DESIGN AND IMPLEMENTATION OF AN AURALIZATION SYSTEM WITH A SPECTRUM-BASED TEMPORAL PROCESSING OPTIMIZATION

by Frank Filipanits Jr.

Master's Research Project
University of Miami
May 1994

An auralization software system for performing static (non-moving) sound source placement with headphone playback is developed using present theory and algorithms. Several caveats for auralization system design are identified and addressed. One method of temporal computation optimization is then presented. It is shown that bandwidth analysis of the raw sound source greatly reduces the computation time necessary for auralization synthesis, by identifying frequency ranges which contain zero information and can be ignored during processing. A residue method is used to evaluate the resulting algorithm.

To Merrily -

"The Light illumine you."

ACKNOWLEDGMENTS

The author would like to thank the myriad persons who provided support and encouragement throughout the pursuit of this project. In particular:

Tom Zudock for piquing my interest in three-dimensional audio and guiding my efforts to code it into reality; Ken Pohlmann and John Monforte at UM for providing inspiration and enforcing the deadlines, as well as excellent instruction and a general helping hand; Jim Boyk at Caltech for opening my ears and convincing me that I could actually make a career out of audio; Jonathan Abel at Crystal River Engineering for discussing auralization and optimization as implemented in present products; John Antony for providing "software support"; Sean Richards for knowing when to drag me away from my computer and when to let me be, and for keeping me excited about learning; Tim Onders for being a technical "sounding board" and good friend.

Thanks also to my parents, who provided the item of greatest worth — opportunity. Thank you for standing by me through the many trials and decisions of my educational career.

This version of the thesis has been prepared for general distribution with a mind toward paper conservation; therefore it does not meet the thesis style guidelines, nor does it include the (lengthy) appendices detailing the .WAV file standard and source code listings. The full text and relevant source, executable, and data files are available on request from franko@alumni.caltech.edu, and may be placed on an ftp site when a suitable one is identified.

CHAPTER 1: INTRODUCTION TO AURALIZATION

1 INTRODUCTION TO AURALIZATION

1.0 What is Auralization?

Virtual Reality is one of the hottest buzzwords today in the electronics industry. Products of all varieties are claiming to be a part of this emerging technology. Unfortunately, much of the hype is just that — hype. While a great deal of time and money has been spent in the virtual arena, the fact remains that the enabling technologies are still in their infancy. We have yet to fully understand the underlying human perceptual systems, much less develop our own devices to fool them beyond simple effects such as conventional stereo audio.

Until recently, the focus of these efforts was in providing stereoscopic three-dimensional graphics to stimulate our sense of vision. This is understandable, considering the extent to which human perception relies on visual information. We use our eyes as the primary tool for exploring our world, and when presented with contradictory information it is our visual perceptions that take precedence in our mental processing [1].

After much effort by researchers, stereoscopic displays are now available which can provide some degree of representation of a virtual world. However, the early pioneers found that though visual cues are a predominant part of our perceptions, they alone are not sufficient to create believable worlds. Coupling a three-dimensional visual display with conventional stereo sounds presents a very unnatural experience for the user. As a result, the field of auralization — three-dimensional sound — was drawn to the forefront of audio.

There are a number of motivations for developing auralization systems in addition to the advent of virtual reality. True three-dimensional processing adds an auxiliary creative element to be manipulated by the commercial musician or record producer, providing a new realm of entertainment to explore. Auralization also allows end-users to benefit from the "cocktail party" effect, the brain's ability to use localization cues to isolate a single conversation from a multitude of similar sounds. This is very useful in designing systems where a user must monitor several communication channels at once. Applications include air traffic control, NASA mission control, and fighter pilot communications.

Virtual audio displays also potentially utilize auralization algorithms, with promise for use in a general office environment as well as applications for the visually impaired.

The newfound importance of this field is evidenced by the recent explosion of auralization publications, both technical and pedestrian. While only a few years ago pioneers in virtual audio had difficulty getting papers published [2], the past few years have seen numerous articles in the Journal of the Audio Engineering Society (JAES) and the Journal of the Acoustical Society of America (JASA), as well as various virtual reality, telepresence, and human factors publications. The October 1993 AES convention in New York was titled "Audio in the Age of Multimedia" and featured numerous workshops and panel discussions on auralization. In addition, the popular press has latched on to the terms "3-D" and "Virtual", producing a plethora of articles on various degrees of auralization processing available to the consumer and semi-professional musician. Unfortunately, many of these systems are more accurately classified as surround sound or enhanced stereo, rather than true auralization systems [3][4][5].

It is in the AES journal's recent special auralization issue [6] that Kleiner proposed a definition for the term at the center of this activity:

> "Auralization is the process of rendering audible, by physical or mathematical modeling, the sound field of a source in space, in such a way as to simulate the binaural listening experience at a given position in the modeled space."

1.1 A bit of history

Some of the earliest research into the spatial perception of sounds was performed by Mills during the 1950's, determining the minimum audible angle for perceived source motion [7]. Investigation into the mechanisms and limits of human spatial auditory perception continued through the next several decades. Researchers such as Perrot explored minimum resolution angles for sources of different velocities and spectral content [8][9][10]. More recently, Makous and Middlebrooks have studied sources varying in two dimensions [11].

Through the years, a variety of means for recording three-dimensional soundfields have been developed. The Ambisonics system [12] attempted to record the three orthogonal velocity vectors and one absolute pressure at a given point in space, thus completely defining the soundfield at that point. A matrix network and equalizers manipulated the four channel direct recording (A-format) to a different four channel form (B-format) which could then be manipulated to produce one of the following: a steerable mono output, a stereo pair whose effective angle, vertical tilt and rotation can be manipulated, or a quadraphonic set of outputs whose effective direction and tilt can be manipulated. While it found some success by providing additional control over a conventional stereo pair in multitrack recording, the need for a four-channel recording medium and specialized playback environment limited the usefulness and commercial viability of this system for general use.

Dummy-head recording has demonstrated great success within the limitations of a two-channel stereo-compatible format which requires no unusual playback apparatus. Microphones in a carefully constructed artificial head record sound from the location of the "eardrums" onto a standard two-track medium for playback through headphones. These systems can provide convincing reconstructions of a number of auditory environments, though results vary based on the accuracy of the artificial ear and the correlation to an individual listener's head and ear shape. They are also limited to playback through headphones.

The primary failing of these systems is that they are only capable of recollection, not synthesis. It is not possible for the user to position a pre-recorded sound at an arbitrary position in space; items are locked into their positions as they were recorded.

1.2 The current state of the art

It is only recently that computer processing power has reached a level enabling us to even consider synthesis of three-dimensional soundfields. Unfortunately, the auditory cues used by the brain are fairly fragile with respect to listening environment. Because loudspeaker playback setups vary widely both in physical arrangement and design criteria such as time-alignment, most auralization systems are designed for headphone playback. With headphones, the transducer location is fixed and alignment between multiple drivers is usually no longer a concern. In general, headphones exhibit distortions an order of magnitude smaller than loudspeakers and avoid the difficulty of interaction with widely varying environments.

Wightman and Kistler provided groundbreaking data and validation for free-field simulation over headphones [13][14]. Through extensive experimentation, they recorded the head-related transfer functions (HRTFs) of numerous subjects for an array of sound source locations. The HRTF is one of three cues used by the brain to decipher location information, and is the primary intimation utilized to extract sound source elevation. It is modeled as a filter which accounts for the effect of the reflections off the pinnae (outer ear) and shoulder, as well as the shadowing effect of the head itself. The head-related transfer functions (HRTFs) acquired during their research are still widely used today. Wightman and Kistler's "SDO" HRTF provided a basis for the set utilized in the programs accompanying this research [15].

Begault recently identified a number of challenges to the successful implementation of three-dimensional audio systems [16]. Externalization (distinguishing a source as outside the listener's head) is a problem Begault himself has pursued [17][18]. Often though the listener is able to perceive azimuth and elevation differences, the sound appears to be very close to, or inside, the head. The perception of distance is a difficult illusion to construct. One key to successful externalization is an understanding of the role of room reflections, as noted by Hartmann [19][20]. The human brain utilizes information contained in the first few reflections from a reverberent environment to contribute to a perception of space.

Another obstacle to auralization is user-dependence of the HRTFs. There is a great deal of debate on whether better results are obtained with a "good listener's" HRTF, a "composite average" HRTF, or a "generalized" HRTF such as those developed by Wightman and Kistler [21]. A "good listener" HRTF is the measured HRTF of an individual who exhibits above-average localization ability in free-field conditions. Some listeners actually perform better with such a set than with their own measured HRTF [14]. A "composite average" is simply the HRTFs of many individuals averaged to create a single, representative HRTF. The problem with this approach is that it averages not only the common but the unique filter traits as well. In an effort to extract only the common characteristics across HRTF variation, Wightman and Kistler have performed a principal component analysis of a large number of measured HRTFs, resulting in a small set of basis functions from which HRTFs can be constructed with a high degree of accuracy. Because these functions encompass the shared features, any deviation is a result of the individual's own variation —

typically less than 5%. Despite these efforts, the fact remains that HRTF compatibility varies widely in the general population [1].

Currently, most applied research depends on the Convolvotron, developed by Foster and Wenzel at Crystal River Engineering [2]. It is the principal commercial product available for serious auralization work. It combines a control CPU and DSP convolution engine to process audio signals using a library of measured HRTFs. Up to four sources may be auralized at once. A number of smaller auralization products are also under development at CRE, with development goals similar to the goals of this project: to extract more functionality from limited computing resources (constrained primarily by the inter-sample time period) through the application of superior algorithms.

1.3 Variables in an auralization system

In an auralization system, there are many parameters which may affect the listener's perception of the sound source. In order to address these issues and provide compensation, it is first necessary to identify them. In its simplest form, an auralization system may be represented as four steps: source recording, HRTF recording (or synthesis), a convolution means, and a playback system, as shown in Figure 1.1. Each of these has a set of parameters which must be controlled and defined to achieve accurate results.

Source recording:

- Distance

The distance at which the source was recorded influences the relative SPL level as well as other parameters. Sources recorded at different distances present a difficulty during playback because they retain that discrepancy in depth. A reference distance of one meter is recommended as a standard. For directly synthesized sources (i.e. from a drum machine), the level should be set to correspond to the SPL of a similar natural source at one meter.

- Level

Here level is defined as the conversion from sound pressure levels (SPL) to digitally represented values (-32768...+32767 for a 16 bit system). It is necessary to know the conversion factor so that it may be reversed precisely for playback. Without this knowledge, a sound may be perceived as "too close" if the playback level is higher, or "too far away" if it is lower.

- Microphone response

The microphone used for recording the source will exhibit a characteristic frequency and phase response, which may be directional as well.

- Room response

The environment in which the recording is made will affect the recorded sound. Room reflections and resonant modes will force an undesired sense of environment onto the source recording. This can be avoided with direct recording of synthesized signals, or by performing recording in an anechoic chamber.

- A/D converter response

The converters necessary to move the analog signal into the digital domain for processing will also have a characteristic frequency and phase response. Fortunately, most modern converters utilize oversampling to reduce the filter constraints, resulting in passbands which are very flat across the spectrum and also exhibit linear phase. Consequently, the error added here is negligible.

HRTF recording/synthesis:

- Distance

Similar to the concerns presented for source recording, this also presents a twist; because the HRTFs are almost always recorded in pairs, the distance reference is to the center of the head. This is a problem since the distance computation for Interaural Level Differences (ILDs) and Interaural Time Differences (ITDs) is referenced to the appropriate ear rather than the center of the head. This can be compensated for in software, though a more accurate result would be accomplished by measuring the HRTFs for each ear individually, with the source located in a one meter radius from the ear. This would generate HRTFs consistent with the ILD and ITDs, and is discussed in more detail in Chapter 3.

- Level

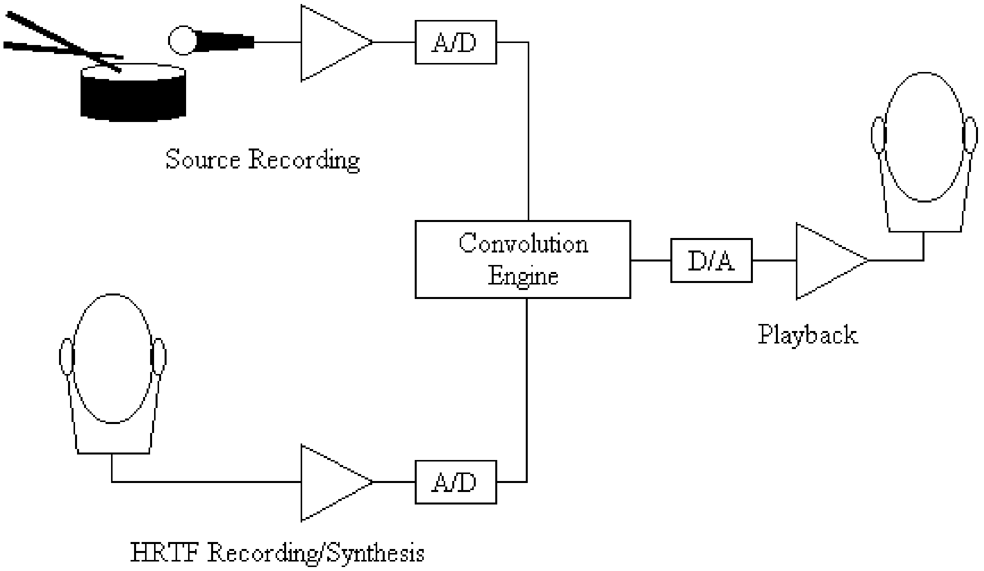

Figure 1.1: An auralization system

Again, the level of the impulse response must be specified. A sound source recorded and replayed through systems with identical gain should have the same perceived level as a natural source at the specified position. For frequency domain filtering, the "zero dB" level must be established.

- Impulse non-idealities (recording)

When HRTFs are recorded by playing an "impulse" through a speaker at various positions [13], the fact that this stimulus is non-ideal must be acknowledged. The spectral content of this impulse signal will be reflected in the HRTFs generated by it. If the impulse source is significantly non-ideal, a compensation filter should be applied to the HRTFs to remove some of the effects of the stimulus characteristics. Provided the deviations in the stimulus are consistent and well-defined, this does not present a significant obstacle.

- Microphone response (recording), room response (recording), A/D converter response

These are identical to corresponding concerns for source recording.

- Individuality of HRTFs

The unique nature of each individual's HRTF presents a considerable stumbling block to the implementation of systems for general use. The frequency and phase characteristics of the HRTF vary from person to person, much like fingerprints. Since measuring each listener's HRTFs is not currently practical, alternate solutions must be found, based on a common set of HRTFs. Methods for realizing this are discussed in Chapter 4.

Convolution:

- Windowing and phase effects (fast convolution)

When HRTF filtering is performed in the frequency domain, there are two main concerns. First, it is erroneous to apply a magnitude-only filter to perform the HRTF, as phase effects are critical for successful implementation [22]. The other matter is windowing effects resulting from the segmentation of the audio stream. Naive implementations using a rectangular (no) window will result in added clicks, pops, and discontinuities in the output stream. The use of a Hamming or Blackman-Harris window and the proper amount of overlap can greatly reduce these artifacts [23].

- Linear phase response (direct convolution)

When utilizing FIRs for direct convolution, phase response is often overlooked since most classical design methodologies assume linear phase is desired. Using directly measured impulse responses instead of synthesizing an FIR from a frequency magnitude plot eliminates the problem of attempting to reconstruct phase variations during synthesis.

Playback:

- Level

Symmetric to the need for recording level specification is the output gain; this should be a fixed gain equal to the conversion used in recording. The user should not have a volume control! If the system is engineered for accurate reproduction, the digital level to SPL conversion must complement the SPL to digital level conversion executed earlier.

- D/A converter response

Similar to the concern for A/D response, this too has become less of an issue with the advent of high-quality parts which use oversampling and digital filtering.

- Amplifier response

Non-linearities and frequency dependence in the amplifier must be accounted for and compensated. Again, problems can be drastically reduced by using high quality components.

- Headphone response

The single-driver construction of most headphones leads to a some frequency dependence, as the 10 octave range of human hearing is beyond the flat-response capabilities of most speaker elements. There will be some rolloff on low frequencies (typically below 40 Hz) as well as on high frequencies (typically 15 kHz). While professional quality headphones greatly reduce this effect and generally approximate the full bandwidth of human hearing, the fluctuation in response across the full spectrum of headphones available to the listener is too significant to ignore, and should be countered by an inverse filter prior to playback (Chapter 4).

- Head position

There must be feedback regarding the listener's head position; if the head is moved even slightly, all of the auralization information (HRTFs, ITDs and ILDs) must change to accommodate this change. Head tracking is discussed in detail in Chapter 8.

- Expectation

Expectation plays a very large role in human sensory perception in general; as it applies to auralization, listeners will rely on visual cues, their knowledge of the present environment, and memory of recent events to assist in localization. Augmenting auditory stimulation with stereographic visual representations of the environment provides enhanced results. If a listener sees a saxophone in front of her and to her right, she will not confuse the position of a saxophone sound as coming from behind. The problem of front/back reversals disappears as the listener can associate the sounds with items in view (in front) or not in view (in back).

Providing a means for tracking and compensating for head movement also contributes to resolution of these reversals. By combining movement and memory of recent stimuli, the listener can "triangulate" to determine the true location of the source. Since motion affects front source interaural differences contrary to rear sources, the direction of change is sufficient to indicate the hemisphere in which the source is located. This method also has the advantage of relying on the fairly robust ITDs and ILDs, rather than the delicate HRTFs, to provide front/back differentiation.

For more discussion of integration with visual displays and head tracking, see Chapter 8.

CHAPTER 2:
OVERVIEW

2 OVERVIEW

2.0 Objective

While visual three-dimensional synthesis for virtual environments has seen a great expenditure of time and effort, audio spatialization is in its infancy. Present systems incorporate crude algorithms running on several high-powered DSPs to accomplish simple placement of a single sound source. The goal of this project is to develop refinements to the current state of the art, reducing the temporal computation demands placed on processing systems.

A software implementation of current algorithms will be developed to serve as a baseline. Several points of consideration for increasing realism in auralization systems will be identified and implemented. A module will then be added which incorporates bandwidth analysis to identify and eliminate unnecessary computation. These refinements demonstrate a marked savings in computation time for auralization processing.

2.1 Facilities

The primary development environment for this project was a 486dlc/25 personal computer, operating under DOS 5.0 and Windows 3.1 with 5MB of RAM and 440MB of disk storage. A Media Vision Pro Audio Studio 16 soundcard was used to record and playback Microsoft Type I Wave (.WAV) format sound files through JVC HA-D500 headphones. Turtle Beach WAVE for Windows was utilized extensively for viewing the .WAV files, and MATLAB for Windows was used for algorithm development and graphical analysis. Software development was completed with the GNU C++ compiler version 2.5.7, which offers true 32-bit executables (for speed) and a flat memory model (avoiding DOS's 640k memory restrictions), as well as enhanced portability across platforms. A Sun workstation was also used at various stages, both for file transfer and for fast execution of tested code segments.

Sound samples used in development were recorded directly to the PAS 16 soundcard from an Alesis D4 drum module. Voice samples were recorded from a pre-recorded CD, again directly to the PAS 16. Sine waves and noise samples were synthesized in software.

2.2 Methods

Available facilities rendered the development of a real-time auralization system unreasonable; instead, a "preprocessing" software system was created. Sound sources were recorded (in mono) using a PC soundcard, and were stored in Microsoft Type 1 .WAV format files. An "auralized" stereo .WAV file was generated by invoking one of the programs developed in chapters 4 and 5, along with a desired source position. This .WAV file was then ready for playback though the soundcard and headphones.

The provided source code was written with portability in mind. Every effort was made to avoid non-standard C++ functions and conventions. As a result, executables may be generated on a wide variety of machines. The code has been tested on 486-based systems, and a Sun workstation under UNIX.

Once an executable has been compiled from the source code, the command for processing a file from the system prompt (independent of computing platform) is

```
auralize input.wav output.wav [θ] [ϕ] [r]
```

where

`input.wav` is the name of the mono source file
`output.wav` is the name of the stereo output file
`[θ]` is the desired azimuth in degrees [default 0]
`[ϕ]` is the desired elevation in degrees [default 0]
`[r]` is the desired distance in meters [default 1]

2.3 Definition of the coordinate system

When working with auralization concepts, it is much more intuitive to work in spherical rather than Cartesian coordinates. For clarity, the specific definition of this coordinate system is described here.

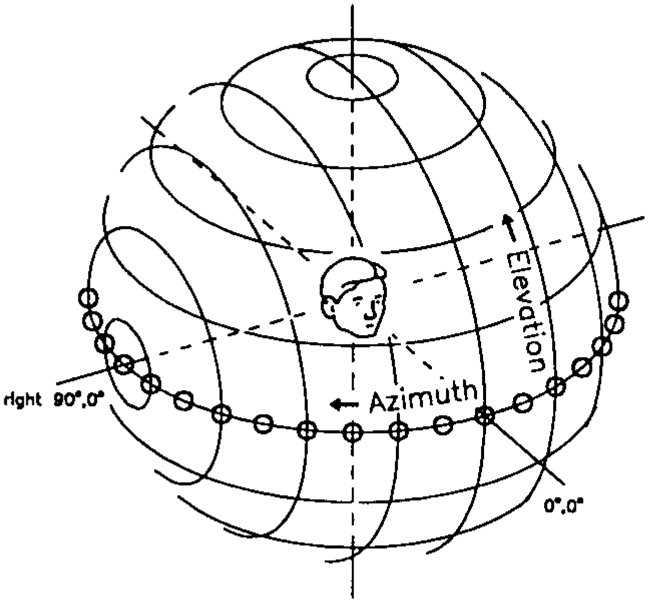

Figure 2.1: Diagram of spherical coordinate system
(from Makous & Middlebrooks [24])

All locations (source and ear positions) are referenced to the center of the head, and are given as a triplet of azimuth (θ), elevation (ϕ), and distance (r).

Azimuth $(-180° \leq \theta \leq 180°)$ is defined as the deflection from front center (0°) in the horizontal plane, with positive angles defined to the right. Therefore, 90° is directly to the right and -90° is directly left. Positions directly behind the head may be described as either 180° or -180°; the two are functionally equivalent

Elevation $(-90° \leq \phi \leq 90°)$ is defined as the deflection from horizontal (0°), with positive values defined above and negative below. Therefore, 90° is directly overhead and -90° is directly below. Angles greater than 90° are redundant and are not used.

Distance $(0 < r < \infty)$ is defined in meters from the center of the head. The reference distance for all level calculations is one meter.

For many of the manipulations involved in auralization, it is necessary to compute the distance between two arbitrary points in space — the ear and the source. While the distance formula for three-dimensional Cartesian coordinates is well known, one must be derived for our spherical system. This derivation has the following result:

Given two points in spherical coordinates (θ_1, ϕ_1, r_1) and (θ_0, ϕ_0, r_0):

$$d = \sqrt{r_1^2 + r_0^2 - 2r_0 r_1[\cos\phi_0 \cos\phi_1 \cos(\theta_1 - \theta_0) + \sin\phi_0 \sin\phi_1]}$$

This distance computation is generalized and implemented as a function for flexibility. Within the context of the auralization programs developed here, it is used for measuring the distance from each ear to the source; this distance is then used to compute the ILD and ITD for that ear. The distance is recomputed each time the source changes location, for non-static sources.

The distance calculation is also used to remove intrinsic ILDs and ITDs from HRTFs recorded using a centro-cranial origin (see Chapter 3).

CHAPTER 3:
SPATIAL PLACEMENT USING ILDS AND ITDS

3 SPATIAL PLACEMENT USING ILDS AND ITDS

3.0 The role of interaural level and time differences

By examining the physics underlying the travel of sound waves in air, two location cues become apparent. As sound radiates outward from a source, the power (and resulting perceived level) drops with increased distance. If the distances to each ear are unequal, an interaural level difference (ILD) will be noted. Likewise, since the speed of sound in air is finite, sound which must travel different distances to each ear will arrive at different times. This is referred to as an interaural time difference, or ITD.

3.1 Calculating the ILDs

The basis for calculating ILDs is the inverse-square law [25]:

$$I = \frac{W}{4\pi d^2}$$

where

I is the sound intensity in watts per square meter,
W is the sound power of the source in watts,
d is the distance from the source in meters.

The law assumes that the source is a point source and is radiating uniformly into a free field. The amount of power flowing through a given solid angle is constant, and allows us to equate the sound power at two radii:

$$W_1 = W_0$$

$$(I_1) \times (4\pi d_1^2) = (I_0) \times (4\pi d_0^2)$$

rearranging, we get

$$\frac{I_1}{I_0} = \frac{4\pi d_0^2}{4\pi d_1^2}$$

$$= \frac{d_0^2}{d_1^2}$$

which states simply that the intensity of sound in a free field is inversely proportional to the square of the distance from the source. While intensity is difficult to measure and manipulate, sound pressure level (SPL) is relatively easy to deal with. Since SPL is proportional to the square root of the intensity, the inverse-square law reduces to

$$\frac{L_1}{L_0} = \frac{d_0}{d_1}$$

where *L* is the sound pressure level.

3.2 ILD implementation

The necessity of computing the distance from each ear to the source is now apparent. The distance formula provides the distance *r* for the left and right ears. It is tempting to simply use the ratio between the two to find the ILD. In fact, many systems do process ILDs in this manner. This technique is adequate for static (non-moving) sources. However, if the sound source is moving radially with respect to the listener, it is necessary to add another level compensation for the change in distance (the source will appear louder as it approaches the listener). Another possibility is the use of impulse responses which implicitly include the level difference. Unfortunately, the source distance is then dictated at the time of HRTF generation and cannot be accurately compensated by a simple level shift, since the ILD may change at a different rate from the absolute level. Therefore, it is more reasonable, from a systems perspective, to adjust levels for each ear individually according to some reference distance at which the sound sources have been recorded. This results in automatic generation of ILDs and attenuation of sounds as they travel further from the listener's position, and maintains flexibility to specify distance parameters at run-time.

The level for each ear is adjusted by the reciprocal of the distance to that ear, in meters:

$$Gain_{R,L} = \frac{1}{d_{R,L}}$$

3.3 Calculating the ITDs

ITD calculation depends on the speed of sound in air, and the distance traveled. The speed of sound in air can be approximated as [26]:

$$v = 331 + 0.6T$$

where

v is the speed of sound, in meters per second,

T is the ambient temperature, in degrees Celsius.

This approximation holds true for conditions near room temperature and pressure. The time delay from the source to the ear is simply the distance divided by the speed of sound.

3.4 ITD implementation

As with ILDs, it is tempting to compute the ITD directly using the difference in path lengths to the right and left ears or by incorporating it in the HRTF impulse response. Similar problems arise. If the ITD is computed directly, a separate overall delay must be computed for the time it takes sound to reach the first ear. This is critical for integration with visual elements. A sound synchronized to a visual event perceived as several hundred meters away should have an inherent delay; the presentation of sound with the correct ITD but incorrect absolute delay introduces an anomaly which inhibits the willing suspension of disbelief. Alternately, if the ITD is intrinsic to the HRTF it reduces accuracy from the filtering function; any delay simply fills the beginning of the lagging ear's FIR with zeros, reducing the effective filter length without reducing computational load. A better approach is to compute the delay separately for each ear; ITD's and absolute delay are computed in a single operation, and both ears benefit from full-length FIR filters.

Because it is impractical and unnecessary to incorporate temperature variations in this project, a static sound velocity of 346 m/sec was selected for the purposes of computation. The resulting formula is:

$$t_{delayR,L} = \frac{d_{R,L}}{346}$$

where t is the delay in seconds. The delay in terms of samples is given by multiplying by the sample rate f_s:

$$z_{delayR,L} = \frac{d_{R,L}}{346} f_s$$

3.5 Compensation for intrinsic ITDs and ILDs

HRTFs recorded with traditional methods (as in Wightman and Kistler [13]) incorporate both a time delay and a level change, as the impulse source used for measurement is located at some distance d from the center of the head. To allow source positioning within this radius and a more generalized algorithm, it is necessary to remove these biases.

Removal of intrinsic ILDs is relatively simple, and involves computing the equivalent level shift for a source at the HRTF radius (1.43 meters for the set used here) and given angular displacement. This level shift is then divided out of the ILD shift computed in section 3.1.

ITDs present a more significant challenge, since the delay line must be modified to be anti-causal. That is, the system must be aware of samples both before and after the current sample being processed. The maximum number of anticausal samples needed is derived by calculating the longest distance traveled by the HRTF measurement source. This is equal to the HRTF recording radius plus half the interaural spacing. For this setup, this is 1.43+0.06 = 1.49 meters. Dividing by the speed of sound gives the delay in seconds and multiplying by the sample rate provides the maximum embedded delay in samples:

$$z_{HRTFmaxdelayR,L} = \frac{1.49}{346} f_s$$

For our 44.1 kHz sample rate, the maximum delay embedded in the HRTF equals 184 samples. Once this anticausal z-buffer offset is established, the actual HRTF intrinsic delay for a specific source position must be removed; this is accomplished by subtracting the HRTF equivalent distance from the current source-ear distance before the ITD calculation described in section 3.4. Note that negative distances are possible, and resolve to negative delays — hence the need for an anti-causal system.

CHAPTER 4:
INTRODUCING THE HRTF

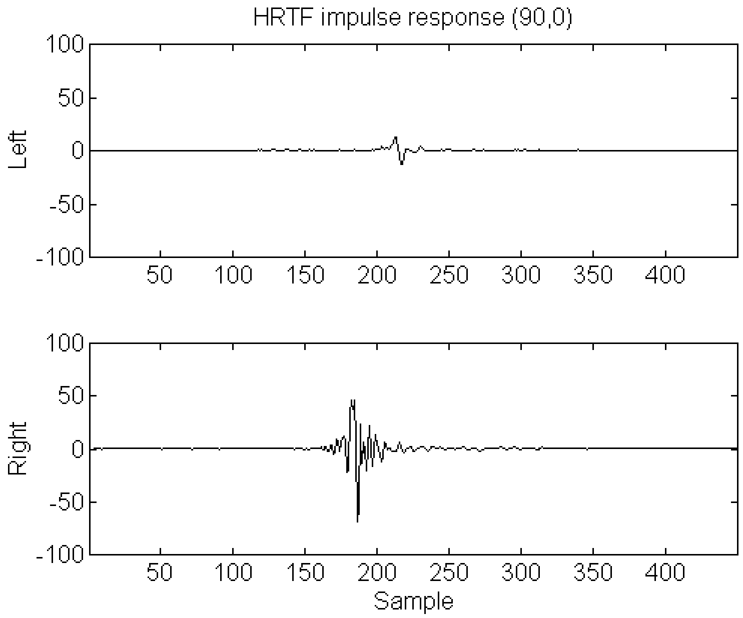

Figure 4.1: Example impulse response for 90° azimuth, 0° elevation from SDO set.

4 INTRODUCING THE HRTF

4.0 Derivation of the HRTF

While ILDs and ITDs play a dominant role in presenting azimuthal information, there are limits to the information they carry. The ILDs and ITDs for any equidistant point are equal; this is most problematic in contributing to front/back reversals. Listeners cannot distinguish between sounds in front of the head and the "mirror image" position behind the head (i.e. +30° and +150°) without additional information.

The head-related transfer function is the source of this data. It accounts for diffraction around the head, reflections from the shoulders and most significantly, reflections from the pinnae. It is these structures of the outer ear which act as a direction-dependent filter to add elevation and front/back information to the sound signal each eardrum receives. Unfortunately, the physical composition of the pinnae varies widely across the general population. As a result of this diversity, HRTFs are also quite different for each individual.

4.1 Generalization

Since the measurement of HRTFs is a time-consuming and difficult practice, it is impractical to construct a full set of them tailored to each user of an auralization system. Instead a generalized set, or possibly a choice of generic HRTFs, is implemented. There are numerous methods of arriving at these default sets. One may choose the HRTF of an individual who has demonstrated above-average localization ability, or possibly an average taken over many listeners of different background. Wightman and Kistler have performed extensive HRTF measurements and have recently proposed a set of "principal components" from which a generalized set of HRTFs may be constructed for an arbitrary source position [21]. The consensus of experiments using generalized HRTFs is an increase in front/back and up/down reversals over free-field or individualized transforms. Hence, there is a trade-off for the simplicity of using a single set of filters. Fortunately, localization ability is usually otherwise very accurate when using generalized HRTFs. Many methods have been developed for reducing the frequency of these reversals, including the addition of head-tracking, visual stimuli, and the addition of synthetically generated reverberation; these are discussed in chapter 8.

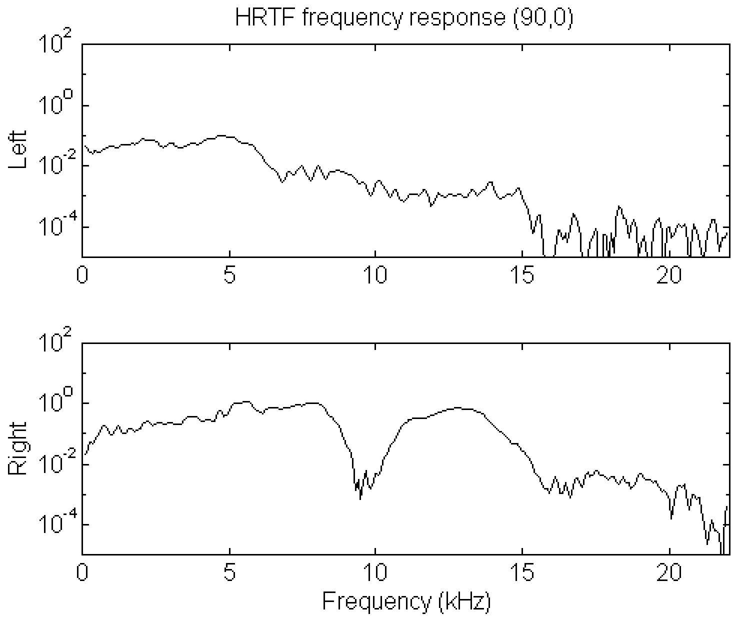

Figure 4.2: Example HRTF magnitude for 90° azimuth, 0° elevation from SDO set.

4.2 HRTF Implementation

Because the HRTF represents a filter function to be applied to the source signal, a convolution is necessary. There are two methods to achieve this: first, a direct convolution may be performed with measured impulse responses (Fig. 4.1) implemented as an FIR.

Alternately, since convolution in the time domain is equivalent to multiplication in the frequency domain, an FFT may be performed, followed by multiplication by the desired filter response (Fig. 2), followed by an inverse FFT [27][28]. It is important, however, to realize the role of temporal and phase cues in the HRTF; it is not accurately represented by a magnitude (real-valued) filter [22]. The phase information must be maintained through the use of complex FFTs and a complex frequency domain filter.

The first approach, direct convolution by FIR, was the method implemented in this project. FIR coefficients may be directly retrieved from measured impulse responses utilizing either human subjects [13] or a dummy head [29]. For the purposes of this project, Wightman and Kistler's HRTFs for a representative subject (identified by her initials, "SDO") were adapted for use in the FIR stage of the auralization process. These impulses are 512 samples in length, sampled at 16 bit, 50 kHz resolution and are available by ftp in a columnar ASCII text format [15]. In order to accommodate the limitations of the soundcard used for sample recording and playback, these HRTFs were downsampled to 44.1 kHz.

Due to this sample rate conversion, the impulse responses were reduced to 450 samples per ear. The SDO set was measured at 15° intervals in a full 360° rotation of azimuth, and in 18° intervals in a 90° angle of elevation, from 54° to -36° [13]. These 144 HRTF pairs were stored as a sequential file of raw sound samples with 450 stereo 16-bit samples apiece (SDO44.DAT). At the beginning of execution, the HRTF file was loaded into an array to allow convenient addressing of the 144 positions represented. During computation, the HRTF most nearly corresponding to the specified azimuth and elevation was used. There was no provision in this implementation for interpolation of "in-between" angles.

4.3 Headphone system response compensation

The SDO set of HRTFs made available by Wightman and Kistler [15] have been pre-compensated for "high-fidelity headphone response". A possible future enhancement would be the inclusion of a headphone calibration file, allowing the user to specify the make and model headphone in use. An appropriate compensation filter would be added before playback, to nullify the effect of using different headphones.

4.4 Error identification for HRTFs recorded as a pair

It is common practice to measure, store, and recall HRTFs as a stereo pair. They are recorded with a source displaced along a sphere centered on the middle of the head. The resulting impulses are stored as a pair referenced to the coordinates of the source relative to the center of the head. A simple example illustrates the error inherent in this method; an error which has simply been ignored in existing systems.

Let us examine the behavior of a sound source at 0° azimuth, 0° elevation (directly in front of and level with the listener). When the HRTFs are recorded at this position at a reference distance of 1 meter, it is evident that the direction vector from each ear is actually $\tan(\frac{0.06m}{1.0m})$ = 3.5° from center (assuming an interaural spacing of 12cm). The left ear is actually measuring the HRTF for the (θ, ϕ, r) triplet (-3.5°, 0°, 1.002m) and the right is measuring the HRTF for (3.5°, 0°, 1.002m). While the distance variation from the 1 meter reference is negligible (one sample at 44.1 kHz is equivalent to 0.0078 meters assuming sound travels at 346 m/sec), the angle variation is not. As the source moves closer, the angular discrepancy increases. At 0.5m, the actual angle between the left ear and the source is -6.84°, with the right ear at +6.84°.

This error only manifests itself when the source is being placed inside the radius at which the HRTFs were recorded. It increases as the desired source placement becomes closer to the center of the head, to a maximum of approximately 45° *for each ear* at the surface of the head.

The solution to this problem is to break the artificial link between right and left side HRTFs. There is no physiological reason to link them as a stereo pair; it is only to simplify HRTF measurement and playback. To avoid this error, HRTFs should be recorded individually for left and right ears, with the source displaced along a sphere centered on the ear, not the center of the head. During playback, individual left and right HRTFs should be chosen based on the angle between the source and the corresponding ear. The latter is the most important consideration: individually selecting left and right HRTFs for playback (from a stereo-recorded set) will limit the error to ±3.5° — the error incurred during recording.

The angle to the source from each ear can be computed as follows:

$$\varphi_{R,L} = \tan^{-1}\left(\frac{r\sin\theta - d_{ear}}{r\cos\theta}\right)$$

Where d_{ear} is equal to the distance from the center of the head to the ear (one half the interaural spacing); positive for the right ear, negative for the left.

CHAPTER 5:
FIRST-CUT OPTIMIZATION: BAND-LIMITED SOURCES

5 FIRST CUT OPTIMIZATION: BAND-LIMITED SOURCES

5.0 Real-world sources

By examination, it is evident that in current auralization algorithms, the computation time is independent of the characteristics of the source. A complex sound requires the same processing time as a simple sound, or even a period of silence. It is here we find the first optimization to explore: source-dependent algorithms.

In auralization systems, a great deal of effort is given to accurate localization of sound sources. Rarely, however, are such systems called upon to place a broad bandwidth sound. When such occasions do occur, it is questionable whether pinpoint spatial accuracy is necessary. As an example, consider that most naturally occurring sounds *which emanate from a single point in space* (speech, animal vocalizations, etc.) are limited to a relatively small bandwidth, typically a few thousand Hertz. Sources which maintain large bandwidth for an extended period of time (such as an orchestra) produce sound from numerous spatial positions, presenting a diffuse general perception of location. There are some sounds, such a drum, which nominally have a large bandwidth yet present a localized origin. On further examination, however, it is apparent that the wide spectral content attributed to such sounds are primarily due to the impulse caused by a sharp attack transient. Once this transient has passed, the sound settles into a more band-limited range. The following algorithm is a first attempt at dynamic allocation of resources to match the spectral content and psychoacoustic "importance" of source signals.

5.1 Bandwidth identification

There are several methods available to identify spectral content in the upper frequency ranges. One direct approach is to implement an FFT and evaluate the magnitude response in the high frequencies. Another approach is to use a high-pass time domain filter, and evaluate the magnitude of the remaining signal. For this project, the FFT approach was adopted. A 256-point FFT was computed every 4096 samples to re-evaluate the bandwidth of the signal. The computation interval of this evaluation is open to manipulation; it exerts an direct effect on the execution time of the algorithm. However, decreasing the rate at which frequency snapshots are taken will reduce the system's ability to react quickly to changes in spectral content.

5.2 Algorithm modification

Once the bandwidth of the signal is known, the algorithm switches between two states. The first state, when the bandwidth of the signal is high, operates at a full 44.1 kHz sampling frequency but uses a 225 point approximation of the HRTF, effectively halving the "accuracy" of the filter, while also reducing our computation by half. The second state, for low bandwidth signals, cuts the sampling frequency in half (only processes every other incoming signal), which results in a 225 point sub-sampled HRTF. The output uses linear interpolation to upsample back to the original frequency. Because the signal was determined to have little high-frequency content, the aliasing introduced by this sample rate conversion was negligible.

The overall result of this optimization is a reduction in computation by nearly a factor of four (halving the length of an FIR quarters the time required to process it) at a cost of a FFT performed at adjustable intervals. Examining the complexity of the algorithms:

A straight N-point FIR convolution is of order

$$\mathrm{o(N\text{ - }point\ FIR)} = N^2$$

Cutting the resolution in half results in order

$$\mathrm{o(\tfrac{N}{2}\text{ - }point\ FIR)} = \left(\frac{N}{2}\right)^2 = \frac{N^2}{4}$$

The M-point FFT adds a *MlogM* term every 4 cycles, so the total order is

$$\mathrm{o}(\mathrm{Optimization}) = \frac{N^2}{4} + \frac{M \log M}{4}$$

Since M and N are known for this case (M = 256, N=450), we can directly compute the approximate number of operations for each algorithm. For the straight convolution, the number of operations needed to process 450 samples is

$$450^2 = 202{,}500$$

For the optimized algorithm, the necessary number of operations drops to

$$\frac{450^2 + \frac{450}{2}\log_2 \frac{450}{2}}{4} \cong \frac{202{,}500 + 1758}{4}$$
$$\cong 51{,}064$$

which is very nearly a factor of four. The implications of this savings are tremendous; four times the number of sources may be computed with the same resources, or three indirect reflections may be computed for each source (see section 8.2).

CHAPTER 6: RESULTS

6 RESULTS

6.0 Auralization Caveats

A number of caveats have been identified and addressed throughout this paper; items which are easily forgotten when constructing an auralization system, but are essential for its proper operation. A summary of these findings follows:

The need for specification at every level of the auralization process was addressed in section 1.3, with a list of the pertinent variables in an auralization system and the effect each has on the final result. While some of these issues (such as headphone response and head movement) have been addressed in present systems, many have escaped mention in the literature. The necessity of meticulously specifying the recording conditions for both HRTF and source had not previously been brought to light.

The problems introduced by ILDs and ITDs intrinsic to the HRTF were discussed in Chapter 3, along with means for a solution. Methods for the removal of these imbedded characteristics were developed and implemented.

The HRTFs were examined in Chapter 4, including a suggestion for a new method of recording them. Problems related to the artificial pairing of individual ear responses were observed and again, a solution proposed.

These items are all results of the research work performed on this project, though in a somewhat different way from the audible products of the software. They are however, a significant consequence of this effort.

6.1 Baseline — ITD, ILD, HRTF

The conventional implementation presented here successfully generates source positioning for nearly all angles, bounded only by the limitations of the SDO HRTFs. These restrictions affect only large deviations in elevation, and are not a primary concern. The approximation used for these unusual cases is sufficient.

A number of different sources were used, varying from percussive drum sounds to human voice. Figure 6.1 shows a raw sample of the word "sound" spoken by a male voice, before processing. The sound was recorded as a mono, 16-bit, 44.1 kHz file. Figure 6.2 is the raw sound's spectral content, with the confidence interval indicated as well. Spectral content was analyzed with an FFT using a Hanning window and 50% overlap. An examination of the spectral content shows that at one quarter the sampling frequency (11 kHz), the signal content is down to -60 dB ($20\log_{10} 10^{-3}$). The inherent signal-to-noise ratio (SNR) of the soundcard used to record and play back these samples is approximately 62dB. Therefore, at 11 kHz, this voice sample has effectively vanished into the noise floor. This supports the statements made in Chapter 5 regarding the bandwidth-limited nature of commonly used signals.

This sample was then processed with the conventional auralization algorithm, for a source position of 90° azimuth, 0° elevation, 1 meter distance. The resulting stereo .wav file is displayed in figure 6.3. The power spectrum for this file is also given, in figure 6.4. The effects of the HRTF filtering are visible in the difference between the left and right channel spectra. The ILD is evident from the time domain graph, and zooming in on a section of the signal displays the ITD, as in figure 6.5. It is most noticeable by comparing the position of the trough just before 6300 on the right channel with the same trough just past 6300 on the left.

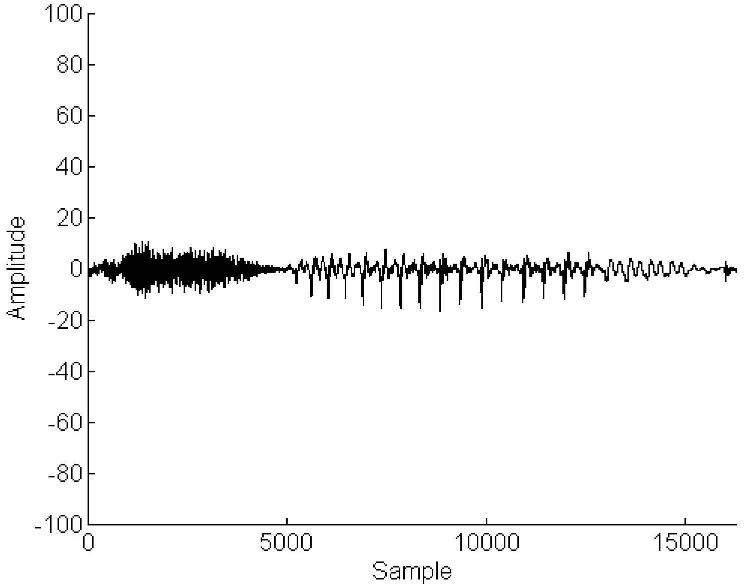

Figure 6.1: Raw sample of "sound" spoken by a male voice.

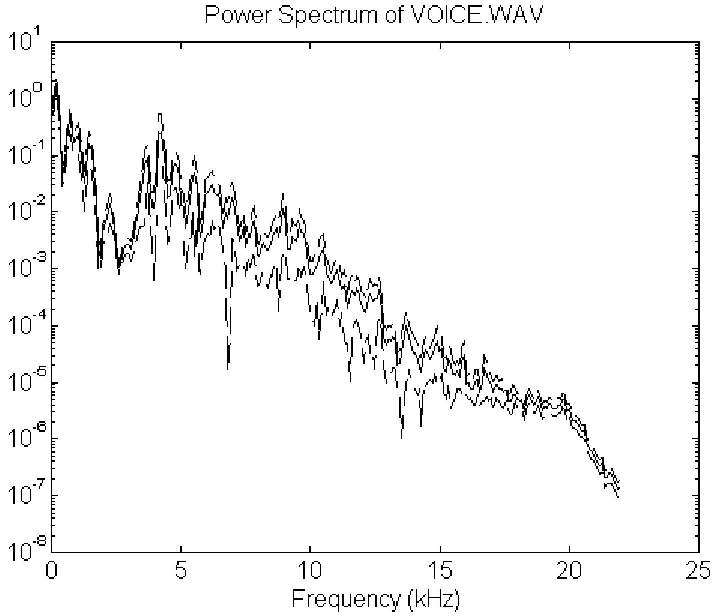

Figure 6.2: Frequency spectrum of "sound" spoken by a male voice (before processing).

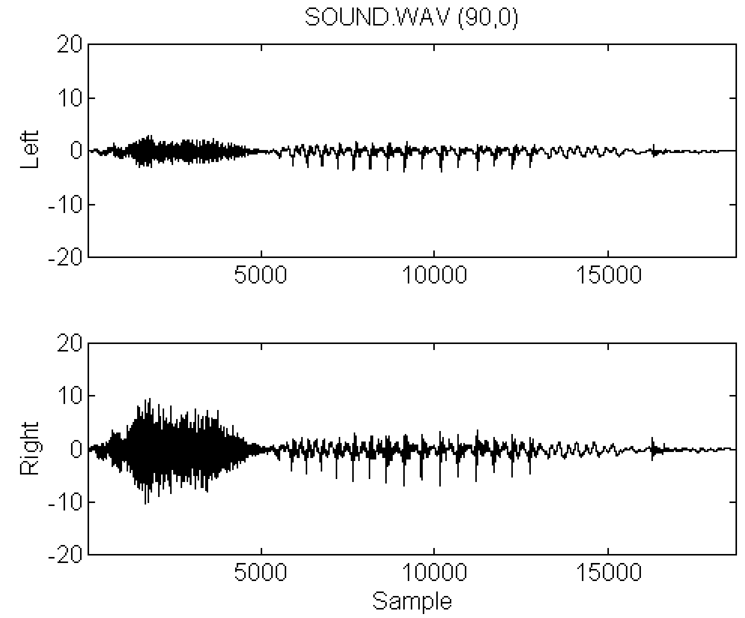

Figure 6.3: Male spoken "sound" after auralization placement at (90,0,1).

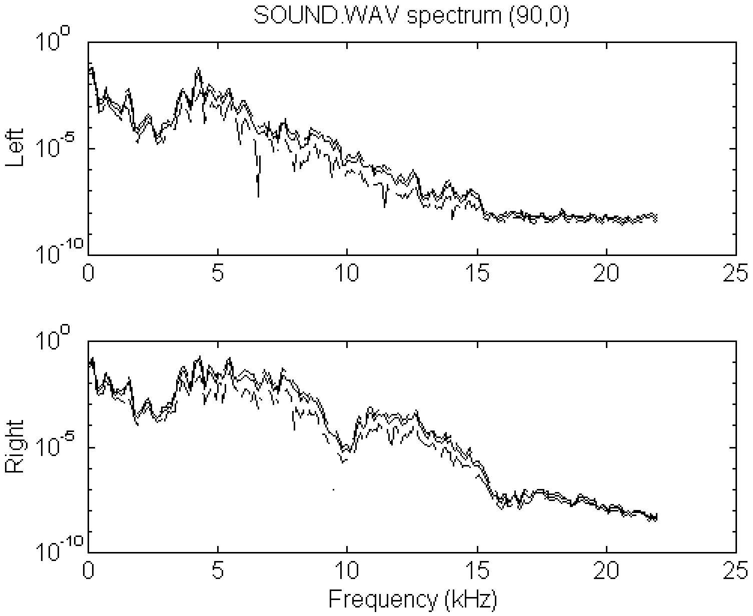

Figure 6.4: Spectral content of "sound" after auralization to (90,0,1).

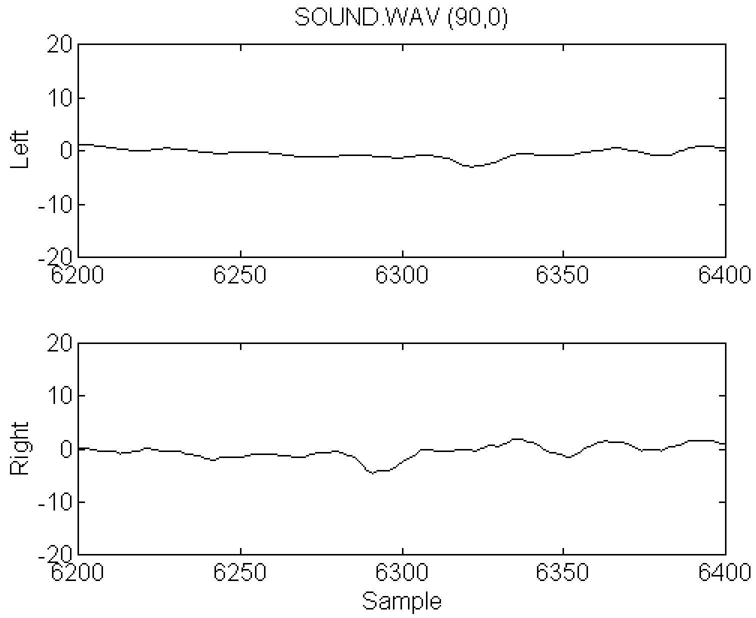

Figure 6.5: A close-up of the middle of "sound", exhibiting interaural delay.

6.2 Bandwidth limiting

The results from the optimized code are visually very similar to those from the classic algorithm. Figure 6.6 shows the time-domain response, and figure 6.7 is the spectral power of the "sound" sample after processing by the optimized algorithm. The aliasing visible about 11 kHz is the result of using a simple linear interpolation to upsample the output of the program. A short FIR low-pass filter could be added to further reduce these aliased frequencies, if necessary. The aliasing visible here contains little power and will likely have a negligible effect on the perceived sound.

To validate this approach, it is necessary to look at the overall effect of the optimizations; this is performed in the next chapter.

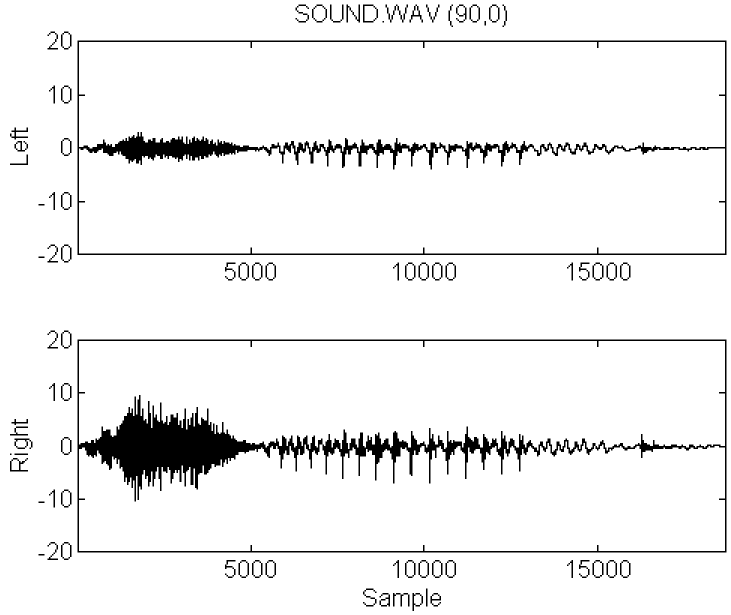

Figure 6.6: The same voice sample ("sound") processed with optimized code.

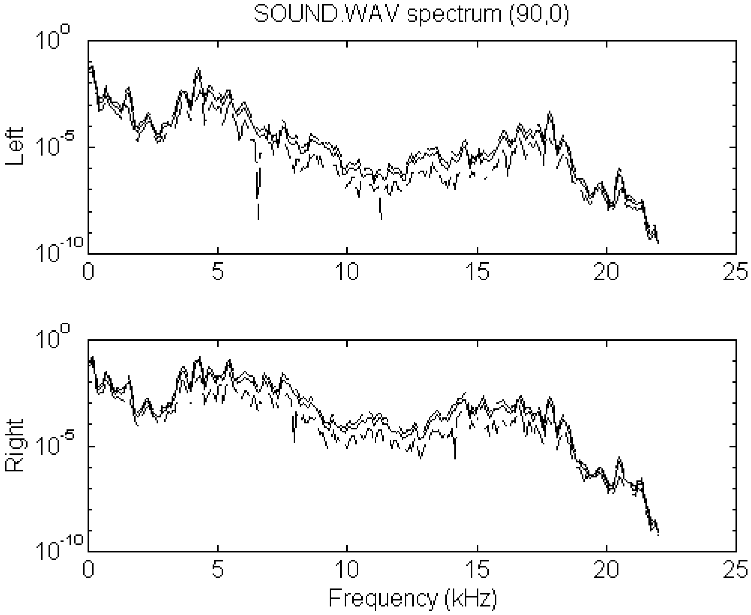

Figure 6.7: Frequency response of "sound" using optimized algorithm.

CHAPTER 7:
VALIDATION

7 VALIDATION

7.0 Residue measurement

The optimizations presented in Chapter 5 have been shown to greatly reduce the computation time necessary for auralization processing, but at what cost? To examine the effects of the optimization, it is useful to employ a residue evaluation. The residue is simply the difference between the original (conventionally auralized) result and the result generated by the optimized algorithm. This now contains only those signal features which have changed.

Figure 7.1 is a graph of the residue for the spoken word "sound". The only place where there is significant change is at the beginning of the word, at the sibilant "s". This is to be expected somewhat, as the "s" is a broadband sound closely related to white noise. The amount of residue present during the "s" may be symptomatic of a non-optimal threshold for switching between the two modes of the optimized algorithm; it may indicate that the algorithm did not switch to the "high bandwidth" mode at the beginning of the word. However, the lack of appreciable residue during the rest of the word does demonstrate the validity of the technique for band-limited signals.

An examination of the spectral content of the residue (figure 7.2) reveals several items of note. First, the spectral energy of all components removed is very low, below -60dB ($20\log_{10} 10^{-3}$). Second, there is aliasing about 11 kHz. This was expected, as a side effect of reducing the sample rate. Since the source was already determined to have very little power above 11 kHz, the removal of more of these energies is inconsequential.

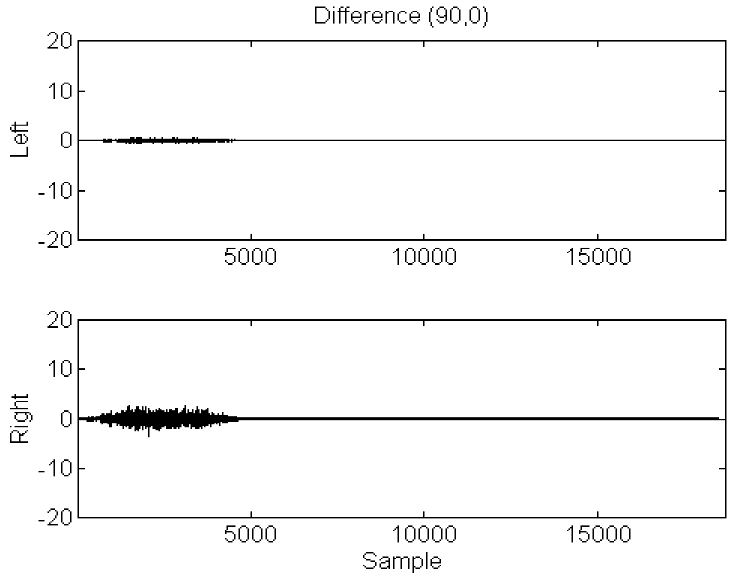

Figure 7.1: Difference signal between original and optimized algorithms for "sound".

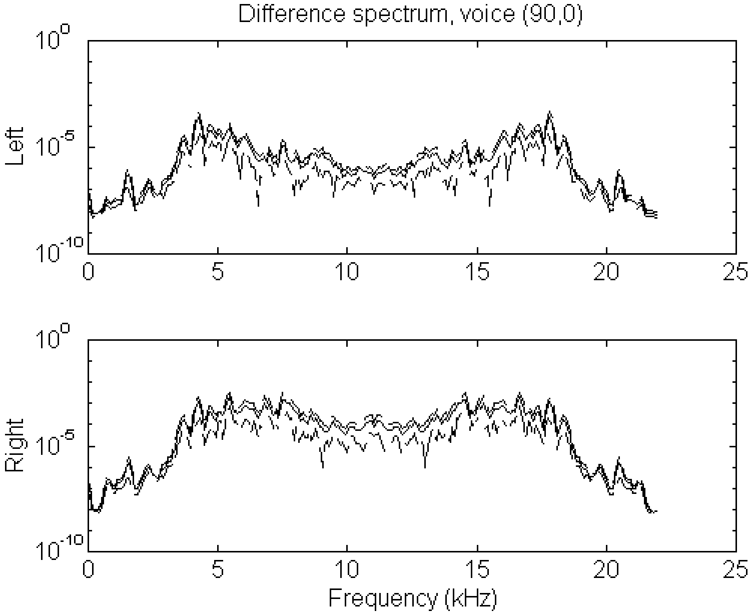

Figure 7.2: Spectral content of difference signal.

CHAPTER 8:
DIRECTIONS FOR FURTHER STUDY

8 DIRECTIONS FOR FURTHER STUDY

8.0 Further code refinements

While the algorithm utilized in this project incorporates several improvements over conventional programs, there are additional enhancements which could increase performance but were not implemented because of time limitations. These include a provision for specifying and processing moving sources, and interpolation at intermediate HRTF positions.

One method for supporting moving sources would involve a separate file providing a "script" of the object's motion, given as a description of the object's state at various points in time. The state includes the x, y and z positions of the source as well as the time derivatives dx, dy and dz (the directional velocities). The file need only contain entries at the appropriate times when the velocity vector changes or if the source makes a discontinuous jump in its path.

The coarse sampling of the HRTFs used (15° azimuth, 18° elevation) is a cause for some concern. An algorithm to perform even a simple linear interpolation for intermediate angles would increase the accuracy of the HRTFs used and prevent any sense of discrete change with moving sources.

The linear interpolation used during sample-rate conversion of the HRTFs is also a place for improvement. Though the symptoms of this simplification are relatively minute (they are correlated to the degree of rate change, which is only 12% from 50 to 44.1 kHz.) a first-order interpolator does add some distortion, most notably aliased terms in the high frequency ranges [27]. Since the sample rate conversion of the HRTFs is performed only once and has no real speed limitation, a true oversampling/decimation conversion is feasible and will be implemented in the future.

The assumption of planar wavefronts (free-field conditions) has been made throughout auralization systems. Unfortunately, this is not necessarily the case. A point source located near the listener will actually present a spherical wavefront, which may not reflect off the pinnae in the same manner as a planar wavefront. The effect is, however, very slight for reasonable source distances from the ear, due to the relatively small solid angle formed by the surface area of the ear. It is mentioned here only for completeness; the computation for accurate modeling of spherical wavefronts is prohibitive, particularly in light of the small magnitude of this effect.

8.1 Indirect reflections

Externalization is a considerable problem in auralization; it is often difficult to generate sounds which appear to originate at any large distance from the head. Much of this difficulty may be attributed to the brain's reliance on indirect reflections to determine distance [17]. When a sound source is in a physical room, the human localization system determines some information from the initial (direct) sound, but also uses the reflected (indirect) sound from the walls to assimilate information regarding the environment and the relation of the source position to that environment.

The use of artificial reverberation with ray-traced early reflections has been shown to increase perceptions of auditory distance [18][29]. In these systems, the specifications of the room were entered into a ray-tracing software package, which computed the positions of "phantom sources" to represent the first few early reflections. Ray-tracing works by assuming sound travels in a direct line and exhibits specular reflection. While this is only a first approximation for sound waves (ray-tracing originated in the light domain, where it is a more realistic representation), it can produce a fairly accurate characterization of the direction of sound reflected off the room boundaries. These reflections were represented by "virtual" or "phantom" sources located along the angle of incidence at a distance equivalent to the total path length of that reflection. The phantom sources were then processed using conventional auralization techniques, resulting in signals comprised of a combination of direct and indirect sound, all of which contained location and environmental cues similar to those found in a physical environment.

8.2 Head tracking

It is unreasonable to expect auralization systems to perform at par with nature when one of the primary localization cues is removed. Yet without a head-tracking system to compensate for head movements, this is the result. The innate human response to an unexpected sound is to turn towards it; it is in the movement of the head that critical and incontrovertible information about location is discerned. While

HRTFs do contribute somewhat to front/back differentiation, they are very frequency dependent. If the source does not contain the particular frequency affected by the difference in front and rear HRTFs, these cues simply do not exist. Instead the brain uses memory and comparison to discern hemispherical location. For a source in front of the head, clockwise movement (positive azimuthal deflection) will result in a decrease in distance to the left ear and an increase to the right (along with the appropriate ITD and ILD changes). A source in the rear will exhibit the opposite behavior; clockwise rotation will result in an increase in distance to the left ear and a decrease to the right. These distance-related cues are robust and rely solely on the laws of physics, not on the characteristics of the source.

The addition of a head-tracking system to constantly monitor head position and update relative source position would greatly increase the success of any auralization system, particularly with regard to front/back reversals. There are numerous options available for head-tracking, ranging from highly accurate multi-thousand dollar systems to low-budget home-brew devices. As the technology matures, high quality systems will become reasonably available.

8.3 Virtual reality

A logical extension of providing a simulated aural environment is to provide a simulated visual environment as well. The generalization of this is the generation of a total simulated sensory environment, or "Virtual Reality". The inclusion of artificially generated visual stimuli with a simulated aural environment allows the participant to become fully immersed in the generated world, and contributes to the principle of "willing suspension of disbelief." This principle states simply that apriori knowledge of the true physical surroundings creates an expectation which greatly inhibits acceptance of the artificially generated sensory cues, since they contradict this expectation. A preponderance of artificial cues — such as combining aural and visual stimuli — can override the natural tendency to remain "in the real world".

Without a combined sensory input, good results are difficult at best. Often the success of such a system depends a great deal on the imagination of the listener. As an example, take a listener seated in a small classroom. If presented with a simulated aural environment of a fighter jet, he must first overcome the visual stimuli which are in overwhelming conflict. In a joint system, presenting visual and aural stimuli and incorporating head-tracking "removes" the user from the room. If the visual and aural senses of the real world are replaced with synthetic substitutes, the only significant obstacles to a true feeling of immersion are tactile sensory input and the knowledge — the *belief* — that the room entered just moments ago is still there. Without visual and aural reinforcement, the degree of certainty for that belief decreases rapidly. It is much easier to gain an auditory perception of being in a fighter jet when that is not in conflict with your visual and rational senses.

8.4 Custom transforms

A great deal of work remains to be done to optimize auralization processing, both for accuracy and for speed. One area worthy of investigation is the use of alternative transforms. The dependence of the HRTF upon frequency bands and the logarithmic nature of human hearing make the conventional FFT an inefficient choice for frequency analysis or fast convolution. Much of the processing time and power is wasted gaining information from frequencies that are of little interest. Wavelet transforms appear to have properties which would make them ideal for addressing individual directional bands.

Another possibility is the construction of a new transform utilizing basis functions derived from the HRTFs themselves. The principle component analysis of HRTFs performed by Wightman and Kistler [21] is a first step in this direction.

8.5 Customization of the HRTF

As discussed in Chapter 4, the use of a general set of HRTFs is a non-optimal compromise necessitated by the difficulty of measuring individual HRTFs. At the present time, the HRTF recording process is cumbersome, requiring the use of an anechoic chamber and other specialized equipment which preclude incorporation into end-user auralization systems. One approach to increased HRTF compatibility is to provide a selection of several general HRTF sets, with a means for the user to select the HRTF which generates the most effective results for him or her. The principle disadvantage to this method is the increased storage requirement for the additional HRTF data.

Another alternative is the creation or modification of HRTFs based on pinnae structural information extracted by computer imaging techniques. Although presently prohibitively expensive, it is conceivable that a system could be developed to scan each individual's ear to determine the major physical structures; an HRTF could then be constructed from the physical dimensions of the ear, as in the work by Han [30]. This would allow for true individualization of the HRTF, and potentially higher accuracy in spatial localization.

There are indeed many paths to explore, and a vast number of optimizations to discover. Auralization is still truly in its infancy and there is much we have yet to accomplish. The challenges of simulating three-dimensional audio fields will remain a valid topic of exploration for a considerable time in the future.

Bibliography

BIBLIOGRAPHY

[1] E. 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).

[2] L. Wirbel, "'Virtual Audio' Finally Sounds Like Music to the Ears," *Electronic Engineering Times*, issue 717, pp. 9-10 (1992).

[3] R. Goldberg, "3-D Audio", *Electronic Musician*, vol. 8(10), pp. 38-47 (1992)

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

[5] J. La Grou, "Studios Go Beyond Stereo," *EQ*, vol. 5(2), pp. 56-109 (1994).

[6] M. Kleiner, B.I. Dalenbäck, and P. Svensson, "Auralization - an Overview," *Journal of the Audio Engineering Society*, vol. 41(11), pp. 861-875 (1993).

[7] A.W. Mills, "On the Minimum Audible Angle", *Journal of the Acoustical Society of America*, vol. 30(4), pp. 237-246 (1957).

[8] D.R. Perrott and A.D. Musicant, "Minimum Auditory Movement Angle: Binaural Localization of Moving Sound Sources," *Journal of the Acoustical Society of America*, vol. 62, pp. 1463-1466 (1977).

[9] D.R. Perrott and S. Pacheco, "Minimum Audible Angle Thresholds for Broadband Noise as a Function of the Delay Between Onset of the Lead and Lag Signals," *Journal of the Acoustical Society of America*, vol. 85(6), pp. 2669-2672 (1989).

[10] D.R. Perrott and J.C. Tucker, "Minimum Audible Movement Angle as a Function of Signal Frequency and the Velocity of the Source," *Journal of the Acoustical Society of America*, vol. 83, pp. 1522-1526 (1988).

[11] J.C. Makous and J.C. Middlebrooks, "Two-dimensional Sound Localization by Human Listeners," *Journal of the Acoustical Society of America*, vol. 87, pp. 2188-2200 (1990).

[12] J. Eargle, *Microphone Handbook* (Elar Publishing, Plainview, NY, 1981).

[13] F.L. Wightman and D.J. Kistler, "Headphone Simulation of Free-field Listening I: Stimulus Synthesis," *Journal of the Acoustical Society of America*, vol. 85(2), pp. 858-867 (1989).

[14] F.L. Wightman and D.J. Kistler, "Headphone Simulation of Free-field Listening II: Psychophysical Validation," *Journal of the Acoustical Society of America*, vol. 85(2), pp. 868-878 (1989).

[15] F.L. Wightman and D.J. Kistler, "sdo_l.dat" and "sdo_r.dat", available by FTP at *waisman.wisc.edu*, in the directory *[anonymous.public]*.

[16] D.R. Begault, "Challenges to the Sucessful Implementation of 3-D Sound," *Journal of the Audio Engineering Society*, vol. 39(11), pp. 864-870 (1991).

[17] D.R. Begault, "Control of Auditory Distance", dissertation, UCSD (1987)

[18] D.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).

[19] W.M. Hartmann, "Localization of Sound in Rooms," *Journal of the Acoustical Society of America*, vol. 74, pp. 1380-1391 (1983) .

[20] W.M. Hartmann, "Localization of Sound in Rooms II: The effects of a single reflecting surface," *Journal of the Acoustical Society of America*, vol. 78, pp. 524-533 (1985).

[21] F.L. Wightman and D.J. Kistler, "A Model of HRTFs Based on Principal Component Analysis and Minimum-phase Reconstruction," *Journal of the Acoustical Society of America*, vol. 91(3), pp. 1637-1647 (1992).

[22] M.D. Wilde, "Temporal Localization Cues and Their Role in Auditory Perception", presented at the 95th convention of the Audio Engineering Society, New York, 1993 Oct 7-10, preprint 3708.

[23] C. S. Lindquist, *Adaptive and Digital Signal Processing* (Steward & Sons, Miami, FL, 1989)

[24] J.C. Middlebrooks, J.C. Makous and D.M. Green, "Directional Sensitivity of Sound-Pressure Levels in the Human Ear Canal," *Journal of the Acoustical Society of America*, vol. 86(1), pp. 89-108.

[25] G. Ballou, *Handbook for Sound Engineers: The New Audio Cyclopedia* (Howard Sams, Indianapolis, IN, 1987), pp. 14-15.

[26] R.L. Lehrman and C. Swartz, *Foundations of Physics* (Holt, Rinehart and Winston, New York, NY, 1965), pp. 297-299.

[27] A.V. Oppenheim and A.S. Willsky, *Signals and Systems* (Prentice Hall, Englewood Cliffs, NJ, 1983).

[28] A.V. Oppenheim and R.W. Schafer, *Discrete-time Signal Processing* (Prentice Hall, Englewood Cliffs, NJ, 1989).

[29] T. Zudock, "Minimization of Impulse Response Duration in Three Dimensional Sound Image Processing", research project, University of Miami (1993).

[30] H.L. Han, "Measuring a Dummy Head in Search of Pinnae Cues," *Journal of the Audio Engineering Society*, vol. 42(1), pp. 15-37.

[31] Microsoft, "Multimedia Programming Interface and Data Specification v1.0," available by ftp at *ftp.uu.net* in the directory *vendor/microsoft/multimedia*.

N. Ahmed, T. Natarajan, and K.R. Rao, "Discrete Cosine Transform," *IEEE Transactions on Computers*, vol. C-23, pp. 90-93 (1974).

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

S.B. Lippman, *C++ Primer* (Addison-Wesley, Reading, MA, 1993), 2nd ed.

B.W. Kernighan and D.M. Ritchie, *The C Programming Language* (Prentice Hall, Englewood Cliffs, NJ, 1988), 2nd ed.