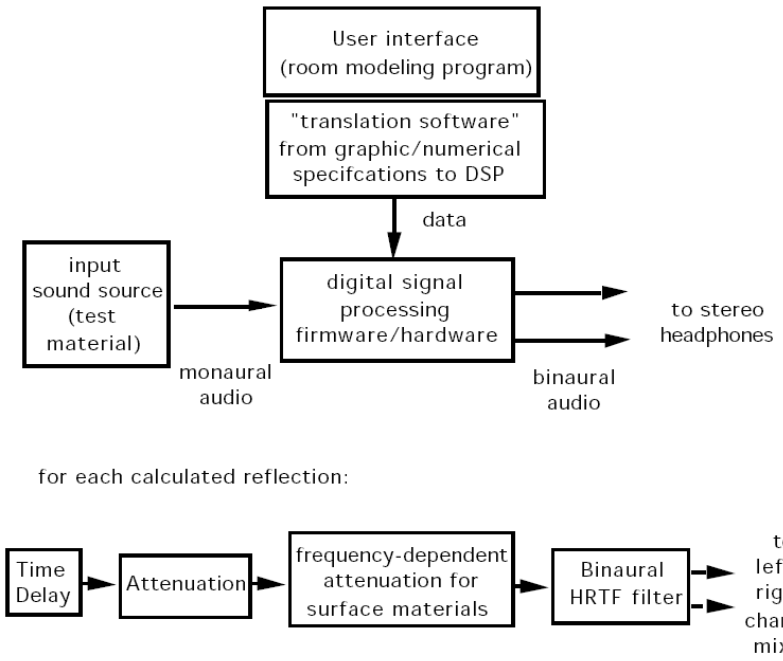


Figure 3.13 Basic components of an auralization system [BEG00]

In the figure above is shows the basic process involved in a computer-based auralization system intended for loudspeakers audition. A room modelling program is used to calculate the binaural room impulse response, based on the positions of the source and the listener and details of the environmental context. Note that here there are not any data about the azimuth or the elevation of the sound source explicitly; this data will result as a function of the specified model. The model also will include details about relative orientation and dispersion characteristics of the sound source, information on transfer functions of the room's surfaces, and data on the listener's location, orientation and HRTFs.

Here is explained how the program determine the scaling, time delay and filtering of the direct sound and each modelled early reflection. For each reflection a copy of the direct sound is obtained from a tapped delay line. The path of the reflection from source to listener results in a time delay that is a function of the speed of sound and an attenuation coefficient determined from the inverse square law. The frequency dependent absorption of the room's surfaces will depend on a complex transfer function that can be approximated by a filter. Finally, the angle of reflection relative to the listener's orientation is simulated by a HRTF filter pair. [BEG00]

3.3 - Binaural audio using loudspeakers

Once we have our 3D sound system implemented, we are going to see in this part of the thesis how to obtain the 3D sound illusion using loudspeakers instead of headphones. This approach requires far fewer transducers than a system that attempts to reconstruct a complex sound field within a volume of space. Existing loudspeakers systems that deliver binaural audio to a listener have the serious constraint that the listener may not move because otherwise the 3D illusion disappears. There are systems that track the listener and adjust the loudspeaker signals to maintain the binaural transmission.

The system that allows 3D audio trough speakers is created by combining a binaural synthesizer with a circuit that inverts the acoustic transmission path to the ears. The primary goal of the transmission path inversion is to eliminate crosstalk from each speaker to the opposite ear, and these circuits are called crosstalk cancellers. Although it's possible to merge the operation of the binaural synthesizer and crosstalk canceller into a single filter operation, there are many reasons, according to the studies of William Gardner [WG97a], for logically separate these operations.

- The binaural synthesizer and crosstalk canceller both require head models. It' possible to use different head models for each function and may each may be individualized or non-individualized.

- The head model used by the binaural synthesizer is intended to be perceptually correct, whereas the head model used by the binaural crosstalk canceller must be acoustically correct.

- Efficient implementations can results from suitable factorizations of the separate systems.

- Reverberation is properly handled by bypassing the binaural synthesizer and using only the crosstalk canceller.

- Headphone compatibility is easily achieved when the systems are separately implemented. The headphones can be driven from the output of the binaural synthesizer.

- When analyzing the performance of the total system, errors may be attributed to deficiencies in either the binaural synthesizer or crosstalk canceller.

For all these reasons is desirable to separately implement the binaural synthesizer and crosstalk canceller.

3.4 - Theory of crosstalk cancellation

Crosstalk cancellation is a technique for sending arbitrary, independent signals to the two ears of the listener from conventional stereo speakers. It involves cancelling the crosstalk that transits the head from each speaker it the opposite ear. The technique was first introduced by Bauer [BB61], put into practice by Schroeder and Atal [SA63], and later used by Schroeder to reproduce concert hall recordings for a comparative study. Essentially, the transfer functions from the two speakers to the two ears from a 2x2 system transfer matrix. To send arbitrary binaural signals to the ears require pre-filtering the signals with the inverse of this matrix before sending the signals to the speakers. The inverse filter, or the crosstalk canceller, is a two input, two output filter which Schroeder implemented using a lattice topology.

The filter functions were derived from head responses measured using a dummy head microphone. For the comparative study, binaural impulse responses of concerts halls were convolved with anechoic music to create binaural signals. These were filtered with the crosstalk canceller and presented to a listener seated in an anechoic chamber with loudspeakers at ± 22.5 degrees. The results of the experiment was described as amazing, cause listeners could perceive sound originating from all directions around them but head rotations of ± 10º were sufficient to ruin the spatial illusion.

The usual method of creating crosstalk cancelling filters is to invert head responses obtained by direct measurement or modelling. Damaske [DA71] described an alternative method whereby the cancellation filters were specified through a calibration procedure adjusted by the listener. The results specified a 90º filter which was superposed with the mirror 90º filter to build a symmetric crosstalk canceller. The results of this crosstalk filter showed excellent localization for all azimuths in the horizontal plane and the vertical localization was also good. Cooper and Bauck [CB89] simplified the crosstalk canceller by exploiting the symmetry of the listening situation. This yields a crosstalk canceller implemented with only two elementary filters, one that operates on the sum of the left and right binaural inputs (L+R), and one that operates on the difference of the inputs (L-R). This topology is called Shuffler. This technique only requires two filters for its implementation.

Several other filter topologies for implementing symmetric crosstalk cancellers have been proposed. Iwahara and Mori [IM78] described a recursive circuit where each output channel is fed back to the opposite channel via a filter consisting of the ratio of the contralateral to ipsilateral HRTFs. Other authors have described crosstalk cancellers based on the HRTFs measured from humans or dummy head microphones [ML89]. Koring and Schmitz described crosstalk cancellers based on individualized HRTF measurements and implemented using a lattice filter topology. The authors noted exceptionally high fidelity reproduction of binaural recordings.

A different approach to multichannel crosstalk cancellation, used by Iwahara and Mori [IM78] is to partition the set of speakers into pairs, each speaker pair becomes a separate 2x2 crosstalk canceller.

Now we are going to examine the theory of crosstalk cancellation, including a review of previous work according to the studies of William Gardner [WG97a]. Many of the equations deal with linear systems and the transfer functions that relate input and output signals. In order that convolution be expressed as a multiplication these equations are expressed in the frequency domain, and for simplicity the frequency variables are omitted wherever possible. Unless otherwise stated, all signals are frequency domain representations of their time domain counterparts. Scalar signals are notated in lower case, transfer functions in upper case. Vectors and matrices are both notated using boldface, vectors in lower case, and matrices in upper case.

Binaural synthesis is accomplished by convolving an input signal with a pair of HRTFs:

$$\mathbf{x} = \mathbf{h}\,x \tag{1}$$

$$\mathbf{x} = \begin{bmatrix} x_L \\ x_R \end{bmatrix},\ \mathbf{h} = \begin{bmatrix} H_L \\ H_R \end{bmatrix} \tag{2}$$

Where x is the input signal, **x** is a column vector of binaural signals, and **h** is a column vector of synthesis HRTFs. This is a general specification of the binaural synthesis procedure; there are many efficient ways to implement the synthesis filters. We call the vector **x** a binaural signal because it would be suitable for headphone listening, perhaps with some additional equalization applied.

The binaural signal may be a sum of multiple input sounds rendered at different locations:

$$\mathbf{x} = \sum_{i=1}^{N} h_i x_i \tag{3}$$

Where $\mathbf{h}_i$ is the HRTF vector for source x_i. The figure below shows the circuit that implements the multiple source binaural synthesizer. For simplicity, in the ensuing discussion the binaural synthesis procedure will be specified for a single source only.

When the binaural signal is being reproduced, rather than synthesized the individual signals will have been recorded with spatial cues encoded, in which case the synthesis HRTFs have been already applied. Using a pre-recorded binaural signal constrains the subsequent processing that can be done because it is not possible to manipulate the individual synthesis HRTFs without first performing a complicated unmixing procedure.

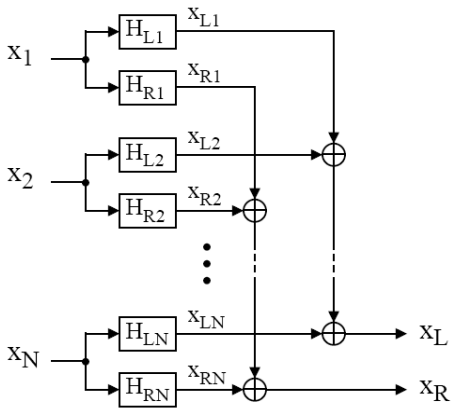


Figure 3.14 Multiple source binaural synthesizer. [WG97a]

In order to deliver the binaural signal over loudspeakers, it's necessary to filter it properly with a 2x2 matrix **C** of transfer functions:

$$\mathbf{y} = \mathbf{C}\mathbf{x} \tag{4}$$

$$\mathbf{y} = \begin{bmatrix} y_L \\ y_R \end{bmatrix}, \ \mathbf{C} = \begin{bmatrix} C_{11} C_{12} \\ C_{21} C_{22} \end{bmatrix} \tag{5}$$

We will call the vector of loudspeaker signals **y** a loudspeaker binaural signal and the filter **C** the crosstalk canceller. Because much of our discussion will concern different implementations of the crosstalk canceller, we have chosen **x** and **y** to be the input and output variables. The standard two channel listening situation is described in the figure below.

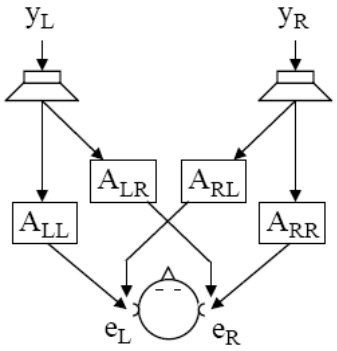


Figure 3.15 Acoustic transfer functions between the two loudspeakers and the listener's ears [WG97a]

The ear signals are related to the speaker signals through the equation:

$$\mathbf{e} = \mathbf{A}\mathbf{y} \tag{6}$$

$$\mathbf{e} = \begin{bmatrix} e_L \\ e_R \end{bmatrix}, \ \mathbf{A} = \begin{bmatrix} A_{LL} A_{RL} \\ A_{LR} A_{RR} \end{bmatrix} \tag{7}$$

Where **e** is a column vector of ear signals, **A** is the acoustical transfer matrix, and **y** is a column vector of speaker signals. The ear signals are considered to be measured in an ideal transducer somewhere in the ear canal such that all direction-dependent features of the head response are captured. The functions A_{XY} give the transfer function from speaker $X \in \{L, R\}$ to ear $Y \in \{L, R\}$ and include the speaker frequency response, air propagation, and head response. **A** can be factored as follows:

$$\mathbf{A} = \mathbf{HS} \tag{8}$$

$$\mathbf{H} = \begin{bmatrix} H_{LL} & H_{RL} \\ H_{LR} & H_{RR} \end{bmatrix}, \ \mathbf{S} = \begin{bmatrix} S_L A_L & 0 \\ 0 & S_R A_R \end{bmatrix} \tag{9}$$

H is the head transfer matrix which is a matrix of HRTFs normalized with respect to the free-field response at the center of the head, with no head present. The measurement point of the HRTFs and hence the definition of the ear signals **e**, is left unspecified to simplify the discussion. **S** is the speaker and air transfer matrix which is a diagonal matrix that accounts for the frequency response of the speakers and the air propagation to the listener. S_X is the frequency response of speaker X and A_X is the transfer function of the air propagation from speaker X to the center of the head with no head present. A simplify assumption is that each speaker response S_X affects the ipsilateral and contralateral ears equally.

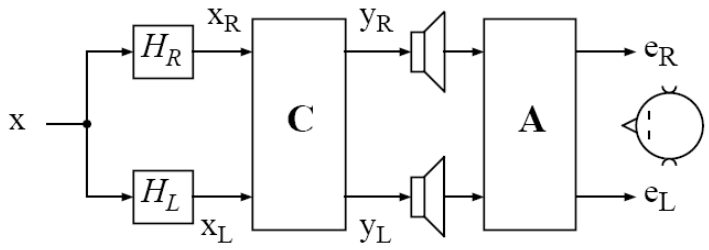


Figure 3.16 Schematic of playback including binaural synthesizer, crosstalk canceller and acoustic transfer to the listener. [WG97a]

In order to deliver the binaural signals to the ears, the crosstalk canceller **C** is chosen to be the inverse of the acoustical transfer matrix.

$$\mathbf{C} = \mathbf{A}^{-1} = \mathbf{S}^{-1}\mathbf{H}^{-1} \tag{10}$$

This implements the transmission path inversion. $\mathbf{H}^{-1}$ is the inverse head transfer matrix, later discussed in detail. $\mathbf{S}^{-1}$ associates an inverse filter with each speaker output:

$$\mathbf{S}^{-1} = \begin{bmatrix} 1/(S_L A_L) & 0 \\ 0 & 1/(S_R A_R) \end{bmatrix} \tag{11}$$

The $1/S_X$ terms invert the speaker frequency responses and the $1/A_X$ terms invert the air propagation. In practice, this equalization stage may be omitted if the listener is equidistant from two well-matched, high quality loudspeakers. However, when the listener is off axis, it's necessary to delay and attenuate the closer loudspeaker so that the signals from the two loudspeakers arrive simultaneously at the listener and with equal amplitude. This signal alignment is accomplished by the $1/A_X$ term above.

In a realtime implementation, it is necessary to cascade the crosstalk canceller with enough modelling delay to create a causal system. Adding a discrete-time modelling delay of m samples to the equation 10 we obtain.

$$\mathbf{C}(z) = z^{-m}\mathbf{S}^{-1}(z)\mathbf{H}^{-1}(z) \tag{12}$$

The amount of modelling delay needed will depend on the particular implementation. In order to simplify the following discussion, we will omit the modelling delay and the speaker equalization $\mathbf{S}^{-1}$ terms, and consider only the inverse head transfer matrix. Thus while we recognize that the first equation 10 is the general solution, we will consider crosstalk cancellers of the form.

$$\mathbf{C} = \mathbf{H}^{-1} \tag{13}$$

We will use the general form of equation 10 whenever the complete playback system is discussed. The inverse head transfer matrix is:

$$\mathbf{H}^{-1} = \begin{bmatrix} H_{RR} & -H_{RL} \\ -H_{LR} & H_{LL} \end{bmatrix} \frac{1}{D} \tag{14}$$

$$D = H_{LL}H_{RR} - H_{LR}H_{RL} \tag{15}$$

Where D is the determinant of the matrix **H**. The inverse determinant $1/D$ is common to all terms and determines the stability of the inverse filter. However; because it is a common factor, it only affects the overall equalization and does not affect crosstalk cancellation. When the determinant is 0 at any frequency, the head transfer matrix is singular and the inverse matrix is undefined.

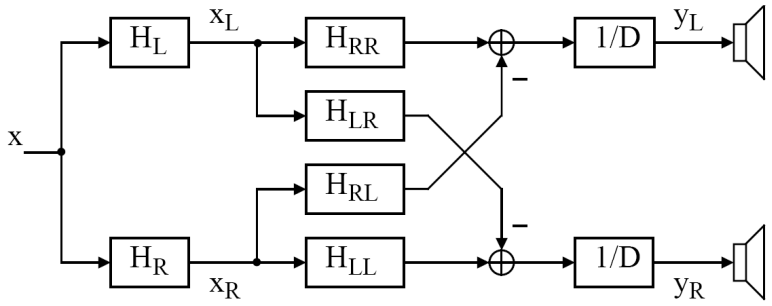


Figure 3.17 Single source binaural synthesizer cascaded with crosstalk cancellation filter [SA63]

Figure 3.17 shows the schematic of a single source binaural synthesizer and the crosstalk canceller of equation 14. This flow diagram was described by Schroeder and Atal in 1963 [SA63]. In their implementations the inverse determinant filter was commuted to the input of the circuit before the binaural synthesis stage.

Dividing numerator and denominator by $H_{LL}H_{RR}$ equation 15 can be rewritten as [ML92] :

$$\mathbf{H}^{-1} = \begin{bmatrix} 1/H_{LL} & 0 \\ 0 & 1/H_{RR} \end{bmatrix} \begin{bmatrix} 1 & -ITF_R \\ -ITF_L & 1 \end{bmatrix} \frac{1}{1 - ITF_L ITF_R} \tag{16}$$

Where:

$$ITF_L = \frac{H_{LR}}{H_{LL}}, ITF_R = \frac{H_{RL}}{H_{RR}} \tag{17}$$

are the interaural transfer functions (ITFs). An examination of equation 16 reveals much about the crosstalk cancellation process. Crosstalk cancellation is effected by the $-ITF$ terms in the off-diagonal positions of the right-hand matrix. These terms predict the crosstalk and send an out-of-phase cancellation signal into the opposite channel. For instance, the right input signal is convolved with ITF_R which predicts the crosstalk that will reach the left ear, and the result is subtracted from the left output signal. The common term $1/(1 - ITF_L ITF_R)$ compensates for higher-order crosstalks, in other words the fact that each crosstalk cancellation signal itself transits to the opposite ear and must be cancelled. It is a power series in the product of the left and right interaural transfer functions, which explains why both ear signals require the same equalization signal: both ears receive the same high-order crosstalks. Because crosstalk is more significant at lower frequencies, this term is essentially a bass boost. The left-hand diagonal matrix, which we call ipsilateral equalization, associates the ipsilateral equalization, associates the ipsilateral inverse filter $1/H_{LL}$ with the left output and $1/H_{RR}$ with the right output. These are essentially high frequency spectral equalizers that facilitate the perception of rear sources using frontal loudspeakers. The use of the ITF to predict crosstalk at the contralateral ear requires that each output be equalized with respect to ipsilateral incidence. The ipsilateral equalization filters also compensate for any asymmetries in path lengths from speakers to ears when the head is rotated.

Using equation 16, the transfer functions for the circuit of figure 3.17 can be written as [ML92] :

$$\begin{bmatrix} y_L / x \\ y_R / x \end{bmatrix} = \begin{bmatrix} \left(\frac{H_L}{H_{LL}}\right) - \left(\frac{H_R}{H_{LL}}\right) ITF_R \\ \left(\frac{H_R}{H_{RR}}\right) - \left(\frac{H_L}{H_{RR}}\right) ITF_L \end{bmatrix} \frac{1}{1 - ITF_L ITF_R} \tag{18}$$

An examination of equation 18 reveals that it is composed entirely of ratios of HRTFs which correspond to either ITFs or free field equalized HRTFs. This is an important point, because it means that any factor common to the HRTFs will cancel. Thus, the HRTFs can be measured at any location within the ear canal or at the entrance of the blocked ear canal. Similarly, the HRTFs may be free field or diffuse field equalized. All of these possibilities yield the same position. The only constraint is that the HRTFs used for the binaural synthesizer be equalized the same as the HRTFs used for the crosstalk canceller.

In practice, the listener's HRTFs may not be exactly equal to the head model used by the crosstalk canceller. In this case, the condition for perfect crosstalk cancellation is that the matrix $\mathbf{HM}^{-1}$ be diagonal, where **H** is the true head transfer matrix, and **M** is an analogous matrix of model head transfer functions. The matrix $\mathbf{HM}^{-1}$ is diagonal when

$$\frac{H_{RL}}{H_{LL}} = \frac{M_{RL}}{M_{LL}}$$

and (19)

$$\frac{H_{LR}}{H_{RR}} = \frac{M_{LR}}{M_{RR}}$$

These ratios are not ITFs and they don't have an intuitive physical interpretation.

The matrix **H** is invertible if and only if it is non-singular, i.e. if its determinant $D \neq 0$. Because H is a function of frequency, it's possible that the inverse matrix $\mathbf{H}^{-1}$ exist only for particular frequency ranges where the matrix **H** is non singular. Similarly, if the matrix H is poorly conditioned at some frequency, this will lead to a small value of D, and the magnitude of $1/D$, and in these frequency ranges the inverse matrix only approximates the true inverse.

The form of the inverse matrix given in the equation 16 is obtained by dividing by $H_{LL}H_{RR}$. Thus, an additional constraint for the existence of this form is that $H_{LL} \neq 0$ and $H_{RR} \neq 0$. The inverse ipsilateral filters and the interaural transfer functions both depend on this constraint. As before, we may limit the magnitude of $1/H_{LL}$ and $1/H_{RR}$ in order to obtain approximate inverses in the frequency ranges where the magnitudes of the ipsilateral transfer functions are small.

For the present discussion we will also assume that the crosstalk canceller is a linear, time-invariant (LTI) system. We consider LTI systems whose z-transforms can be expressed as rational polynomials, which correspond to systems expressed as linear, constant-coefficient, difference equations [OP89].

The crosstalk canceller can be implemented using a network of sample delays, constant gains, and summing junctions. If the network contains no feedback loops, then is guaranteed to be realizable. This means that each set of output samples can be computed from the set of input samples and the state of the internal delays. The system will also be stable. The stability and realizability of the network are only issues when the network contains feedback loops. A simple feedback loop is shown in figure 3.18 and it has the following z-transform:

$$H(z) = \frac{1}{1 - A(z)} = 1 + A(z) + A^2(z) + \ldots \tag{20}$$

For the system in figure to be realizable, the feedback loop must contain at least one sample delay, otherwise it is impossible to compute the current output. This means that $A(z)$, expressed as polynomial in z^{-1}, must contain a common factor of z^{-1}. Referring back to the crosstalk cancellation solution in equation 16, if the term $1/1(1 - ITF_L ITF_R)$ is implemented using a feedback loop, then this will be realizable if the cascade of the two ITFs contains at least one sample of delay. Assuming an ITF can be modelled as a causal filter cascaded with a delay, then the condition for realizability is that the sum of the two interaural time delays be greater than zero.

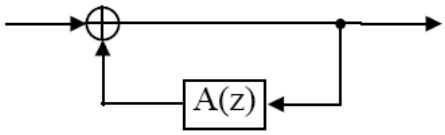


Figure 3.18 Feedback loop [WG97a]

$$ITD_L + ITD_R > 0 \tag{21}$$

The ITD is positive for positive incident angles, and increases monotonically with increasing lateral angle of incidence. It's easy to see that the realizability constraint of equation 21 is met when the listener is facing toward. The figure below shows the standard listening situation when the listener's head is rotated θ_h degrees right. θ_L and θ_R give the incident angles for ITF_L and ITF_R, respectively. When the listener is facing between the speakers both incident angles are positive, therefore both interaural time delays are also positive, and the realizability constraint is easily met. When the head is rotated just beyond a speaker, the ITD for that side becomes negative, while the opposite side ITD stays positive, and because of the monotonicity property, the sum of the ITDs stays positive. According to a spherical head model for ITDs, the ITDs become equal and opposite in sign when the head is oriented at ±90º. Thus, the realizability constraint of equation 21 is met when $-90 < \theta_h < 90$.

In the figure 3.19 θ_L and θ_R are incident angles of left and right speakers, respectively and θ_h is angle of head. In this example, all three angles are positive.

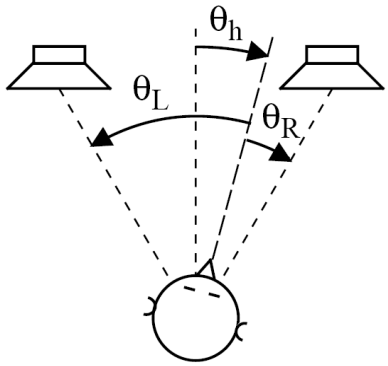


Figure 3.19 Incident angles of speakers for rotated head. [WG97a]

A necessary condition for stability of the crosstalk canceller is that all poles of the system's z-transform have magnitude less than 1. The region of convergence (ROC) then includes the unit circle, from which it follows that the system impulse response is absolutely summable, and therefore the system is stable in the bounded-input bounded-output (BIBO) sense [OP89]. An equivalent condition for stability is that the gain of all feedback loop be less than 1 for all the frequencies. For example consider again the simple feedback loop in figure whose z-transform is given in equation 20. The geometric series will converge if and only if $\left|A(e^{j\omega})\right|$ <1 for all ω.

Applying this constraint to equation 16 the crosstalk canceller will be stable if and only if

$$\left|ITF_L(e^{j\omega})\right|\left|ITF_R(e^{j\omega})\right| < 1, \forall\omega \tag{22}$$

The ITF describes head shadowing, and for positive incident angles and low frequencies the ITF is basically a lowpass filter which rolls off with increasing frequency. Furthermore, at low frequencies and small incident angles, the ITF magnitude decreases monotonically with increasing lateral angle of incidence. For a symmetrical head model, ITFs at negative incident angle are the inverses of the corresponding positive incident ITF, and are highpass filters. At high frequencies the magnitude of the ITF should be greater than 1, even at positive incident angles, because of notches in the ipsilateral response. Thus to ensure a stable crosstalk canceller it may be necessary to either limit the gain of the ITF model, or to use a band limited ITF model, the latter being the approach we will take. Considering then only the low-frequency portion of the ITF, we find that the constraint in equation 22 is met for frontal head orientations. When the listener is facing between the speakers, both incident angles are positive, and both ITFs are lowpass filters. When the head rotates just beyond a speaker, the ITF for that side becomes highpass, but the opposite side ITF is still lowpass, and because of the monotonicity property, the product of the ITFs will still have magnitude less than 1. According to a spherical model for ITFs, the ITFs in equation 22 become reciprocals when the head rotates to ± 90º. Thus the stability constraint of equation 22 is met for low frequencies when -90< θ_h <90.

A simpler way to reach this result is to consider the head transfer matrix **H** for a spherical head model. When the head is rotated to ±90º, both speakers fall in the same “Cone of confusion” the columns of the matrix **H** become equal, and **H** therefore becomes singular and non-invertible. We expect that a real head model will behave similarly at low frequencies, for instance that **H** will become singular, or at least ill-conditioned for head orientations near ±90º.

Note that when the head is rotated ±90º, to face the rear, both the realizability and stability constraints can be met if the left and right output channels are swapped. This corresponds exactly to implementing a crosstalk cancellation system using a pair of rear loudspeakers.

The straightforward way to implement the 2x2 inverse matrix of equation 14 is using four feedforward filters. Two recursive filter topologies which can also implement the inverse matrix are shown in figure below. The symmetric form of the figure 3.20a has been used by Iwahara and Mori [IM78] to implement crosstalk cancellers. These recursive topologies are also commonly used to implement adaptive filters for blind source separation.
The systems equations for the topology in the figure 3.20a are:

$$y_L = Ax_L + Cy_R \qquad (23)$$
$$y_R = Dx_R + By_L$$

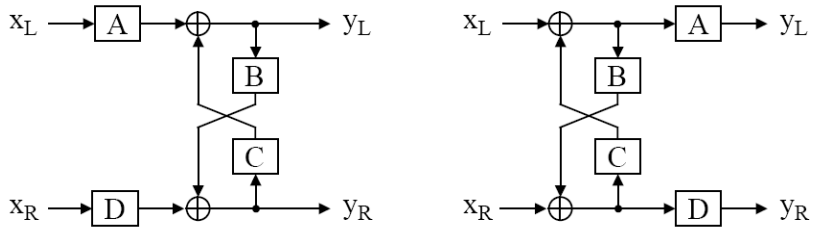


Figure 3.20 Recursive topologies for implementing the 2x2 inverse matrix. [IM78]

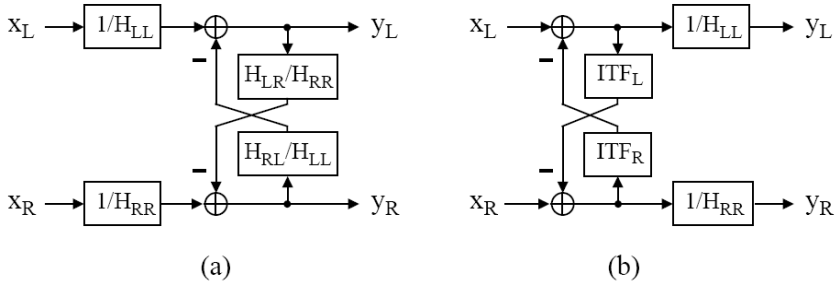


Figure 3.21 Recursive implementations of the asymmetric crosstalk cancellation filter. [IM78]

The coefficients in equation 23 can be solved to satisfy equation 16. The solutions are:

$$
\begin{aligned}
A &= \frac{1}{H_{LL}} \\
B &= -\left(\frac{H_{LR}}{H_{RR}}\right) \\
C &= -\left(\frac{H_{RL}}{H_{LL}}\right) \\
D &= \frac{1}{H_{RR}}
\end{aligned}
\tag{24}
$$

The implementation is showed in figure 3.21 the cross-coupled feedback filters are the HRTF ratios encountered in equation 19 and the feedforward filters are the inverse ipsilateral responses.

The system equations for the topology in figure 3.20 b are:

$$
\begin{aligned}
y_L &= A\left(x_L + \frac{C}{D} y_R\right) \\
y_R &= D\left(x_R + \frac{B}{A} y_L\right)
\end{aligned}
\tag{25}
$$

The coefficients in equation 25 can be solved to satisfy equation 16, the solutions are:

$$
\begin{aligned}
A &= \frac{1}{H_{LL}} \\
B &= -\left(\frac{H_{LR}}{H_{LL}}\right) \\
C &= -\left(\frac{H_{RL}}{H_{RR}}\right) \\
D &= -\left(\frac{1}{H_{RR}}\right)
\end{aligned}
\tag{26}
$$

This implementation is shown in figure 3.21b. The cross coupled feedback filters are ITFs. Although both implementations are mathematically equivalent, figure 3.21b is far more intuitive. As described earlier, convolving either channel with the appropriate ITF predicts the crosstalk that will reach the contralateral ear. The crosstalk is then cancelled by feeding the negative of this predicted signal into the opposite channel. An important feature of this circuit is that it feeds the cancellation signal back to the opposite channel's input rather than its output, and thus higher-order crosstalks are automatically cancelled. Finally, each channel output is equalized with the corresponding inverse ipsilateral response.

Most of the implementations discussed in the literature assume a symmetric listening situation. Obviously, the symmetric solution is simply a particular case o the general solution, but consideration of symmetry can lead to simplified implementations. When the listening situation is symmetric, we define:

$$H_i = H_{LL} = H_{RR}, H_c = H_{LR} = H_{RL} \tag{27}$$

Where H_i is the ipsilateral transfer function, and H_c is the contralateral transfer function. Substituting the symmetric variables into equation 15 we obtain:

$$H^{-1} = \begin{bmatrix} H_i & -H_c \\ -H_c & H_i \end{bmatrix} \frac{1}{H_i^2 - H_c^2} \tag{28}$$

Dividing by H_i^2, we obtain:

$$H^{-1} = \begin{bmatrix} 1 & -ITF \\ -ITF & 1 \end{bmatrix} \frac{1/H_i}{1 - ITF^2} \tag{29}$$

Where

$$ITF = \frac{H_C}{H_i} \tag{30}$$

is the natural transfer function for the symmetrical situation. This symmetric formula was described by Schroeder [SC73]. The corresponding flow diagram is shown in the figure 3.22.

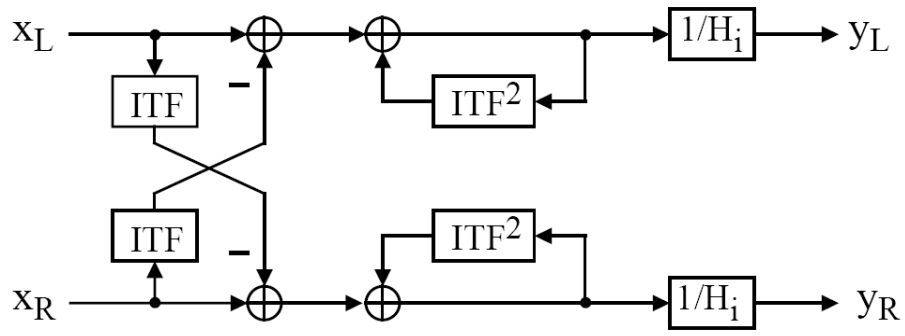


Figure 3.22 implementation of symmetric crosstalk cancellation filter [SC73]

Cooper and Bauck [CB89] later proposed using a “shuffler” implementation of the crosstalk canceller, which involves forming the sum and difference of the binaural inputs, filtering these signals and then undoing the sum and difference operation. The generic shuffler filter circuit is shown in figure. The sum and difference operation is accomplished by the unitary matrix U below, called a shuffler matrix:

$$U = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \frac{1}{\sqrt{2}} \tag{31}$$

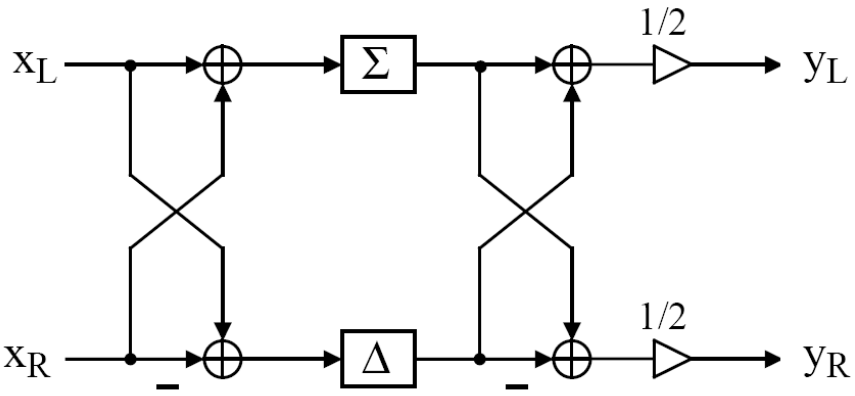


Figure 3.23 Shuffler filter structure. This has been used for implementing crosstalk cancellers [CB89]

Columns of the matrix **U** are eigenvectors of the symmetric 2x2 matrix, and therefore the shuffler matrix **U** diagonalizes the symmetric matrix $\mathbf{H}^{-1}$ via a similarity transformation:

$$\mathbf{H}^{-1} = \mathbf{U}^{-1} \begin{bmatrix} \frac{1}{H_i + H_c} & 0 \\ 0 & \frac{1}{H_i - H_c} \end{bmatrix} \mathbf{U} \tag{32}$$

Thus the crosstalk canceller is implemented with shuffler filters Σ and Δ that are the inverses of the sum and difference of the ipsilateral and contralateral responses [CB89].

$$\begin{aligned} \Sigma &= 1/(H_i + H_c) \\ \Delta &= 1/(H_i - H_c) \end{aligned} \tag{33}$$

The shuffler topology is shown in figure 3.23. The $1/\sqrt{2}$ normalizing gains have been commuted to a single gain of ½ for each channel. Note that $\mathbf{U}=\mathbf{U}^{-1}$, so the same sum and difference operation appears on both sides of the Σ and Δ filters.

A further simplification to equation 32 can be made by factoring out $1/H_i$, which yields:

$$\mathbf{H}^{-1} = \mathbf{U}^{-1} \begin{bmatrix} \frac{1}{1+ITF} & 0 \\ 0 & \frac{1}{1-ITF} \end{bmatrix} \mathbf{U} \frac{1}{H_i} \tag{34}$$

This formulation has been suggested by Jot [JO92] and subsequently by Gardner [WG95]. The ITF can be modelled as an interaural time delay cascaded with a lowpass head-shadowing filter. The shuffler filters are then seen to be simple comb filters with low-pass filters in the feedback loops, with the following transfer functions:

$$\begin{aligned} \Sigma &= \frac{1}{1+ITF} \\ \Delta &= \frac{1}{1-ITF} \end{aligned} \tag{35}$$

In practice, the inverse ipsilateral response in equation 34 can be commuted back to the binaural synthesis stage by using synthesis HRTFs which are free-field equalized with respect to the loudspeaker direction.

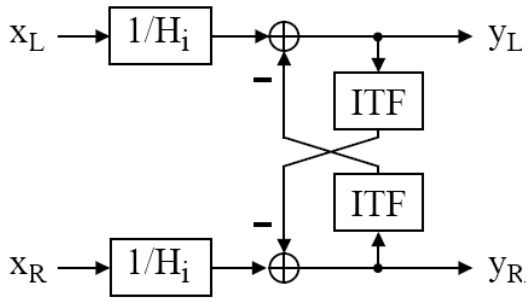


Figure 3.24 Symmetric recursive structure [IM78]

The recursive structures in figure 3.21 can of course be used for the symmetric solution, and this has been described bi Iwahara and Mori [IM78]. When the system is symmetric both feedback filters become the ITF, and the inverse ipsilateral filter can be associated with either the inputs or outputs of the system. In a symmetric implementation, it always makes sense to commute the inverse ipsilateral filter to the binaural synthesis filters by using a free field equalized HRTFs. Figure 3.24 shows this symmetric recursive structure.

For the symmetric case the condition for crosstalk cancellation analogous to the constraint in equation 19 is that the ITF of the listener equal the ITF of the crosstalk cancellation head model.

3.5 - Inverse filtering of room acoustics

The crosstalk cancellers described in the preceding section invert only the listener's head response, and do not compensate for the acoustics o the listening space. It's possible to invert a room's acoustic impulse response with a casual, stable filter only when the room response is minimum phase. However, room responses are seldom minimum phase, and therefore it's necessary to incorporate significant modelling delay into the inverse filter in order to obtain an approximate inverse. This works at the same point in the room where the impulse response measurement was taken, but all other points in the room are subject to the pre-response of the inverse filter. For this reason, most techniques equalize only the minimum phase portion of the room response leaving the excess phase portion unchanged.

A method for exactly inverting acoustic impulse responses in a room is described by Miyoshi and Kaneda [MY88]. Because of the non-minimum phase nature of room responses, it is not possible to realize an exact inverse when the number of sources is equal to the number of equalization points one wishes to control. However, by adding one extra source it becomes possible to realize an exact inverse using FIR filters. This principle is called the multiple-input/output inverse theorem (MINT). This important result is obtained when the transmission convolutions are formulated as a matrix multiplication, and then the inverse of this matrix yields the set of FIR inverse filters. The MINT principle follows from the requirement that the system transfer matrix be square in order to be invertible.

A multiple error least mean square (ME-LMS) algorithm is developed to adaptively design a matrix of FIR inverse filters. These filters can perform crosstalk cancellation and also equalize for the room response. A practical implementation would require that the listener have microphones positioned in the ear canals in order to adapt the inverse filters to the optimal solution, although the microphones could be removed after the filters converge to a solution. Nelson [NE95] demonstrated that the MINT and the ME-LMS method are equivalent, and have derived conditions that must be fulfilled for an exact inverse to exist. The spatial extent of the resulting equalization zone was shown to depend on the acoustic wavelength of the highest frequency of interest.

3.6 - Head tracking system

Head tracking is necessary if the listener is not going to be constrained to a single fixed location. Both the binaural synthesizer and the crosstalk cancellation are affected by the location of the listener's head. The modification to the binaural synthesizer is rather trivial. As the listener moves his head, the synthesis HRTFs must be adjusted so that the rendered scene remains fixed to an external frame of reference, otherwise the rendered scene will move with the listener's head. The head model within the crosstalk canceller must also be updated so that the crosstalk canceller is inverting the current transmission path from the speakers to the ears.

Implementing the head tracking and adaptation has two benefits. First the 3D effect will function over a large listening area because the sweet spot is steered to the location of the listener's head. Secondly, if the tracking is fast enough, the listener will have the additional benefit of dynamical localization cues. These are very powerful cues, particularly for the resolution of front-back confusions.

A complete head-tracked 3D audio loudspeaker system is created by combining a multiple source binaural synthesizer with a crosstalk filter and a head tracker as shown in the figure below.

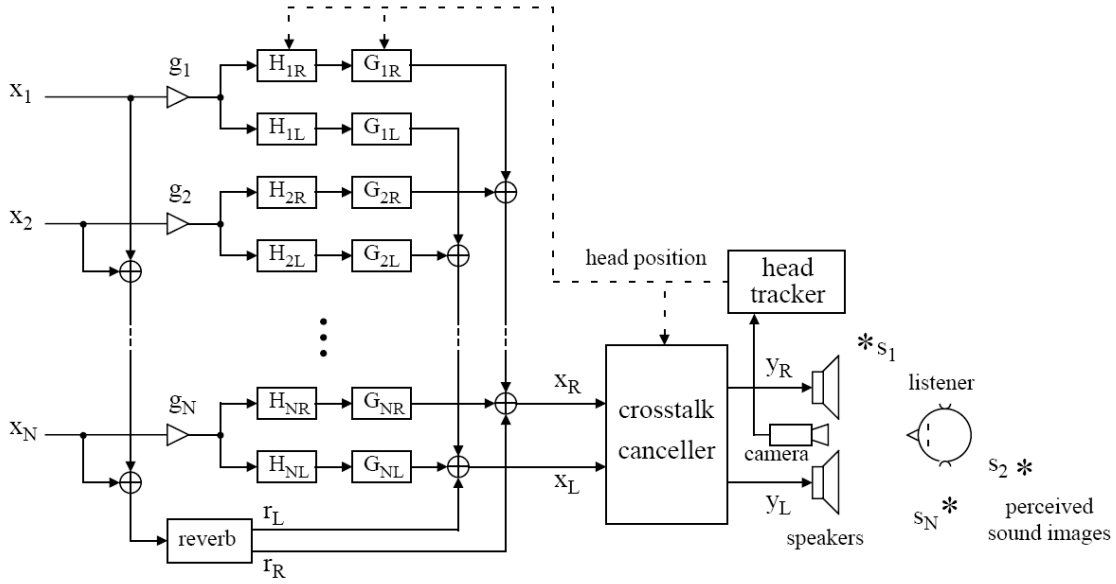


Figure 3.25 Complete implementation of a Head tracked 3D loudspeaker audio system [WG97a]

Following Gardner [WG97a] at the left are the N inputs sounds X_i whose spatial positions are to be separately synthesized. Each sound is filtered with an appropriate HRTF pair (H_{iL}, H_{iR}) to encode directional cues. The equalization of the HRTFs depends on the particular implementation of crosstalk canceller.

After binaural synthesis, the individual binaural signals are processed with the high-frequency shelving filters (G_{iL}, G_{iR}). For simplicity there are separate shelving filters for each channel. The high frequency adjusted binaural signals are summed to a single binaural pair which is input to the crosstalk canceller. The output of the crosstalk canceller is sent to the loudspeakers. As described earlier, the parameters of the crosstalk canceller, binaural synthesizer, and shelving filters depend on the current head position. This dependency is indicated in the figure with dashed lines. The figure also includes the reverberation processing suitable to achieve control of perceived source distance. Prior to binaural synthesis, each source is scaled by a gain g_1 intended to simulate the attenuation of direct sound due to air propagation. The unscaled sources are summed and fed to a reverberator that outputs a binaural signal.

The circuit allows the direct-to-reverberant ratio of each source to be controlled by the scaling gains g_1, which provide independent distance control for each source.

Finally to determine the position of the listener's head a tracker system is necessary. A tracker system consists on a source, sensor and a hardware interface that sends positional data to the system. Currently tracker technology usually involves electromagnetic sensors. Alternate systems include optical, mechanical, and acoustic tracking systems, each with advantages and disadvantages in terms of the volume of the space one can move, in accuracy and response latency. In the system shown in the figure above is used a camera attached to the head tracker system. However most part of 3D sound systems works with magnetic systems due to cost, hardware configuration and the ability to move 360 degrees.

In the case of the magnetic tracker sensor, this is attached to the body, normally in the top of the head. The sensor measures the low-frequency magnetic field generated by the source, which is located outside the body a few meters away and can be used to symbolize the position of the virtual sound source. Positional data about the user's position in relationship to the source can then be obtained in terms of pitch, yaw, roll and *xyz*. The 3D system then can compensate for the relative position of source a sensor to give the illusion of a sound that remains fixed in the space independent of head movement.

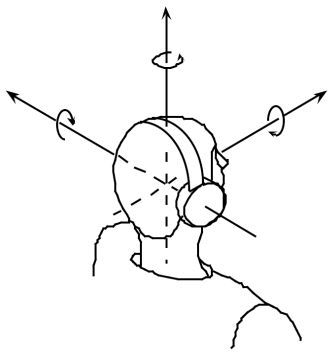


Figure 3.26 Coordinates for pitch-yaw-roll and *x, y, z* in a head tracked system [BEG00]

4 - Validation
4.1 - Implementation of the Crosstalk Canceller

For the purpose of this thesis and to obtain an overall idea of the efficiency and performance of these systems a binaural synthesizer and a crosstalk filter were programmed in matlab.

The first part of the program consists on a binaural synthesizer. In this part the program asks the user about four parameters, first the library of HRTF that is going to use and then three localization parameters, the distance, the elevation and the azimuth. With these parameters the program will attenuate the signal in function of the distance, based on the inverse square law, and it will search and read the HRTF that correspond to the data the user has typed in the program.

For this program we used two libraries of HRTF published by the MIT media lab, one called compact and another one called full. All these HRTF were measured using one KEMAR mannequin with the loudspeaker placed at 1.4 meter distance from the KEMAR, the impulse responses were sampled at 44.1 KHz. The compact data is a set of impulse responses which has been reprocessed to compensate the recording equipment response and they are ready to be used directly. The full data is what the people of the MIT recorded when they were generating the data. The compact data set has 128 taps for the FIR filter and the full data has 512 taps for the FIR filter.

Once the program has read the HRTF it will read a monaural test sound saved in a wav file and it will make the convolution between the HRTF and the test sound for the left and the right channel. Following this the program will combine the two channels of audio and it will store the processed sound as Test_Sound_Headphones and beside the values used for distance, elevation, azimuth and HRTF used.

The second part of the program is the crosstalk filter itself. To cancel the crosstalk filter the program makes an estimation of how the crosstalk signal is going to be and then it inverts the signal and it adds to the opposite channel. The block diagram of the crosstalk filter used is shown below.

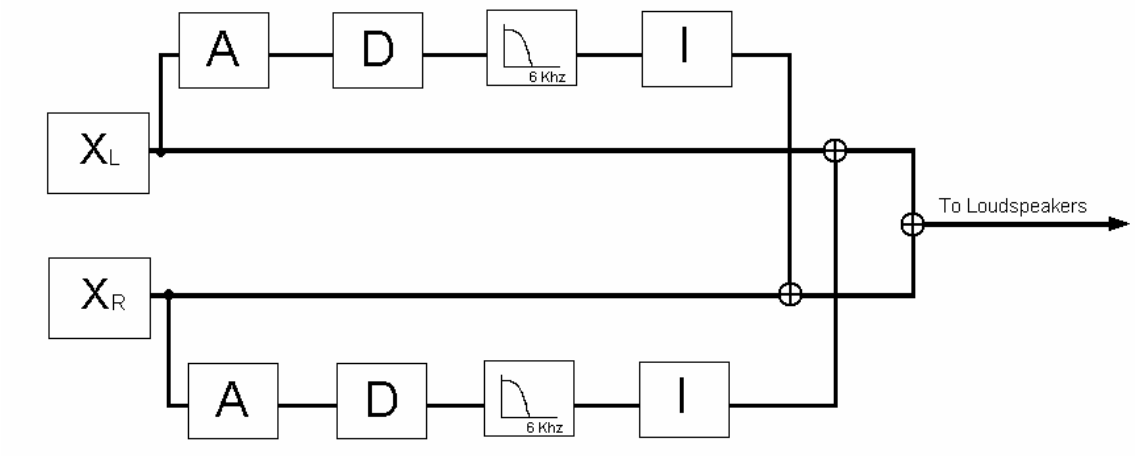


Figure 4.1 Crosstalk Filter Used

The program takes the right and the left signal and performs the same process on each one. First it attenuates the signal, after that it applies on the signal a delay corresponding to the difference of distance between the speaker and the opposite ear. In the next step it filters the signal with a lowpass filter with a 6 KHz cut off frequency. In the last step it inverts the signal and adds the signals to its respective opposite channel.

The program process this way the crosstalk signal because we know that the signal is going to be attenuated, delayed because the signal arrives later to the opposite ear and the high frequencies are going to be attenuated because of the shadowing of the head. If we adjust properly the parameters of attenuation, delay and cut off frequency we obtain an estimation of the signal we want to cancel. Once we know this we just have to invert and add the signal to the opposite channel.

Once the program has done all the process it will combine and save the audio channels in one wav file as Test_Sound_Loudspeakers and besides the values about the distance, elevation, azimuth and HRTF used. At the end the program will perform a graphical representation of the channels saved in the wav files of headphones and loudspeakers.

The following data correspond to the code used in the matlab program.

```
function crosstalk;
% AUTOR: Miguel David Botia Fernandez

clear all

%BINAURAL SYNTHESIZER

fprintf('               WELCOME TO THE CROSSTALK PROGRAM\r' );
fprintf(' PLEASE FOLLOW THE NEXT STEPS TO OBTAIN THE 3D AUDIO
SIMULATION\r\n' );


select=input('\nPlease type L to use the full data from the left
pinna, R to use the full data of the right pinna \nand H for compact
data \n','s'); % Ask for the HRTF


d=input('\nPlease introduce the distance for the simulation in meters
\n'); % Ask for the elevation

if (d<0)
   error('The value introduced is an invalid value');

end

elev=input('\nPlease introduce the degrees of elevation in steps of 10
degrees  from -40° to 90° \n'); %Ask for the elevation
if ((elev < -40) | (elev > 90))
    error('elevation must be between -40 and 90 degrees');

end
```

```
azim=input('\nPlease introduce the degrees of azimuth in steps of 5
degrees from 0° to 180° for compact data \nand from 0° to 360 for full
data \n'); % Ask for the azimuth
azim = round(azim);
if (select == 'H');
    if   ((azim < 0) | (azim > 180))
          error('azimuth must be between 0 and 180 degrees');
    end
else
    if   ((azim < 0) | (azim > 360))
          error('azimuth must be between 0 and 360 degrees');
    end
end


% Read HRTF
% Processing with compact HRTFs. They are symmetrical for left/right
% incidence, so only HRTFs for azimuths 0 <= azim <= 180 are provided.
root = 'C:\Archivos de programa\MATLAB\R2006a\work\HRTF';
dir_ch = '\';

flip_azim = 360 - azim;
if (flip_azim == 360)
    flip_azim = 0;
end
ext = '.dat';
if (select == 'L')
    pathname = hrtfpath(root,dir_ch,'full',select,ext,elev,azim);
    x(1,:) = readraw(pathname);
    pathname = hrtfpath(root,dir_ch,'full',select,ext,elev,flip_azim);
    x(2,:) = readraw(pathname);
elseif (select == 'R')
    pathname = hrtfpath(root,dir_ch,'full',select,ext,elev,flip_azim);
    x(1,:) = readraw(pathname);
    pathname = hrtfpath(root,dir_ch,'full',select,ext,elev,azim);
    x(2,:) = readraw(pathname);
elseif (select == 'H')
    pathname = hrtfpath(root,dir_ch,'compact',select,ext,elev,azim);
    tmp = readraw(pathname);
    x(1,:) = tmp(1:2:length(tmp));
    x(2,:) = tmp(2:2:length(tmp));
else
    error(sprintf('%s not a valid selection, use L, R, or H',select));
end




fs = 44100; % Sample rate
nbits = 16;% Cuantification bits
[i,fs,nbits]= wavread ('Test_sound.wav'); % Read the wav file


at = 20*(log10(1.4/d)); % Calculates the attenuation in dB in function
of the distance, 1.4 is de distance where the HRTF were measured
attenuation = 10^(at/20);
in = i * attenuation; ; % PROVISIONAL NEEDS REVISION
```

```
leftout = conv( in(:,1), x(1,:) ); %Convolution left channel
rightout = conv( in(:,1), x(2,:) );%Convolution right channel


comb= [leftout rightout]; %Combine channels


sound(comb,fs) % Plays the processed sound


s = num2str(elev); % Transform the variables azimuth and elevation
into string
c = num2str(azim);
p = num2str(d);
File = ['Test_sound_Headphones_' p 'dist_' s 'elev_' c 'azim_' select
'.wav']; % Stores the processed sound with the data about elevation
and azimuth

wavwrite(comb,fs,nbits,File)

%Representations of the signals for Headphones
fvec = 1:size(i,1);
fvec = 1000/44100 * fvec; % frequency vector in ms
subplot(2,2,1:2); plot(fvec,i);
title( sprintf( 'Test sound') ); % Represents the test sound
xlabel( 'Time [ms]');
ylabel( 'Normalized amplitude');

fvec2 = 1:size(leftout,1);
fvec2 = 1000/44100 * fvec2; % frequency vector in ms
subplot(2,2,3); plot(fvec2,leftout);
title( sprintf( 'Left HRTF convolved output') ); % Represents the left
output
xlabel( 'Time [ms]');
ylabel( 'Normalized amplitude');
axis([0 6000 -1 1]);

fvec3 = 1:size(rightout,1);
fvec3 = 1000/44100 * fvec3; % frequency vector in ms
subplot(2,2,4); plot(fvec3,rightout);
title( sprintf( 'Right HRTF convolved output') ); % Represents the
right output
xlabel( 'Time [ms]');
ylabel( 'Normalized amplitude');
axis([0 6000 -1 1]);


%CROSSTALK FILTER
% To implement the crosstalk filter we are goin to take the right and
left
% signals and we are going to add to the inverse channel a version of
this
% signals attenuated, delayed, filtered and inversed
```

```
%ATENUATION
%To sinthesize the crosstalk signal I suppose there is a attenuation
of the
%80%

leftcrossA = leftout * 0.8; %This is the signal from the left speaker
that interfiere with the right ear
rightcrossA = rightout * 0.8;% This is the signal from the right
speaker that interfiere the left ear


%DELAY

%Left signal

R=9;%Number of samples in the delay 0.00029 seg
leftcrossAD=[zeros(R,1); leftcrossA];

%Right signal

rightcrossAD=[zeros(R,1); rightcrossA];

%Add to the original signal the same samples added in the delayed
signal to make
%possible the addition

F= R+62;% Number of samples delayed plus the samples of the Lowpass
filter of 62th level

leftoutlong = [leftout;zeros(F,1)];

rightoutlong = [rightout;zeros(F,1)];



%LOW PASS FILTER

%Left channel
Wn = 0.272; % Cut off frequency at 6Khz (6000/(44100/2))Hz
N = 62;% Level of filter
gLP = 1; %Gain
LP = fir1(N,Wn);
y1 = conv(LP,leftcrossAD);
leftcrossADF= gLP * y1;

%Right Channel
Wn = 0.272; % Cut off frequency at 6Khz (6000/(44100/2))Hz
N = 62; % Level of the filter
gLP = 1;%gain
LP = fir1(N,Wn);
y2 = conv(LP,rightcrossAD);
rightcrossADF= gLP * y2;
```

```
%INVERSE

Lcross = leftcrossADF * -1; % Inverse of the left crosstalk signals
Rcross = rightcrossADF * -1; % Inverse of the right crosstalk signal

%COMBINE

leftoutsp = leftoutlong + Rcross;

rightoutsp= rightoutlong + Lcross;

Speakerout = [leftoutsp rightoutsp];


File = ['Test_sound_Speaker_' p 'dist_' s 'elev_' c 'azim_' select
'.wav']; % Stores the processed sound with the data about elevation
and azimuth
wavwrite(Speakerout,fs,nbits,File) ;

%Representations of the signals for Speakers
fvec4 = 1:size(leftoutlong,1);
fvec4 = 1000/44100 * fvec4; % frequency vector in ms
figure(2),subplot(2,2,1); plot(fvec4,leftoutlong);
title( sprintf( 'Signal Left Speaker') ); % Represents the test sound
xlabel( 'Time [ms]');
ylabel( 'Normalized amplitude');
axis([0 6000 -1 1]);

fvec5 = 1:size(rightoutlong,1);
fvec5 = 1000/44100 * fvec5; % frequency vector in ms
figure(2),subplot(2,2,2); plot(fvec5,rightoutlong);
title( sprintf( 'Signal Right Speaker') ); % Represents the test sound
xlabel( 'Time [ms]');
ylabel( 'Normalized amplitude');
axis([0 6000 -1 1]);

fvec6 = 1:size(Lcross,1);
fvec6 = 1000/44100 * fvec6; % frequency vector in ms
figure(2),subplot(2,2,3); plot(fvec6,Lcross);
title( sprintf( 'Left Crosstalk Canceller signal') ); % Represents the
left output
xlabel( 'Time [ms]');
ylabel( 'Normalized amplitude');
axis([0 6000 -1 1]);

fvec7 = 1:size(Rcross,1);
fvec7 = 1000/44100 * fvec7; % frequency vector in ms
figure(2),subplot(2,2,4); plot(fvec7,Rcross);
title( sprintf( 'Right Crosstalk Canceller signal') ); % Represents
the right output
xlabel( 'Time [ms]');
ylabel( 'Normalized amplitude');
axis([0 6000 -1 1]);
```

4.2 Localization experiments

In this part of the dissertation we are going to describe the processes followed to perform the localization experiments on people. We will describe the experimental procedures for each experiment and then discuss the results obtained.

The experiments were conducted in the audiolabor of the University of Applied Sciences St. Pölten. The dimensions of the sound studio are 4,30L x 3,19W x 2,62H meters and the walls of the sound studio were covered by acoustic pyramid foam of 8 cm height. The theoretical reverberation time of the room at 500 Hz is of approximately 0.14 seconds.

The experiments were divided in three parts. The first part consists on a brief introduction to 3D audio and a brief explanation of what the subject is going to hear. A helicopter sound was slowly panned around the head of the subjects from 0 to 360 degrees of azimuth and then a second sound of a helicopter was panned from -40 to 90 degrees of elevation. The sounds were accompanied by an explanation of where the sound was supposed to come from. This procedure was repeated with the test sound used in 5 aleatory positions to familiarize the subject with the test sound. The test sound used in these experiments was a monaural sound that contains 6 burst of pink noise. Every burst has duration of 300 msec with 30 msec onset and offset ramps, with 600 msec gap between bursts. This first part can be considered as a little training for the subject.

In front of the subject were placed three localization charts to help the subject to localize the sound. These charts consist on two clocks showing azimuth and elevation angles and a distance diagram. The azimuth clock was divided in increments of 15 degrees from 0 to 180 degrees in left and right hemisphere and the elevation clock was divided in increments of 20 degrees from 40 degrees down up to 80 degrees up. Distance judgements were given in a somatocentric coordinate system: Position inside the head, position close to the surface of the head and external locations shoulder, elbow, arm and beyond reach length. All this distances were numbered from 1 to 6 where 1 was the “In head” position and 6 was the “beyond reach” position. The subject was asked to report in reference to these charts.

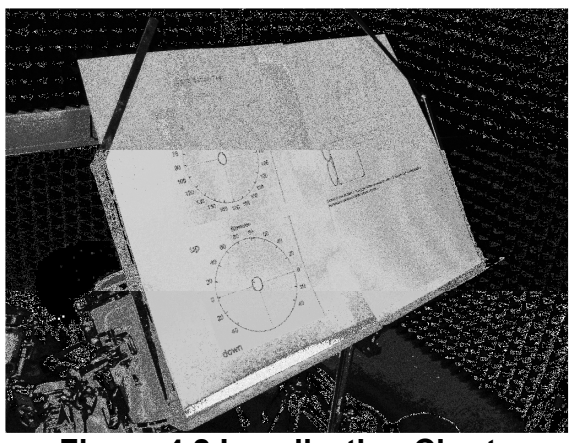


Figure 4.2 Localization Charts

The second part was the localization test over headphones. In this part 5 different processed test sounds were presented to the listener over headphones. The source was the same 6 burst of pink noise described before. The source was processed with the program made with matlab to encode directional cues for a target location and then was presented to the subject over headphones at a listening level of approximately 70dBA. The signal was presented to the subject over BEYERDYNAMIC DT-990 PRO circumaural headphones. These models are open diffuse field headphones.

The targets locations were randomly chosen from a set of 5 possible locations. Each location was tested once per subject. Although the program in matlab can include distance cues all the sounds were processed at 1.4 meters, excluding this way any distance cue because that is the distance where the HRTF were measured. In this experiment 10 subjects were tested. None of these subjects had any prior experience with 3D audio systems or sound localization experiments and none reported any known hearing loss.

Fig 4.3 Headphones localization test

The third and last part of these experiments was the localization test over loudspeakers. The loudspeaker experiment procedure was similar to the headphones experiment, differing only in the test sound used. The test sound was the same 6 pink noise bursts used with the headphones but this time the signal was also processed by the crosstalk filter from the matlab program.

To carry out the experiment, one equilateral triangle, with 80cm on every side, was drawn with tape in the floor of the sound studio. In the central vertex of the triangle was placed the center of the subject's head and in the remaining vertexes were placed one GENELEC 8020A active loudspeakers on each one. The speakers were positioned at ±30º azimuth and 0º elevation with respect to the subject.

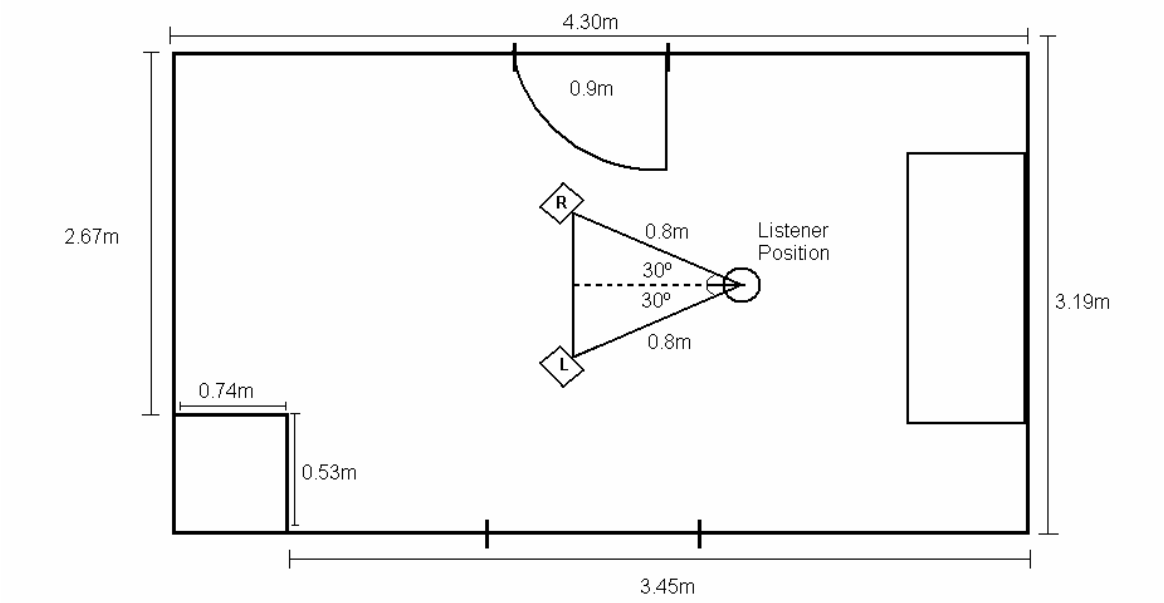


Figure 4.4 Situation of the room

The height of the loudspeakers was measured to align the center of the speaker with the center of the subject's ears. To maintain the listener's head in the same position during the entire test the head of the listener was tied with a strap to a metal bar fixed to the chair.

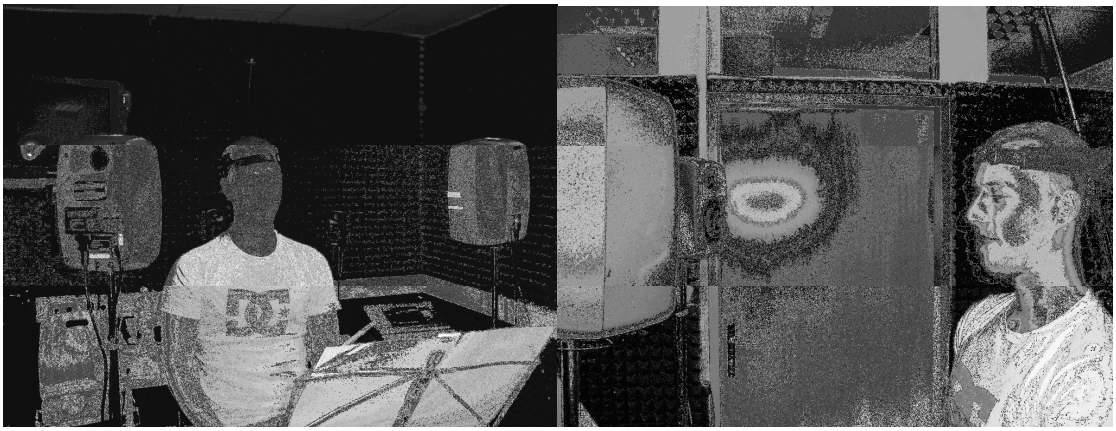


Figure 4.5 Loudspeakers test

The stimuli used during the test were in the same target positions than in the headphones test and they were presented to the subject in an aleatory order. Every test sound was tested once per subject and the subjects were the same 10 subject tested in the headphones experiment. We will refer to the subjects as A, B, C, D, E, F, G, H, I and J along the analysis of the test.

4.2.3 - Analysis of the tests

According to Blauert [BL83] and Wenzel [WW93] there are three principal types of localization errors:

- Systematic errors between the mean judged location and the target location, these errors are called localization errors. In the free-field listening the smallest errors are seen in azimuth in frontal targets and the biggest errors are seen in elevation judgements for the rear targets.

- Front-back and up-down reversals, where a target location is confused with the mirror symmetric location obtained by reflecting the target across the frontal plane, for a front-back reversal, or the horizontal plane for an up-down reversal.

- Variation of the responses around the mean, attributed to perceptual noise.

In addition to these errors there are variations in the distance responses, but as explained before the targets were free of distance cues. That is the reason why distance errors are reported without consideration of a target distance.

The following figures below shows the Azimuth and elevation localization error in degrees of the five stimuli used during the tests across all the subjects. The stimuli used during the experiment are shown in the table below.

	Test 1	Test 2	Test 3	Test 4	Test 5
Elevation	20º	-20º	80º	0º	40º
Azimuth	60º	210º	150º	180º	330º

Figure 4.6 Stimuli used

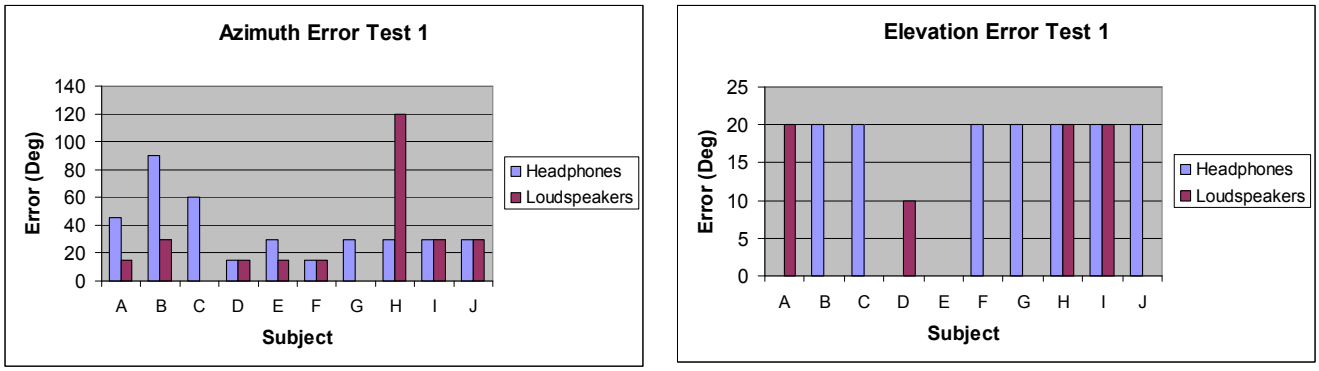
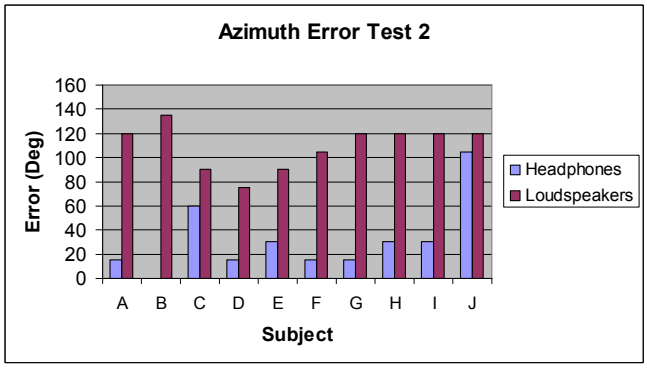


Figure 4.7 Azimuth and Elevation error for test 1



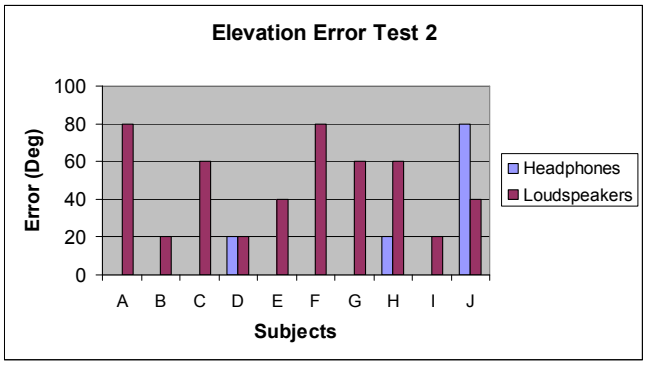
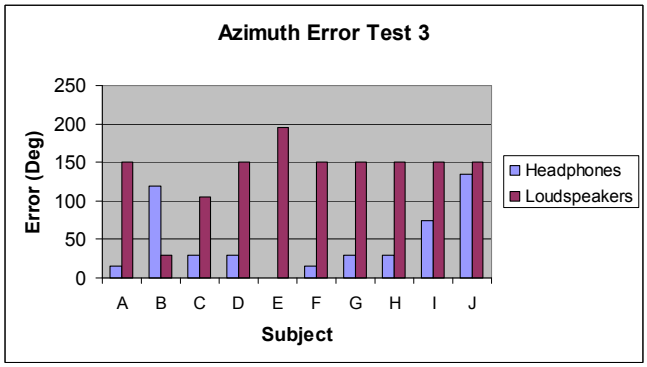


Figure 4.8 Azimuth and Elevation error for test 2



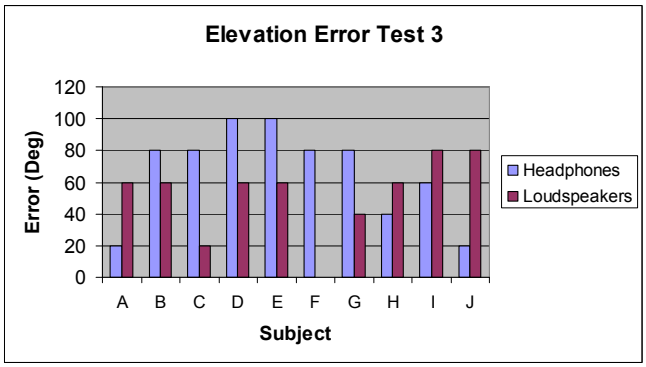
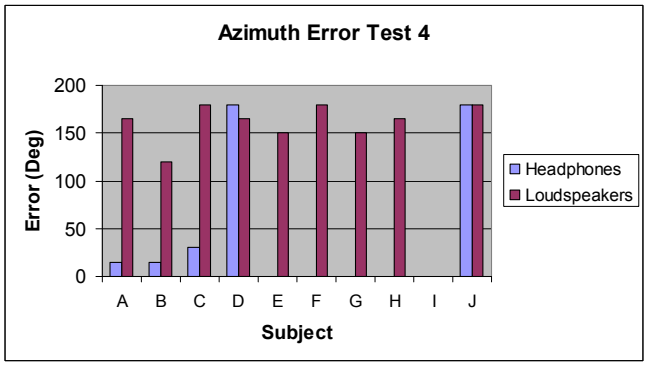


Figure 4.9 Azimuth and Elevation error for test 3



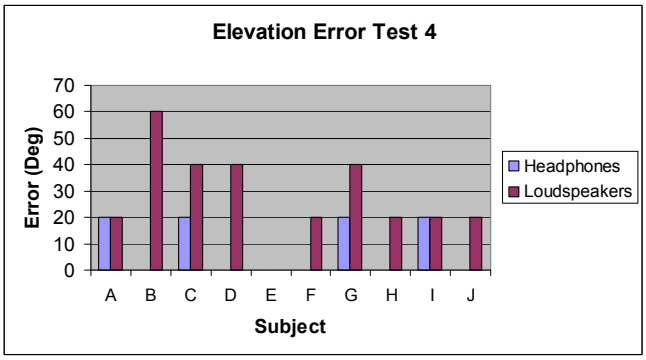
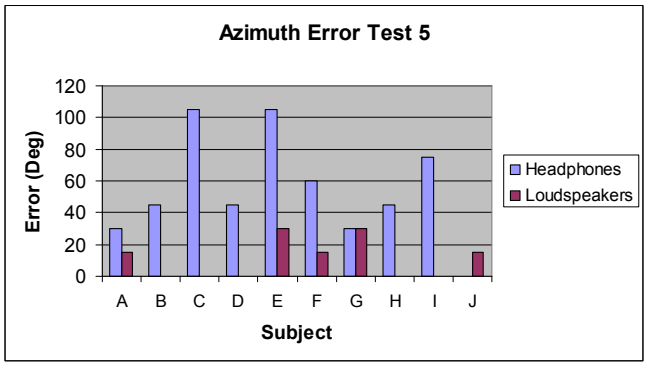


Figure 4.10 Azimuth and Elevation error for test 4



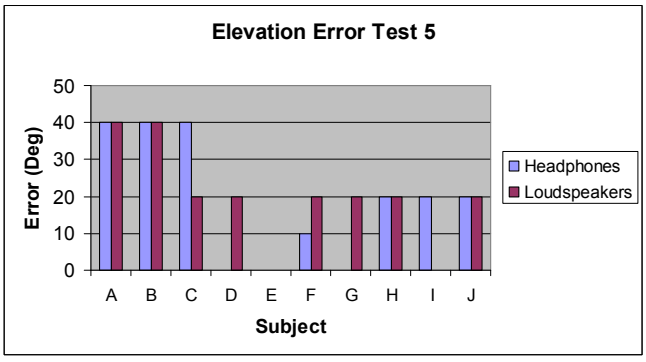


Figure 4.11 Azimuth and Elevation error for test 5

The table below summarizes the average angle error in azimuth and elevation for the headphone and loudspeaker experiments.

	Azimuth	Elevation
Headphones	42,6º	23,8º
Loudspeakers	86,1º	31º

Figure 4.12 Average angle error

As we can see the main error difference is established between the azimuth for headphones and loudspeakers. However error in elevation is similar for both experiments. This may indicate that the KEMAR HRTFs are not as effective for spatial synthesis as the HRTFs of a good human localizer. Many of these azimuth errors are cause by the elevation in the targets, we would expect smaller azimuth errors if targets were restricted to the horizontal plane. The biggest errors in azimuth are shown in rear targets and in targets with high levels of elevation like in test 3 and 4. On targets at 80 º elevation, almost overhead, although the report of elevation was good for both experiments over loudspeakers targets in the rear were reported by many listeners to be in the front.

Over loudspeakers low elevations are perceived like high elevations, for instance in test number 2 over headphones almost all the subjects reported that the elevation was in the range between -20º and 0º but over loudspeakers the values of the range increases from 0º to 60º. For high elevations targets the elevation error average was lower for loudspeakers than for headphones. The best performance, and consequently the lower error, in elevation was established for targets located in the range between 0º and 40º for both experiments. Overall elevation performance over loudspeakers appears to be poorer than over headphones. This is not surprising if we consider that the high frequency cues are corrupted by crosstalk and room reverberation when delivering the signal over loudspeakers. In this study all the subjects had similar error average; however subject E was the most precise when reporting the angles.

A judgement is considered to be a reversal if the reversed judgement, obtaining by reflecting the judgment across the horizontal plane, is closer to the target. The left and right effect is much easier to detect compare to front and back effect. The reason is because the left right effect depends on IID and ITD but front and back effect depends on pinna's response. The reversals in this study are divided into Front-Back and Back-Front reversals. The table below shows the reversal percentages for horizontal targets across all subjects.

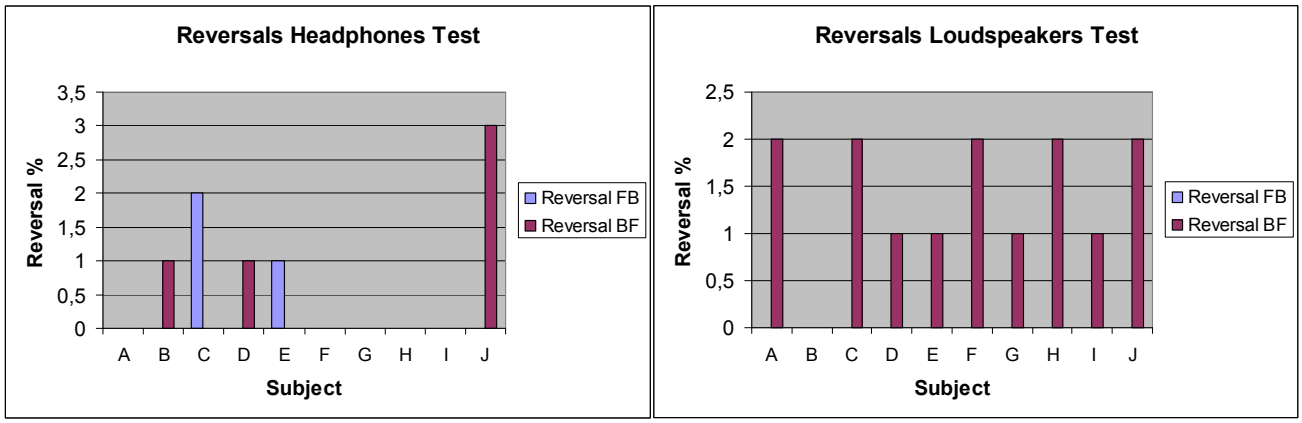


Figure 4.13 Percentages of Front to Back (FB) and Back to Front (BF) reversals.

Over headphones some frontal locations are reversed to the rear and a few rear locations are reversed to the front, but over loudspeakers almost all the reversals are from sounds that are in the back and they are perceived in the front. In both cases the pattern of reversals is specific to the individual. These reversals might be reduced if we use individualized HRTFs instead of Non-individualized HRTFs in the program. In this case subject J clearly has a propensity to perceive the stimuli as frontal in both experiments.

The figure below shows the average of the distance perceptions for each test.

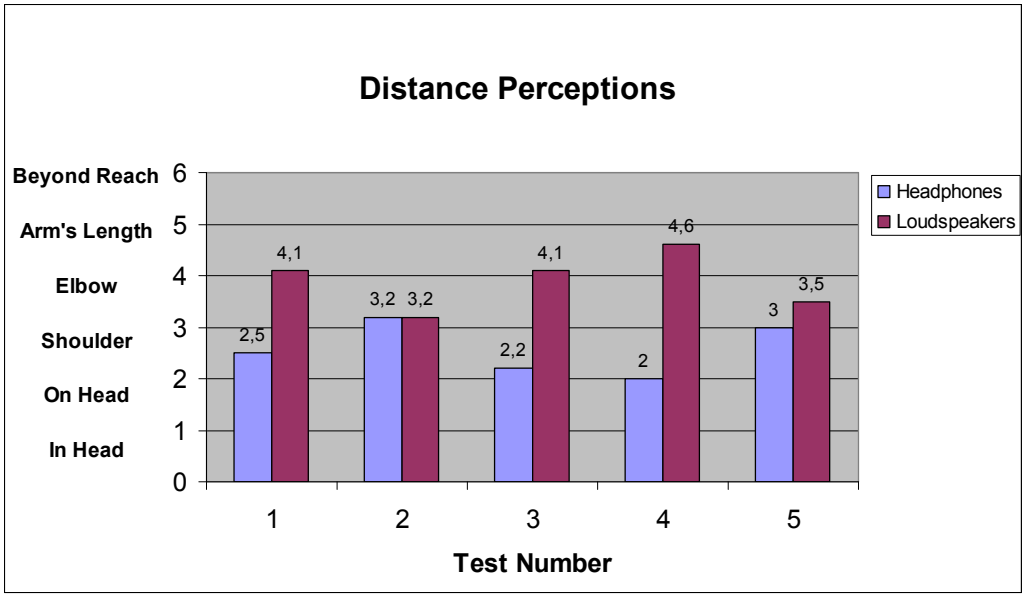


Figure 4.14 Distance perception average

The figure clearly shows closer distance judgements for headphones than for loudspeaker. Over headphones almost all the listeners report that the sound was perceived inside or on their head. Only in a few cases the subject said that the sound was in the shoulder or beyond. But over Loudspeakers we can observe an increase of the distance that goes from the shoulder position to the arms length. Externalization is clearly better using loudspeakers. The headphone data shows that judgment distances depend on target azimuth.

Targets in the front are judged closer than rear targets .The biggest difference is established for targets located in the rear where over headphones they tend to be perceived close to the head and over loudspeakers they were positioned in the arm length position. A further analysis shows that for rear target locations, back to front reversals were accompanied by an increase in the perceived distance.

5 - Conclusions

Overall the results of the experiments and performance of the synthesizer and the crosstalk canceller are satisfying. Although the performance on headphones is better than with loudspeakers the performance over loudspeakers is still satisfying. The results obtained are similar to the results obtained from other researchers, see [WW93] [WG97a] [WK89b], and they were within the foreseen.

However the performance of the system can be improved if we add upgrades to the synthesizer like different pinna sets or individualized HRTF for every subject, we can also improve the crosstalk canceller adding to the filter the inverse response of the room and the of speakers, reverberation cues or if we try another filter topology. Also we could include equalization to the frequencies attenuated by the headphones or loudspeakers. The results also might be better if instead of a sound studio we carry out the experiments in an anechoic chamber. And of course we can implement a head tracking system to achieve a more natural way of hearing 3D audio instead of being with the head permanent in a small area.

3D audio systems are getting better and more complex day by day and crosstalk filters are also part of this development. The DSPs and technology used in 3D sound systems are enhancing every day performing processes much faster and improved. Daily emerge more applications and uses for 3D audio and is still growing. 3D audio through loudspeakers is still developing but who knows what the future will provide.

Bibliography

[AS90] Asano, F., Suzuki, Y., and Sone, T. **"Role of spectral cues in median plane localization".** *Journal of the Acoustical Society of America*, **88**, 159–168. (1990).

[BA68] Batteau, D. W**." Listening with the naked ear".** *In S. J. Freedman (Ed.),* (1968).

[BB61] Bauer, B. B. **"Phasor Analysis of Some Stereophonic Phenomena"**, *J. Acoust. Soc. Am.*, 33(11), pp. 1536-1539. (1961).

[BEG87] Durand R. Begault. **"Control of Auditory Distance"**, *dissertation, UCSD* (1987)

[BEG91] Durand R. Begault. **"Challenges to the Sucessful Implementation of 3-D Sound"** *Journal of the Audio Engineering Society, vol. 39(11), pp. 148-151* (1991).

[BEG92] Durand R. Begault. **"Perceptual Effects of Synthetic Reverberation on Three-dimensional Audio Systems"** *Journal of the Audio Engineering Society, vol. 40(11), pp. 895-904* (1992).

[BEG93] Durand R. Begault. **"The Evolution of 3-D Audio"***, Mix Magazine, vol. 17(10), pp. 42-46* (1993).

[BEG00] Durand R. Begault*.* **"3D Sound for Virtual Reality and Multimedia"** (2000)

[BEW93] Begault, D. R., and Wenzel, E. M **"Headphone Localization of Speech".** *Human Factors*, **35**, (1993).

[BL69] Blauert, J**, "Sound localization in the median plane"**. *Acustica*, **22**, 205–213. (1969).

[BL83] J. Blauert, **"Spatial Hearing: The Psychophysics of Human Sound Localization"** *MIT Press, Cambridge, MA,* (1983).

[BU87] Butler, R. A. **"An analysis of the monaural displacement of sound in space"** *Perception and Psychophysics*, **41**, 1–7. (1987).

[CB89] Cooper, D. H., and J. L. Bauck **"Prospects for Transaural Recording**", *J. Audio Eng. Soc.*, 37(1/2), pp. 3-19. (1989).

[CO63] Coleman, P.**" An analysis of cues to auditory depth perception in free space"***PsychologicalBulletin*, **60**, 302–315. (1963).

[DA71] Damaske, P. **"Head-related Two-channel Stereophony with Loudspeaker Reproduction"**,*J. Acoust. Soc. Am.*, 50(4), pp. 1109-1115. (1971)

[FX05] **"DAFX – Digital Audio Effects"** Edited by Udo Zölzer, *John Wiley & Sons* (2005)

[GA69] Gardner, M. B **"Distance estimation of 0 degree or apparent 0 degree–oriented speech signals in anechoic space".** *Journal of the Acoustical Society of America*, **45**, 47–53. (1969).

[GIE92] Gierlich, H. W. **"The Application of Binaural Technology".** *Applied Acoustics*, **36**, 219–243. (1992).

[GI84] Gill, H. S. **"Review of Outdoor Sound Propogation"** (*Unpublished technical report). Wilson, Ihrig, and Associates, Oakland, CA.* (1984).

[GRI90] Griesinger, D. **"Equalization and spatial equalization of dummy head recordings for loudspeaker reproduction".** *Journal of the Audio Engineering Society*, **37**, 20–29. (1990).

[GRI93] Griesinger, D. **"Quantifying Musical Acoustics through Audibility".** *Unpublished transcript of the Knudsen Memorial Lecture, Denver Acoustical Society.* (1993).

[HA96] Harris, C. M "**Absorption of sound in air vs. humidity and temperature"** *J. Acoust. Soc. Am*, 40, 148–159. (1996)

[HH72] Haas, H **"The influence of a single echo on the audibility of speech"**. *Journal of the Audio Engineering Society*, **20**, 146–159. (1972)

[HK57] Hanson, R. L., and Kock, W. E. **"Intereting effect produced by two loudspeakers under free space conditions"**. *Journal of the Acoustical Society of America*, **29**, 145. (1957)

[IM78] Iwahara, M., and T. Mori. **"Stereophonic sound reproduction system***". United States Patent 4,118,599.* (1978)

[JO92] Jot, J.-M. **"Etude et réalisation d'un spatialisateur de sons par modèles physiques et perceptifs"** Ph.D. thesis, Telecom Paris. (1992).

[KE62] Kellogg, W. N "Sonar system of the blind". *Science*, **137**, 399–404 (1962).

[KL93] Kleiner, M., Dalenbäck, B. I., and Svensson, P **" Auralization—an overview".** *Journal of the Audio Engineering Society*, **41**, 861–875 (1993).

[MA91] Martens, W. L **" Directional Hearing on the Frontal Plane: Necessary and Sufficient Spectral Cues"**. Ph.D. Dissertation, Northwestern University (1991).

[MI72] Mills, W. "**Auditory localization***". In J. V. Tobias (Ed.), Foundations of Modern AuditoryTheory. New York: Academic Press.* (1972).

[MID91] Middlebrooks, J. C., and Green, D. M. **"Sound Localization by Human Listeners"** *Annual Review of Psychology*, **42**, pp135–159. (1991).

[MID92] Middlebrooks, J. C. **"Narrow-band sound localization related to external ear acoustics"**, *J. Acoust. Soc. Am.*, 92(5), (1992).

[MO90] Moore, F. R. **"Elements of Computer Music".** *Englewood Cliffs, NJ: Prentice-Hall.* (1990).

[ML89] Møller, H.**. "Reproduction of Artificial-Head Recordings through Loudspeakers",***J. Audio Eng. Soc.*, 37(1/2), pp. 30-33. (1989)

[ML92] Møller, H. **"Fundamentals of Binaural Technology"**, *Applied Acoustics*, 36, pp.171-218. (1992).

[MY88] Miyoshi, M., and Y. Kaneda **"Inverse Filtering of Room Acoustics"**, *IEEE Trans. Acoust,. Speech, and Signal Processing*, 36(2), pp. 145-152. (1988).

[NE95] Nelson, P. A., F. Orduna-Bustamante, and H. Mamada **"Inverse Filter Design and Equalization Zones in Multichannel Sound Reproduction"**, *IEEE Trans. Speech and Audio Processing*, 3(3), pp. 185-192. (1995).

[OP89] Oppenheim, A. V., and R. W. Schafer **"Discrete Time Signal Processing",** *Prentice Hall, Englewood Cliffs, NJ.* (1989).

[RA07] Rayleigh, L. **"On our perception of sound direction"**. *Philosophical magazine*, **13**, pp 214–232. (1907).

[RV89] Rife, D. D., and Vanderkooy, J. **"Transfer–function measurements with maximum–length sequences"**. *Journal of the Audio Engineering Society*, **37**, 419–444. (1989).

[SA63] Schroeder, M. R., and B. S. Atal **"Computer simulation of sound transmission in rooms"**, *IEEE Int. Conv. Record*, 7, pp. 150-155. (1963)

[SC73] Schroeder, M. R **"Computer Models for Concert Hall Acoustics",** *Am. J. Physics*,41, pp. 461-471. (1973).

[SE82] Sheeline, C. W **"An investigation of the effects of direct and reverberant signal interaction on auditory distance perception***". Ph.D. Dissertation, Stanford University* (1982).

[SH74] Shaw, E. A. G**. "Transformation of sound pressure level from the free field to the eardrum in the horizontal plane"**. *Journal of the Acoustical Society of America*, **56**, 1848–1861. (1974).

[SL61] Schroeder, M. R., and Logan, B. F**""Colorless" artificial reverberation"**. *Journal of the Audio Engineering Society*, **9**, 192–197 (1961).

[ST68] Shaw, E. A., and Teranishi, R, **"Sound pressure generated in an external–ear replica and real human ears by a nearby sound source"**. *Journal of the Acoustical Society of America*, **44**, 240–249. (1968).

[TH67] Thurlow, W. R., and Runge, P. S **"Effects of induced head movements on localization of direct sound"**. *Journal of the Acoustical Society of America*, **42**, 480–487. . (1967).

[VB60] Von Békésy, G. **"Experiments in Hearing"** *(E. G. Wever, Trans.). New York: McGraw–Hill.* (1960).

[WA40] Wallach, H. **"The role of head movements and vestibular and visual cues in sound localization".** *Journal of Experimental Psychology*, **27**, 339–368. (1940).

[WG95] William G. Gardner **"Transaural 3-D audio"**, *M.I.T. Media Lab, Perceptual Computing Section,* Technichal Report no. 342. (1995).

[WG97a] William G. Gardner **"3D Audio Using Loudspeakers"** *Ph.D. thesis, Dept. of Media Arts and Sciences, MIT. pp 52-118(1997a)*

[WG97b] William G. Gardner. **"Head Tracked 3-DAudio Using Loudspeakers"**, *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY. (1997b).

[WG98] William G. Gardner. **"3-D Audio Using Loudspeakers"**, *Kluwer Academic, Norwell, MA.* (1998)

[WG99] William G. Gardner**. "3D Audio and Acoustic Environment Modelling"** *pp 1-8* (1999)

[WW93] Wenzel, F.L. Wightman, and D.J. Kistler, **"Localization Using Nonindividualized Head-Related Transfer Functions"** *Journal of the Acoustical Society of America, vol. 94(1), pp. 111-123* (1993).

[WK89a] Wightman, F. L., and D. J. Kistler **"Headphone simulation of free-field listening.I: Stimulus synthesis",** *J. Acoust. Soc. Am85(2), pp. 858-867.*(1989a)

[WK89b] Wightman, F. L., and D. J. Kistler **"Headphone simulation of free-field listening.II: Psychophysical validation"**, *J. Acoust. Soc. Am* 85(2), pp. 868-878. (1989b)

[WK92] Wightman, F., and D. Kistler **"The dominant role of low-frequency interaural time differences in sound localization"**, *J. Acoust. Soc. Am.*., 91(3), pp. 1648-1661. (1992)

Webs

- **http://sound.media.mit.edu/KEMAR.html**
- **http://www.ecel.ufl.edu/~shassan/courses/eel6539/**
- **www.mathworks.com**
- **http://www.hitl.washington.edu/scivw/EVE/I.B.1.3DSoundSynthesis.html**
- **http://www.ee.psu.edu/reu/All_journal/2003V1/REUV1_p45p53.pdf#search=%22crosstalk%20matlab%22**
- **http://www.pa.msu.edu/acoustics/loc.htm**
- **http://en.wikipedia.org/wiki/Pinna** (Checked on 13-11-2006)
- **http://labrosa.ee.columbia.edu/matlab/**

Glossary of terms

- **DSP**: Digital Signal Processor.
- **FIR** : Finite Impulse Response
- **HRTF**: Head Related Transfer Function.
- **IHL**: Inside the Head Localization.
- **IID**: Interaural Intensity Differences.
- **ITD**: Interaural Time Differences.
- **ITF**: Interaural Transfer Function.

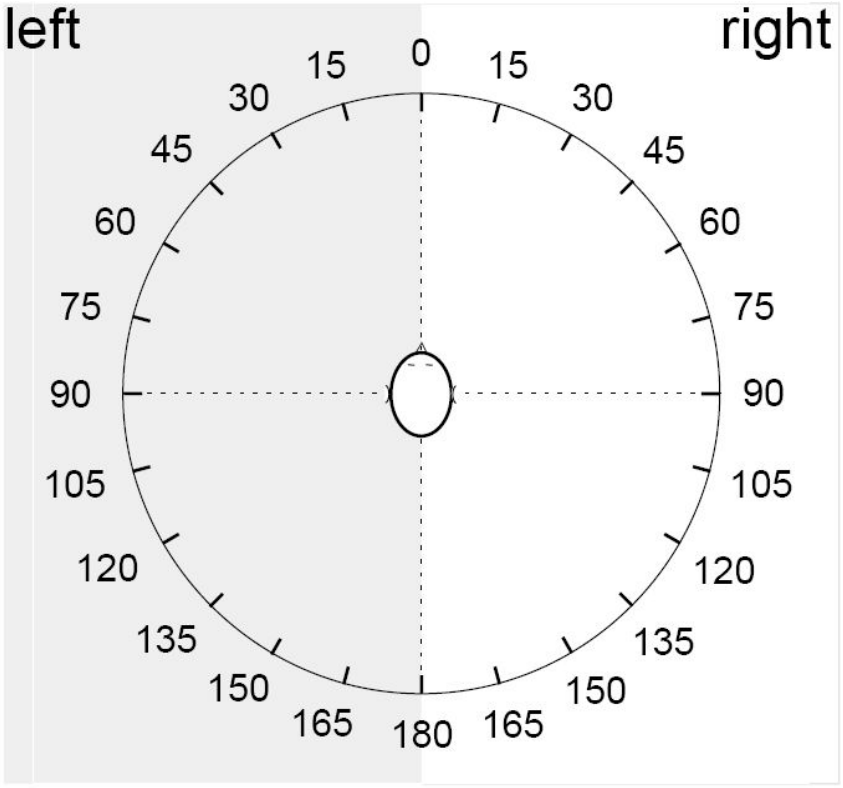
Attached Documents

Document A Speakers specifications

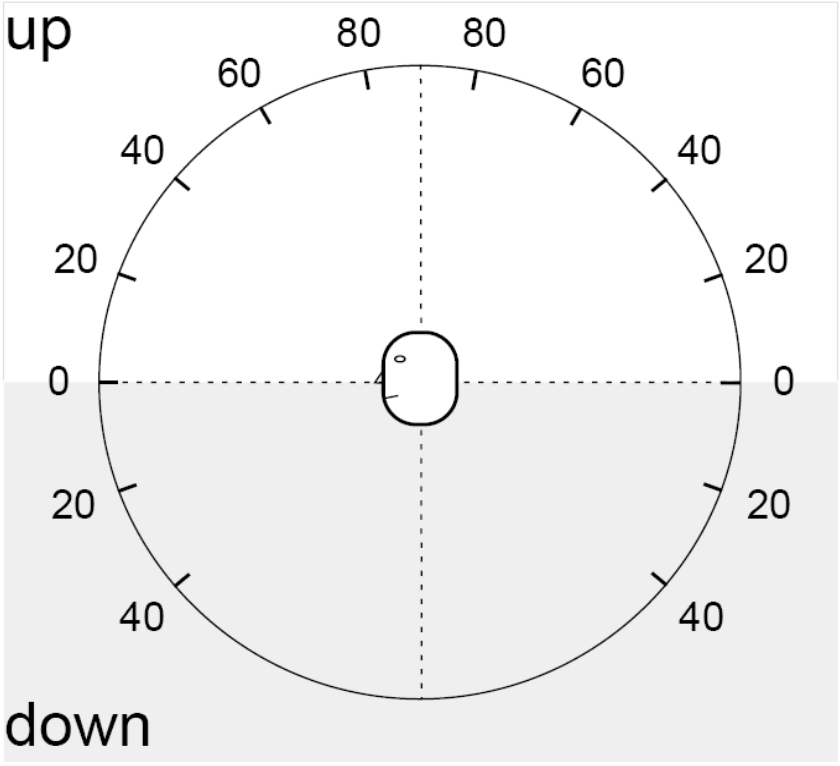
Document B Headphones specifications

Document C Localization charts [WG97a]

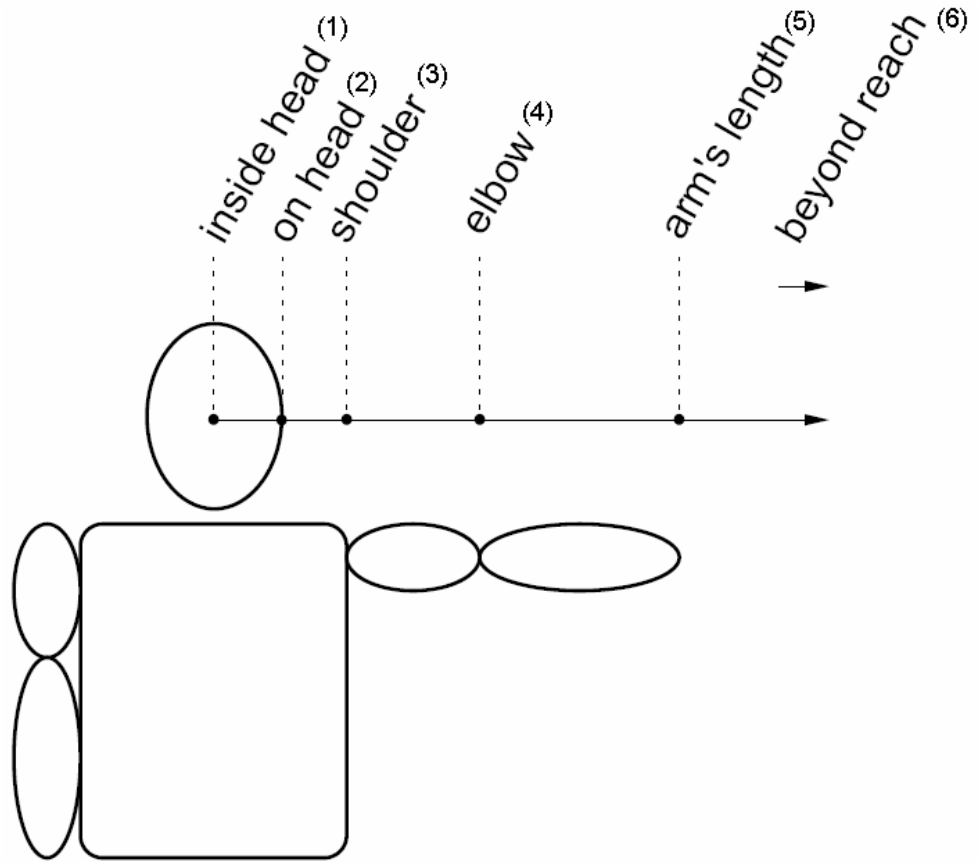
Azimuth



Elevation



Distance



Note: If the distance is bigger than number 6 write a 7 on the table and then the distance in meters within a parenthesis.