Real-Time Audio Spatialization with Inexpensive Hardware

David A. Burgess
Graphics Visualization and Usability Center - Multimedia Group
Georgia Institute of Technology
Atlanta, Georgia 30332
burgess@cc.gatech.edu

Abstract

There are a variety of potential uses for interactive spatial sound in human-machine interfaces, but tremendous computational costs have made most of these applications impractical. Recently, however, single-chip digital signal processors (DSP's) have made real-time spatial audio an affordable possibility for many workstations. This paper describes a spatialization technique based on empirically derived FIR filters. The fundamental performance and quality limits for this technique are discussed as well the minimum bandwidth required for the associated audio channels. It is shown that current single-chip DSP's may be expected to spatialize several sources to different positions in real time. Techniques for improving spatial audio quality and performance are described. As an example application, the spatial sound system of an all-acoustic computer interface is described.

Introduction

In recent years there has been research into what is called binaural[1] recording. Binaural recordings are intended to be reproduced through headphones, and can give the listener a very realistic sense of sound sources being located in the space outside of the head. These virtual sound sources can be in front of, behind, or even above or below the listener. This effect is achieved by using a suitably accurate model of the human acoustic system, such as a dummy head with microphones embedded in the ears (Plenge, 1974). Because of the better model, the sound waves that arrive at the eardrums during playback are a close approximation of what would have actually arrived at a listener's eardrums during the original performance.

Along with greater realism, binaural sound conveys spatial information about each sound source to the listener. Furthermore, when sounds are spatially separated, a listener can easily distinguish different sources, and focus on those sources which are of interest while ignoring others (Cherry, 1953).

If sounds can be recorded in this manner, an obvious next step is to convert monaural sounds to binaural sounds by artificially spatializing them. This goal has lead research interest in the subject by the military, by NASA (Wenzel et al 1988), and by user interface designers (Ludwig et al 1990, 1991).

The work described in this paper is part of the Mercator[2] project (Mynatt & Edwards, 1992). The goal of Mercator is to allow visually-impaired computer users to have access to application software packages with graphical user interfaces under the X Window System[3]. Mercator will accomplish this task by mapping the structures and behaviors of X applications to an auditory and tactile space. An important feature of Mercator will be the use of spatial sound as a primary organizational cue. For Mercator to be useful, this feature must be implemented at a reasonable cost with off-the-shelf hardware.

This paper is written in two sections. The first section describes the general spatialization technique. The second section describes spatial sound as implemented in Mercator.

Section I: Background and Spatialization Technique

How We Localize Sounds

Much of our information about the position of a sound source is based on the effects of resonances inside the ear and refraction in the vicinity of the head and upper body. The dominant cues provided by these head and upper body effects are interaural delay time (IDT) (Rayleigh, 1907), head shadow (Mills, 1972), pinna and ear canal response (Gardener, 1973), and shoulder echoes (Searle, et al, 1976). Together, these cues form the *head-related transfer function* (HRTF). Blauert (1983) provides a comprehensive description of the HRTF, and Searle, et al (1976) give a statistical study of the relative importance of HRTF cues. Of the HRTF cues listed above, only IDT is described in this paper.

[1] In strict terms, "binaural" and "stereo" mean exactly the same thing—two channels of sound. However, in the music recording field, stereo generally means recorded with multiple, spaced microphones and binaural generally means recorded through a dummy head.

[2] Named for Gerhardus Mercator, the cartographer who devised the Mercator map projection.

[3] Recent legislation in the United States mandates that computer suppliers ensure accessibility of their systems and that employers provide accessible equipment (Title 508 of the Rehabilitation Act of 1986, 1990 Americans with Disabilities Act).

IDT (also called interaural group delay or interaural time difference) is the delay between a sound reaching the closer ear and the farther one. IDT provides a primary cue for determining the lateral position of a sound. The delay is zero for a source directly ahead of, behind, or above the listener and roughly 0.63ms for a source to one's far left or right. The delay varies as a sinusoid with azimuth but is also dependent on both the frequency of the sound and the distance of the source (Blauert, 1983). IDT manifests itself as a phase difference for signals below 1.6kHz and as an envelope delay for higher frequency sounds.

Other cues known to be of importance are head motion (Thurlow & Runge, 1967), vision (Thomas, 1940), early echo response (Moore, 1990), and reverberation (Gardner, 1968). We tend to move our heads to get a better sense of a sound's direction (Thurlow, et al, 1967). This "closed-loop" cue can be added to a spatial sound system through the use of a head-tracking device. Furthermore, we rely so heavily on where we see a sound source that we will ignore our auditory directional cues if they disagree with visual cues. Early echo response and reverberation will be addressed later in the paper.

Basic Technique for Synthetic Spatialization

The HRTF represents a linear system that is dependent on the position of a sound source relative to the listener's head. Therefore, it follows that sounds can be spatialized artificially using well-known digital filter algorithms, if the HRTF is known. Although the HRTF can vary substantially from individual to individual, people who localize well tend to have similar HRTF's (Wenzel et al 1988). It is believed that the important features of the HRTF are consistent enough that one such set of filters may be suitable for a large portion of the population (Wenzel, 1992).

The spatialization technique we will focus on is the use of an empirical HRTF measured from the ears of a specific person (Wightman & Kistler, 1989a&b). The technique is similar to that used for binaural recording—the primary difference is that a living human is used instead of a mannequin. Special probe microphones are used to monitor sound near the eardrum of a test subject. Periodic, pseudorandom noise is played through movable speakers at various positions about the subject. The eardrum responses are recorded through the subject's ears with the probe microphones. Additionally, the noise is played though headphones and recorded by the probe microphones as a reference. In each case, several periods of noise are recorded to insure accuracy. The subject's HRTF is computed at each speaker position by dividing the Fourier transform of each speaker-induced eardrum response by the Fourier transform of the headphone-induced reference response. The spectra of the noise burst and microphone cancel in the division, leaving the HRTF and the transfer function of the speaker. This speaker transfer function remains as an artifact. Other possible artifacts from the recording process are echoes from the gimbals used to support the moveable speakers and a distortion of ear canal response due to the fact that the probe cannot actually touch the eardrum (Asano, et al, 1990).

The resulting HRTFs are transformed back to the time domain to yield a set of finite impulse response (FIR) filters. For each position (azimuth, elevation), there is a pair of filters—one for each ear. To place a sound in a given direction, simply apply the corresponding filter pair. The filters are typically measured at 10 to 20 degree intervals, and filters for intermediate angles can be bilinearly interpolated from the original, empirical set. Azimuth separations as great at 60 degrees can provide useful filters although elevation separations should be much smaller (Wenzel, 1992).[4]

Problems with Spatial Sound

A classic problem with spatial sound is an inability of listeners to tell whether sound sources are in front of or behind them. This problem is not necessarily due to some failing of the spatial sound system, because front-back confusion can occur with real sound sources as well. However, as HRTF accuracy is degraded, rates of front-back reversal increase (Asano, et al 1990).

Another common shortcoming is lack of externalization. As a result, sound may appear to emanate from points inside the head. Externalization is lost when signals reaching the ears are not adequately consistent with those from external sources (Plenge, 1974). In practical terms, this means that externalization requires a faithful model of the HRTF.

A minimum requirement for any useful spatial sound system is a monotonically increasing relationship between perceived position and target position.

Position update rates should be high enough to give an illusion of continuous movement. In tests for Mercator, update rates of 10Hz to 12Hz have been found to be adequate for rotational speeds of up to 180 deg/sec. A closely related issue is position update latency, which is especially important when head-tracking devices are used. Update latencies of 90ms have been found acceptable for rotational speeds of up to 360 deg/sec (Wenzel, 1992). If update latency is a problem, predictive filters may be used to improve the performance of head-tracking systems (Liang, et al, 1991).

Limits of Real-Time Performance

A set of head-related impulse responses were provided to the Mercator project by Professor Fredric Wightman of the University of Wisconsin at Madison. The test subject who pro-

[4] It should be noted that due to changing IDT, the linear interpolation of widely separated FIR's can result in a comb filter effect. The offset of the dominant peak of a head-related impulse response in time as a function of azimuth is roughly 0.315ms x sin(azimuth). With a 60 degree interpolation separation, the IDT offset between filters is 0.27ms. With a 15 degree separation, the offset is only 0.081ms. Shoulder and torso reflections may produce additional comb filter effects.

vided these filters is referred to in several references as SDO (Wenzel, et al, 1988, Wightman & Kistler, 1989a&b, Wenzel, 1992). The impulse responses were produced from recordings made in an anechoic chamber at a sample rate of 50kHz and each had a duration of 512 samples (10.24ms).

Because computational cost grows quadratically with sample rate, we stand to gain a great deal by using the lowest rate possible. The filters are bandlimited to 18kHz and contain little or no useful information above 16kHz (Wightman; personal correspondence). Furthermore, most adults hear very poorly at such high frequencies. It follows that these filters can be resampled at rates as low as 32kHz with little or no loss of effectiveness. At 32kHz, the length of the filter is reduced to 328 points, and the real-time computational requirement is reduced to 21M points per second—less than half the original cost.

Much of the measured HRTF above 12kHz is believed to be inaccurate or excessively noisy due to the limitations of recording equipment (Wightman; personal correspondence). A bandwidth of 12kHz corresponds to a sample rate of 24kHz. When resampled at 24kHz, the filters shrink to 247 points. For real-time use, a 24kHz filter set requires only 11.9M convolution points per second. However, it should be pointed out that at rates below 32kHz the filters are not equivalent to the original 50kHz filters, and are not guaranteed to perform as well.

The absolute minimum limits on sample rates for spatialization systems are still not well-defined. The features of the HRTF are not well understood above 10kHz, but this does not mean that they are of no importance. Wightman and Kistler used a sample rate of 50kHz in their original experiments, although their test stimuli had a bandwidth of only 14kHz. In tests for Mercator, sample rates as low as 22.05kHz have been found to give localized images and maintain a monotonic relationship between target and perceived positions, although the externalization and localization accuracy of these filters may not be suitable for all applications.

Another loss of the efficiency of the original filters was due to long periods (at least 3.40ms) of silence preceding each impulse response. Furthermore, these silent periods give an indication of the noise floor of the original filter set. Once this noise floor is known, it can be seen that each filter also ends in at least 0.54ms of silence. With all 3.94ms of silence removed, the 50kHz filters shrink to 315 points (6.30ms). The 32kHz filters shrink to 202 points and require only 13M points per second for real-time use. The 24kHz filters shrink to 139 points and require 6.7M points per second.

It would be useful to know just how small HRTF filters can be before they stop working or, equivalently, how much detail the HRTF must have in the frequency domain to produce a convincing spatialization effect. More aggressive truncation and windowing can be used to further reduce the lengths of the filters. An existing commercial spatial sound product, Focal Point^{tm}, applies an HRTF filter pair to a 44.1kHz sound stream in real time with a 27MHz DSP56001. It follows that the filters used in the Focal Point^{tm} system can be no longer than 153 taps, or 3.47ms. Filters as short as 128 points (2.56ms) are used effectively with the Convolvotron at 50kHz. An equivalent filter at 32kHz would have only 82 points and require 5.3M points per second, which would allow a DSP56001 to spatialize two channels in real time.[5]

We began to speculate about the minimum possible length for a spatializing filter. Two estimates of minimum filter length were made. For the first estimate, we reasoned that the HRTF cue with the longest delay before reaching the eardrum would be shoulder echo. It would follow that a set of complete HRTF filters could be no shorter than maximum shoulder echo delay time + required delay for IDT, and may need to be longer to capture adequate pinna and middle ear reverberations. Based on body dimensions and geometry, the maximum shoulder echo delay was estimated to be in the neighborhood of 1ms. This maximum delay occurs at high elevations and is subject to an additional delay of 0.32ms to allow for IDT. The total delay is 1.32ms, meaning that a working HRTF filter set probably would not be any shorter than this and may need to be longer.

For the second estimate, we reasoned that most of the spectral cues of the HRTF are above 3.5kHz. This assumption was based on the fact below this frequency, the HRTF has little fine detail (Gardner, 1969). The spectral resolution of the inner ear is 1/3 octave (Blauert, 1983). For an FIR to match this spectral resolution at 3.5kHz, it should have a duration no shorter than 1.10ms. This is a rough figure, however, because the cutoff frequency for fine details is dependent on one's definition of “fine” detail. The cutoff is certainly between 2kHz and 4kHz, however, giving a minimum FIR length somewhere between 0.96ms and 1.92ms.

A Test of Extremely Short Fillers

While the Mercator project lacks the facilities required for rigorous evaluations of localization accuracy, a simple test was devised to get some idea of the size limits of HRTF filters. In this test, a subject is presented with a sound which is slowly moved along an arc by the application of filters. The subject is then asked to describe the direction of the sound's movement. From this test, we can determine whether or not the filter set maintains a monotonic relationship between perceived and target positions along the given arc. We can also test for externalization and front-back differentiation.

Two arcs were used in the test. The first, which we will call arc A, had a constant azimuth of 90 degrees (immediately to the left of the subject) with elevation changing from -36 to +72 degrees over a period of 6 seconds. Steps of 12 degrees were used for elevations below 0 degrees (the median plane), and steps of 24 degrees were used for elevations above 0 degrees. The azimuth of 90 degrees was used because listeners seemed to be most sensitive to changes of elevation in the

[5] It has also been found that product currently under development at Crystal River Engineering is expected to use a 74 point filter (1.68ms) at 44.1kHz.

regions to the immediate left and right. The second arc, B, had a constant elevation of 0 degrees with azimuth changing from 90 degrees to 180 degrees (behind the head) in steps of 15 degrees over a period of 6 seconds[6].

A sample rate of 22.05kHz was used for the test. A low rate was chosen to test the viability of 24kHz filters, and 22.05kHz was the closest rate available. Three sets of HRTF filters were prepared. The first had a length of 128 points, which contained the full duration of the original filters—only silence was removed. These 128 point filters were meant as a control set. The second set was truncated to 32 points (1.45ms). A third set was truncated to only 16 points (0.73ms). To produce these filter sets, the original 50kHz filters were resampled at 22.05kHz, giving a length of 226 samples. Rectangular windows were then applied to this resampled set with each window starting at the 87th sample—the point at which the earliest impulse responses start to rise above the noise floor. Only angles represented in the empirical set were used—no intermediate positions were interpolated.

Spatialization was performed with the internal DSP56001 of a NeXT workstation. The NeXT's internal digital to analog converter was used to generate the audio signals. Because this converter has poor antialiasing characteristics at 22.05kHz, an additional first order lowpass filter was used with a cutoff frequency of 11.3kHz. The headphones were Sennheiser HD540, and the filters were not corrected to compensate for their effect. The headphones were driven by a Sherwood RA-1142 amplifier. The test sound was a recording of music that was known to fill the 11kHz bandwidth being used. Filters were tested in order from longest to shortest, but arcs were presented in random order. Nine subjects were tested. All subjects were sighted males in their early twenties with no previous experience with spatial audio. Two subjects claimed to suffer slight, burial hearing loss. Each subject was placed in a chair and asked to sit upright, face straight ahead, and not to move his head during the test.

For the 128 point filter set, there was no loss of monotonicity along either of the two arcs. Estimates of distance ranged from just outside the skull to "20 or 30 feet."

For the 32 point filter set, there were two cases of loss of monotonicity for arc B. In one of these cases, the sound appeared to move to the center of the head. In another case, the sound appeared to move to the front of the head (this could be normal front-back confusion). There was one case of loss of localization at higher elevations along arc A. For the other eight subjects, however, monotonicity and localization were maintained. The sound source was usually perceived as just outside the skull.

For the 16 point filter set, there were two cases of loss of monotonicity along arc B. In both cases, the subjects described the sound as moving toward the center of the head instead of toward the back. What should be noted, however, is that there were six cases of loss of monotonicity along arc A and one case of total loss of localization at higher elevations. In the two cases in which both localization and monotonicity were maintained, the subjects described the sound as moving more to the toward the back of the head than to the top.

From the test results, it can be concluded that along arc A, the 32 point filters provide a monotonic relationship between target elevation and perceived elevation for a majority of listeners, while the 16 point filters do not.

In a second test, under similar conditions, a virtual sound source was moved along a spiral with elevation ranging from -36 to +72 degrees in 24 degree steps and azimuth rotating in 15 degree steps (a total of 120 positions in the spiral). The entire movement was performed in 120 seconds. Three subjects with no known hearing loss listened to the sound through 32 point filters at a sample rate of 22.05kHz and all reported monotonic movement along the intended pattern in regions not in front of the head, although all subjects experienced front-back reversals for target positions with azimuths in the range of -60 to +60 degrees.[7]

Additional testing is needed to determine the localization accuracy of 1.45ms filters, but if found to perform adequately for a given application, they would allow the simultaneous spatialization of up to four sound sources it real time at a sample rate of 32kHz with existing single-chip DSPs. At a 24kHz rate, up to eight channels could be used. Although these filters are known to exhibit poor performance (front-back reversal) within the field of vision, they may still produce very compelling effects when used in conjunction with visual cues, as would be the case in many virtual reality applications.

Improvements

There are a number of possible methods for improving the quality of a spatial sound system. User training is a important factor. The user is, in effect, listening through another person's ears. The signals that reach listener's eardrums are intended to be the same as those which would reach the ears of the subject from whom the filters were produced. New users may need time to adapt.

The addition of early echoes from the walls of a simulated

[6] Arc B was placed behind the head to prevent results from being skewed due to front-back reversals. For target positions in front of the head, the average rate of front-back reversal with a 22.05kHz sample rate was 67% for all of the filter lengths tested. It should be noted that this is significantly higher than the reversal rates observed with a sample rate of 50kHz using the full bandwidth (Asano, et al, 1990), or with a sample rate of 50kHz using a 14kHz bandwidth (Wightman & Kistler, 1989b).

[7] In an identical test with a sample rate of 32kHz and a filter length of 60 points (1.88ms), the front-back reversal rate was significantly lower.

listening room can improve front-back differentiation as well as overall spatialization quality. Generating fully spatialized first order echoes for a small room would require filters of several thousand points. However, experiments for Mercator have shown that an echo from a single wall, computed at modest cost, allows reliable front-back differentiation for even our most difficult test subjects. For descriptions of methods for simulating small-room acoustics, see Kendall, Martins, et al (1989) and Allen & Berkley (1979).

Control of the perceived distance of a virtual sound source is a difficult problem in spatial sound systems. For a familiar sound, a primary cue for distance perception is intensity (Gardner, 1968, Laws, 1973). At low frequencies (below 1kHz) the intensity of a sound varies inversely with the square of the distance from its source. At higher frequencies (above 3kHz), dispersion causes an inverse cubic variation.

As a sound source moves away from a listener in an enclosed space, the amount of reverberant energy remains fairly constant as the amount of direct energy decreases with distance. By controlling the ratio of direct to reverberant energy, we can impart a sense of distance. When used properly, reverberation can also improve externalization (Plenge, 1974). When adding reverberation to Mercator, it was found that the reverberation filters should be slightly different for each ear to prevent reverberation from being perceived as a separate sound source at the center of the head. It was also found that a more realistic effect could be achieved by adding a small amount of reverberation to the original signal before spatialization. Generating realistic reverberation is not a simple task. Moore (1990) provides a summary of several techniques.

Head tracking devices can greatly improve the effectiveness of a spatial sound system. At least one experiment has shown that with head tracking, a full implementation of the HRTF is not necessary for blind navigation in an acoustic virtual environment (Loomis, et al, 1990).

Section II: An Example Application

Spatial Sound for a Computer Interface

The Mercator interface, which is currently under development, is designed to map X Window System applications to an acoustic virtual world in which interface objects, such as pull-down menus and buttons, are represented by sound sources called auditory icons (Gaver, 1986). Just as the interface objects of a graphical user interface are organized by their positions on a screen or in a window, the icons of the Mercator interface are organized by their positions in a virtual acoustic space. A higher level of organization is provided by the notion of virtual rooms, with are logical division of the user's workspace based on the Xerox PARC Rooms interface (Henderson & Card, 1986).

Much of the work performed by Mercator involves tracking and interpreting the actions of application programs (Mynatt & Edwards, 1992), and is outside of the scope of this paper. Here, we describe only the lowest levels of the Mercator spatial sound interface.

Types of sound in Mercator

As a general rule, Mercator is a quiet interface—interface objects produce sounds only when requested to do so or when changing state. Auditory icons are short in duration and formed by passing a base sound though a series of linear and nonlinear functions such as filters, distortion generators, and pitch shifters. The modifying functions provide information to the user by reflecting the state of the interface object which the icon represents (Ludwig, et al, 1990). The base sounds are typically sampled from natural sources, stored on disk until needed, and cached in memory while in use. Auditory icons are spatialized for organizational and navigational purposes. The Mercator spatial sound system provides support for up to two simultaneous, independent channels of auditory icons.

Mercator presents most types of text via synthesized voice from a Digital Equipment DECtalk DTC01[8]. Sound from the DECtalk is digitized in real time and feed into Mercator's digital audio system so that it may be spatialized.

Continuously played background noises are added for navigational purposes—rooms have distinctive background sounds assigned to them, such as running water or fan noise. Like auditory icons, background noises are sampled from natural sources and kept on disk or in memory. Background noises are generated by a CODEC[9] at a 8.013kHz sample rate and mixed into the analog audio stream after all other processing—they are not spatialized or subject to modifying functions. A provision is also made for digitizing real-time background noise and including it in Mercator's digital audio processing network. The feature would allow for a monitor of sounds from the user's physical environment[10]. In the output, the left and right channels of real-time background noises are separated by a delay of 1ms to produce a diffuse sound image instead of one at the center of the head.

The Mercator Spatial Sound System

The current platform for Mercator development is a Sun Microsystems SPARCstation IPX. An Ariel S-56X DSP coprocessor board has been installed to provide additional computing power for audio processing. Sixteen-bit, two-channel digital to analog and analog to digital conversion are performed with an Ariel ProPort Model 656. The system is diagrammed in figure 1. (Only those components related to sound generation are show.)

A sample rate of 32kHz has been chosen for Mercator's spatial sound system to insure adequate spatial sound quality. The S56-X uses a DSP56001 (Motorola, 1990) with a clock rate of 27MHz. Thus, to generate binaural output, the DSP is

[8] Although the DECtalk is expensive, it is already widely installed in the visually-impaired community.
[9] Many installed Unix workstations include CODECs.
[10] A number of potential users expressed concern over not being able to hear sounds from the physical environment.

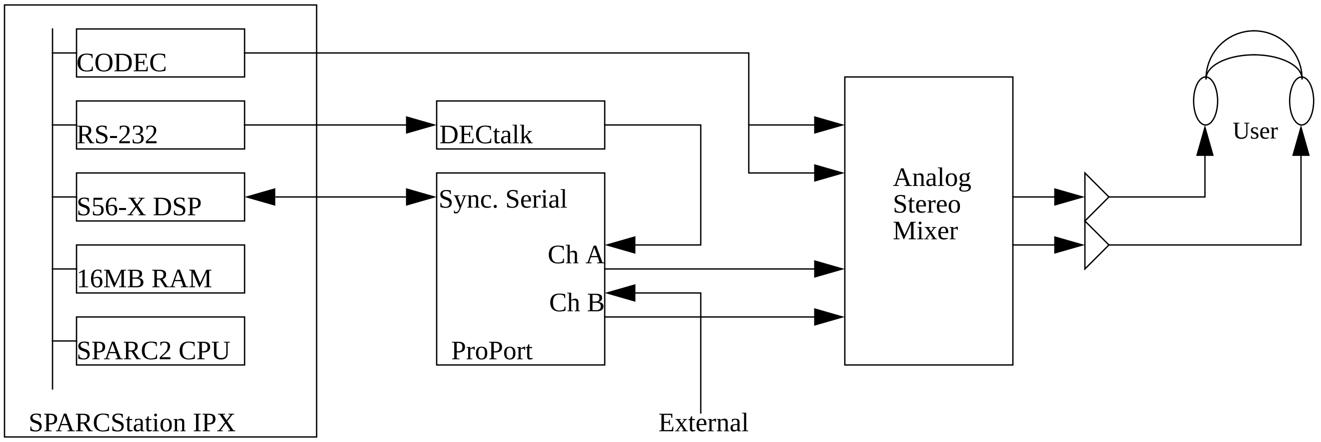

Figure 1. Mercator Audio Hardware

allowed 420 clock cycles to compute each output sample. In this time, the spatialization system must perform the following tasks:

- read incoming auditory icon streams from the host machine
- read synthesized voice and real-time background noises from analog to digital converters
- compute reverberation and modifying functions for the two spatialization channels
- apply a total of four HRTF convolutions
- mix background noises and spatialized channels

Because of limited computing power, a trade-off must be made between the spatialization quality and the variety of modifying functions which can be supported. The decision was made to preserve spatialization quality as well as possible. To facilitate the choice of modifying functions, all of the prospective functions were coded for the DSP, their computational costs were tabulated, and the utility of each was weighed against its cost. Three inexpensive but useful effects were chosen:

- Reverberation—an important part of the spatialization algorithm. Because of DSP processing limitations, a simple recirculating buffer is used for reverberation, although a lowpass filter is included in the feedback loop to allow for warmth. Reverberation is added to the signal after spatialization, and independent reverberators are used on each ear for each audio input stream.
- Nonlinear distortion—used to draw the user's attention to particular events. Any of a variety of distortion functions may be selected at compile time.
- Three-tap FIR—used to create filters for muffling and thinning effects. The filter taps can also be moved to arbitrarily high orders for comb filters.

Remaining computing power was dedicated to spatialization, allowing 55-point (1.72ms) HRTF filters. Filters of this length and sample rate have been tested by a variety of subjects and appear to provide suitable spatialization quality for Mercator.

One input channel of the audio processing network is diagrammed in figure 2. The processing network for the other input channel is identical. After spatialization, the outputs of these networks are summed and background noises are added. In addition to 110 HRTF filter coefficients, a total of 21 control parameters are needed to control distortion, equalization, distance, volume, and reverberation for each spatialization channel.

Control and I/O Mechanisms

There are a total of six audio channels and four control channels passing into and out of the DSP56001 coprocessor. Of the six audio channels, four are synchronous and driven by the ProPort sampling clock: two output channels, DECtalk input, and real-time background noise. All of these channels use the DSP56001's synchronous serial interface (SSI). The DSP software is driven in lockstep with the SSI frame clock for efficiency. The other two audio channels, which carry auditory icons, come from the host and are asynchronous—the host provides data on demand in bursts. These two channels are multiplexed over the DSP56001's host interface, and samples are truncated to twelve bits each so that two samples may be sent in each 24-bit word. Two audio input buffers are used in the DSP—one receives new samples while the other is spatialized. It is important that neither processor spend time waiting on this data stream. To prevent host from having to wait for the DSP to be ready to accept more data, the DSP generates an interrupt (which we call the *data demand interrupt*) to the host when it finishes spatializing a buffer. Inside the DSP, each sample generates an interrupt (host data receive full) as it arrives.

The four control channels carry HRTF filter coefficients and controlling parameters for the spatialization network. These channels are only active when updating source positions and

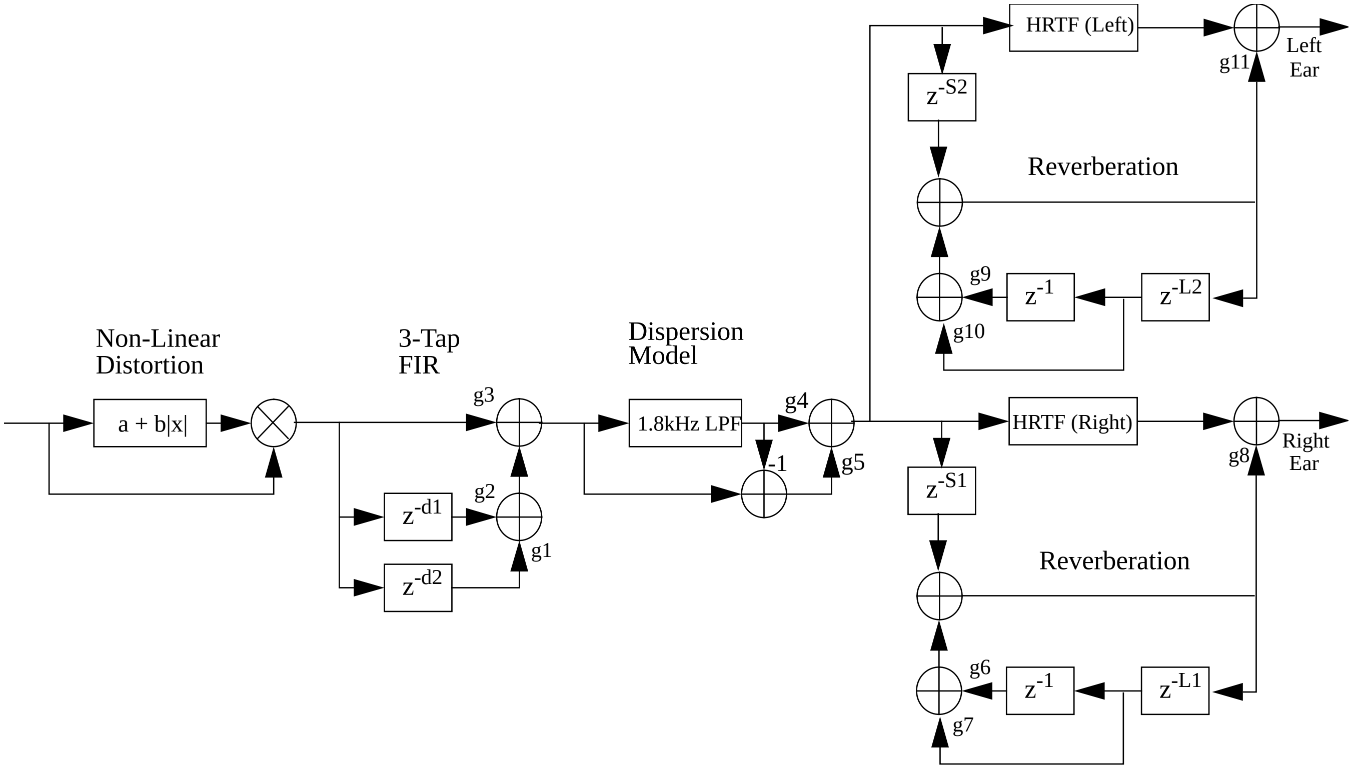

Figure 2. One Channel of Spatialization Network

during user actions, and require very little average bandwidth. It is important to note that the time to transmit a new set of HRTF coefficients is typically several audio clock cycles. To prevent incomplete HRTF tables from causing annoying clicks in the output during these transmissions, the DSP keeps two tables of coefficients. One table is used as a receive buffer; the second is applied to the sound stream. At the end of a transmission, the roles of the two tables are swapped so that the newly completed filter is applied to the sound and the old table is ready to buffer input. This double-buffering scheme is also used with control parameters.

Because the host interface carries six asynchronous, multiplexed channels, a general protocol is used to control host-to-DSP communication. Each transmission to the DSP is preceded by a host command interrupt. Each packet type uses its own particular host command[11]. The purpose of this host command is to set the proper value in the host data receive interrupt vector. The host command is followed by a packet of audio samples or control information, with each incoming word received by the host data receive full interrupt. Audio packets are 4096 words (128ms of sound), HRTF packets are 128 words, and control parameter packets are 32 words. Each transmission is terminated by a second host command which disables host data receive full interrupts, and in the case of control parameters and HRTF coefficients, swaps pointers so that the new parameter and coefficient sets are used.

Host-Side Operation

When designing the host-side support software for the spatialization system, it is important to remember that Mercator is a secondary task. The host computer must still be able to execute application programs at reasonable speeds. To maintain and control the spatial sound system, the host must perform a variety of tasks simultaneously and with little overhead. To meet these requirements, the host-side software is written as a collection of lightweight processes, or threads, which execute concurrently in the same address space. Threads provide the same multiprocessing features as “heavyweight” operating system processes, but at a lower overhead cost. All of the threads in a given address space are considered to be a single process by the operating system. The SPARCstation’s operating system supports threads via its LWP library (Sun, 1990). Many computers in this class provide similar support for multi-threaded programming.

The highest priority task is to keep the DSP supplied with a stream of audio samples. Once the DSP has issued a data demand interrupt, a new audio packet must be completely transmitted within 128ms. If a channel is idle, zeroes must be sent. A realistic time for the receipt of a complete 4096-sample packet by the DSP is about 40ms. To prevent the host from having to wait on this transmission, host-to-DSP audio packets are sent by a dedicated thread of execution which we call the *DSP feeding process*.

The next task, in order of priority, is to supply audio data to

[11] The DSP56001 provides a total of 32 interrupts. Of these, 13 are set aside for host commands.

the CODEC for the production of background sounds. These transfers do not pose the same problems as host-to-DSP audio because of their lower bandwidth and existing support for the CODEC provided in the SPARCstation's operating system and hardware. CODEC sound is also managed by a dedicated thread of execution.

The host-side support software must also compute filter coefficients and control parameters from high-level descriptions of the virtual environment. For example, higher levels of Mercator describe rooms in terms of dimensions and wall materials. From this description, the support software computes gain and delay values for the reverberators and sends the new parameters to the DSP. Related tasks include the control of the DECtalk, the maintenance of a cache of commonly used sounds, and the introduction of new sounds to the auditory icon and background channels. These tasks are performed by a collection of threads called the *control processes*. There is no fixed number of control processes—most are created on an as-needed basis and killed when their tasks are complete.

Because multiple threads may need to communicate with the DSP, it is assigned a *lock*. A thread must have possession of this lock to communicate with the DSP, and no more than one thread may hold the lock at any given time.

A critical problem in this processing model is the synchronization of audio channels and control channels. In order to control the movement of a sound, or to make it initially appear at a given position, we must be able to match the transmission of a control packet with a specific point in the audio signal. The only convenient time to accomplish such synchronization is when the DSP finishes spatializing its current audio input buffer. The following mechanisms are provided for synchronization:

- Host Audio Packet Count—The host maintains a count of outgoing audio packets. This count serves as a real-time clock for the audio system with a resolution of 128ms. The host also keeps the system clock value from the most recent data demand interrupt.
- Signal to the Host—The DSP interrupts the host when the current audio input buffer is exhausted. This is the same as the data demand interrupt used to initiate new audio packets, but it may also be used for synchronization.
- Signal to DSP—The host sends a command interrupt to the DSP instructing it to discard the rest of the current input buffer and optionally wait for a control packet. If the wait option is used, the audio output will be silent during the waiting period.
- Barrier—The host sends a command interrupt to the DSP instructing it to disable further host command interrupts until the current audio input buffer is consumed. By doing so, the DSP refuses to accept new packets from the host. If a host thread attempts to send a packet, it will be blocked until DSP re-enables the interrupts.

Acknowledgments

Most of this work has been sponsored by the NASA Marshall Space Flight Center under Research Grant NAG8-194. Additional work has been supported by Sun Microsystems. The author would also like to thank Professor Fredric Wightman of the University of Wisconsin at Madison for providing HRTF filters and related information, as well as Scott Foster of Crystal River Engineering for information about the Convolvotron and other Crystal River projects.

References

1. Allen, J.B. & Berkley, D.A. (1979) An image method for efficiently simulating small-room acoustics, J. Acoust. Soc. Am., 65, 943-950.

2. Asano, F., Suzuki, Y., Toshio, S. (1990), Role of spectral cues in median plane localization, J. Acoust. Soc. Am., 88, 159-168.

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

4. Buxton, W., Gaver, W.W. & Bly, S. (1991) The Use of Non-Speech Audio at the Interface, Tutorial No. 8, ACM Conference on Human Factors in Computer Systems, CHI '91, ACM Press: New York, NY.

5. Cherry, E.C. (1953) Some experiments on the recognition of speech with one or two ears, J. Acoust. Soc. Am., 22, 61-62.

6. Gardner, M.B. (1968) Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space, J. Acoust. Soc. Am., 45, 47-53.

7. Gardner, M.B. (1973) Some monaural and binaural facets of median plane localization, J. Acoust. Soc. Am., 54, 1489-1495.

8. Gaver, W.W. (1986) Auditory icons: Using sound in computer interfaces, Human-Computer Interaction, 2, 167-177.

9. Gaver, W.W. (1989) The sonicfinder: An interface that uses auditory icons, Human-Computer Interaction, 4, 67-94.

10. Henderson, D.A. & Card, S.K. (1986) Rooms: The use of multiple virtual workspaces to reduce space contention in a window-based graphical used interface, ACM

Transactions on Graphics, July 1986, 211-243.

11. Kendall, G.S., Martins, W.L. & Decker, S.L. (1989) Spatial reverberation: discussion and demonstration, *Current Directions in Computer Music Research*, MIT Press: Cambridge, MA.

12. Laws, P. (1973) Entfernungshören und das Problem der Im-Kopf-Lokalisierheit von Hörereignissen [Auditory distance perception and the problem of in-head localization of sound images], Acustica, 29, 243-259.

13. Liang, J., Shaw, C. & Green, M. (1991) On temporal-spatial realism in the virtual reality environment, Proceedings of the ACM Symposium on User Interface Software and Technology, UIST '91, 19-25.

14. Loomis, J.M, Hebert, C. & Cicinelli, J.G. (1990) Active localization of virtual sounds, J. Acoust. Soc. Am., 88, 1757-1764.

15. Ludwig, L.F., Pincever, N. & Cohen, M. (1990) Extending the notion of a window system to audio, Computer, Aug. 1990, 66-72.

16. Ludwig, L.F. & Cohen, M. (1991) Multidimensional audio window management, International J. Man-Machine Studies, 34(3), 319-336

17. Mills, A.W. (1972) Auditory localization, *Foundations of Modern Auditory Theory*, Vol. II/8, Academic: New York, NY.

18. Moore, F.R. (1990) *Elements of Computer Music*, Prentice Hall: Englewood Cliffs, NJ.

19. Motorola (1990) *DSP56000/DSP56001 Digital Signal Processor User's Manual (Rev. 2)*, Motorola Literature Distribution: Phoenix, AZ.

20. Mynatt, E. & Edwards, W.K. (1992) The Mercator environment: A nonvisual interface to X Windows and Unix workstations, Proceedings of the ACM Symposium on User Interface Software and Technology, UIST '92.

21. Plenge, G. (1974) On the differences between localization and lateralization, J. Acoust. Soc. Am., 56, 944-951.

22. Lord Rayleigh [Strutt, J.W.] (1907) On our perception of sound direction, Phil. Mag., 13, 214-232.

23. Searle, C.L., Braida, L.D., Davis, M.F. & Colburn, H.S. (1976) Model for auditory localization, J. Acoust. Soc. Am., 60, 1164-1175.

24. Sun Microsystems (1990) *SunOS Programming Utilities and Libraries*, Chap. 2, Lightweight Processes, Sun Microsystems: Palo Alto, CA.

25. Thomas, G.J (1940) Experimental study of the influence of vision on sound localization, J. Exper. Psych., 28, 163-177.

26. Thurlow, W.R. & Runge, P.S. (1967) Effect of induced head movements on localization of direction of sounds, J. Acoust. Soc. Am., 42, 480-488.

27. Thurlow, R.W., Mangels, J.W. & Runge P.S. (1967) Head movements during sound localization, J. Acoust. Soc. Am., 42, 489-493.

28. Wenzel, E.M. (1992) Localization in virtual acoustic displays, Presence, 1, 80-107.

29. Wenzel, E.M., Wightman, F.L. & Foster S.H. (1988) A virtual display system for conveying three-dimensional acoustic information, Proceedings of the Human Factors Society - 32nd Annual Meeting, 86-90.

30. Wightman, F.L. & Kistler, D.J. (1989a) Headphone simulation of free-field listening I: stimulus synthesis, J. Acoust. Soc. Am., 85, 858-867.

31. Wightman, F.L. & Kistler, D.J. (1989b) Headphone simulation of free-field listening II: psychophysical validation, J. Acoust. Soc. Am., 85, 868-878.