4. AUDITORY INTERFACES

Historically, discussions of realism for simulation have placed little or no emphasis on auditory requirements. Early work on VEs tended to focus almost exclusively on the visual channel, viewing the VE as a 3-D extension of more 2-D simulations. However, human beings are constantly bombarded with auditory stimuli and, for many applications, eliminating this channel in the development of a VE may be inadvisable.

Auditory stimuli do more than increase the realism of a simulation: In many instances they are essential cues for accurate task performance. For example, Begault (1992) discusses research conducted at NASA Ames Research Center where it was found that pilots had difficulty knowing when they had positively engaged a touch screen "virtual" button without a concomitant auditory cue. When a recording of an aircraft switch engaging or disengaging was coupled with the touch panel, the pilots' preference for the interface was increased. This line of thought can be extrapolated one step further. As in any simulation, in VEs there are some stimuli that should not be used due to their potential danger to the individual. For example, the haptic sensations resulting from crashing an airplane or improperly discharging a high voltage could be fatal to the individual. However, the concomitant auditory cues (for example, the sound of the crash or of the electrical discharge) could be used to simulate the event without any danger to the individual. Indeed, Massimino and Sheridan (1993) have demonstrated the value of auditory cues for substituting for force feedback in various manipulation tasks. In other work, Hendrix (1995) has shown that the addition of spatialized sound significantly increased the reported sense of presence in a VE, although it did not increase the apparent realism of that environment. In an experiment at the Jet Propulsion Laboratory (JPL) Advanced Teleoperation Laboratory, researchers found that auditory feedback, given in addition to visual and kinesthetic feedback, speeded the completion of manipulation tasks (Apostolos et al, 1992).

Although auralization—the 3-D simulation of a complex acoustic field—receives the most attention in discussions of audio interfaces for virtual environments, it would be remiss not to include at least some mention of the field of sonification and its application to virtual environments. Sonification is the audible display of data, such as aids for database or map navigation, or the symbolic representation of error messages or data characteristics. For virtual environment applications having to do with complex multivariate inputs, this technique could greatly aid in data reduction.

One of the most often cited applications of sonification is that of an auditory equivalent to the graphical user interface (GUI). An increasingly experienced problem with GUIs and their associated icons is clutter of the workspace. With the use of sonification and "earcons" for data representation, this clutter could be significantly reduced. Audio objects and icons, representing data, messages, processes, or resources, could be "placed" in 3-D space around the user, and either automatically signal messages or status changes, or reply to user interrogation.

Although there has not, to date, been the volume of research on sonification that there has been on auralization, some studies have been performed that demonstrate its potential usefulness. For example in a direct empirical comparison of physiological data presented via an auditory display versus a standard visual display, Tecumeh and Kramer (1994) found that subjects responded more rapidly and more accurately to simulated operating room emergencies when using the auditory display. They concluded that for systems where large numbers of variables are causally and temporally interconnected in subtle ways, auditory displays may have a distinct advantage over traditional visual displays. This will become increasingly important as the amount and complexity of data that needs to be processed continues to grow.

An interesting phenomenon that has been reported as part of the virtual audio experience is synesthesia, where another sensory organ is stimulated by the 3-D audio input. For example, the experience has been cited for one high-end 3-D audio product that some auditory inputs are perceived as having tactile properties. For example, when an auditory "soda can" is "opened" next to the listener's ear, not only is appropriate sound perceived, but the concomitant bursting "carbon dioxide bubbles" are felt on the skin. Although no experimental studies seem to have been done to date on this phenomenon, it represents a potential area of interest not only for the study of human perception and the interaction between the senses, but also for the applied area of virtual environments.

These various reasons for the use of sounds, combined with advancements in technology that allow more realistic simulation of real-world auditory inputs, has led to research and development for simulation of the auditory channel for VEs. The current technologies are essentially software-based and regarded as highly proprietary by their developers. Hence, it is impossible to provide detailed information about them.

Before discussing the research and products resulting from the application of these principles in the area of VEs, it is first important to understand some of the mechanics of the human auditory system, including how humans hear and the limitations on the stimuli that they can interpret. A more complete discussion of the anatomy and functioning of the human ear can be found in Scharf and Buus (1986).

4.1 The Human Auditory System

The auditory system has three basic functions: (1) to transmit sound through the outer, middle, and inner ears, (2) to transduce sound waves into neural energy in the inner ear, and (3) to perform neural processing within and transmission through the audio-vestibular nerve and four or five neural levels to the auditory cortex.

The outer ear comprises the pinna (or auricle) and the ear canal (or external meatus). When a sound reaches the outer ear it continues through the air in the ear canal where the pinna concentrates it, increases its amplitude, and reflects the sound at the entrance of the ear canal. This results in the intensity of the sound being changed by as much as ±10 dB. These effects have been found to be greatest above 2000 Hz (Shaw and Teranishi, 1968). This means that a complex sound containing many high frequencies will vary at the entrance to the ear canal depending on the direction from which it comes. The effects caused by the pinna, therefore, provide cues to the location of a sound source and help to give the impression that a sound source is external to the listener. It should be noted, however, that when auditory stimuli are transmitted via headphones, the sound bypasses the pinna and arrives directly at the ear canal, so that most of the effect of the pinna disappears. It is also worth noting that the difference in propagation delay to the two ears is a source of localization, especially at low frequencies.

After the sound passes through the ear canal, it reaches the middle ear. The primary purpose of this complex system is to match the impedance of the air in the outer ear with that of the fluid in the inner ear. This function of the middle ear prevents sound loss resulting from reflection by the denser cochlear fluid, and allows sound to reach the inner ear with little attenuation. From here, sound vibrations are transmitted through the ossicles (three small bones connected to the tympanic membrane) to the inner ear (or cochlea). This, in turn, causes movement of the cochlear fluid, which causes the basilar membrane to vibrate. This results in bending and activation of the hair-cell receptors lying between the basilar and tectorial membranes, which are innervated by fibers of the auditory nerve. Axons of these fibers enter the central nervous system and synapse in the cochlear nuclei of the medulla, causing the sound waves to be interpreted as audible sound.

There are limits on what sound frequencies are audible to the human ear. These are determined by the acoustical and mechanical properties of the ear canal, ear drum, and the middle ear bones, which set limits on the efficiency with which sounds of various frequencies are converted to mechanical vibrations and transmitted to the cochlea. In humans, the auditory apparatus is most efficient between 1000 and 4000 Hz, with a drop in efficiency as the sound frequency becomes higher or lower. The absolute degree of sensitivity of the human ear is quite remarkable; for example, a movement of the ear drum of less than one-tenth the diameter of a hydrogen atom can result in an auditory sensation. In fact, persons with very good hearing can detect Brownian movement in a soundproofed anechoic chamber (Scharf and Buus, 1986). If the ear were more sensitive than this, random Brownian movements would produce a constant sound and would tend to mask auditory stimuli.

Table 5 presents the international standard threshold values for the minimum audible pressure (MAP) and minimum audible field (MAF). It should be noted, however, that precise values are dependent upon a number of factors, including the type of earphones used in the test and the manner in which the earphones were calibrated. Table 5 gives the values for both the Western Electric 705-A earphone calibrated on a National Bureau of Standards type 9-A coupler and the Telephonics TDH-39 earphone. The values in this table extend only to 8000 Hz since equipment calibration at higher levels is not reliable. In attempts to overcome this problem, it has been found that for teenagers and young adults the threshold at first slowly rises by 6 to 8 dB, and then jumps another 12 to 14 dB between 14,000 and 16,000 Hz (Stelmachowicz, Gorga, and Cullen, 1982; see also Fausti, Frey, Erickson, Rappaport, Cleary, and Brummett, 1979).

Table 5. Threshold Values in Free Field (MAF) and Earphone Listening (MAP)[a]

Sound Pressure Level (dB)						
Frequency (Hz)	ISO Minimum Audible Field (Standard Deviation)		Modified MAF	ISO (W.E. 705-A) Minimum Audible Pressure (Standard Deviation)[b]		ISO (TDH-39)[c]
50	41.7	(6.0)	43.5	—	—	—
120/125	21.4	(5.0)	28.5	45.5	(5.6)	45.0
250	11.2	(4.5)	17.5	24.5	(5.0)	25.5
500	6.0	(4.5)	8.0	11.0	(5.4)	11.5
1,000	4.2	(4.5)	4.2	6.5	(5.4)	7.0
2,000	1.0	(5.0)	1.0	8.5	(5.9)	9.0
4,000	-3.9	(8.0)	-3.9	9.0	(7.6)	9.5
6,000	4.6	(8.5)	4.6	8.0	(7.4)	15.5
8,000	15.3	(8.5)	15.3	9.5	(9.9)	13.0
10,000	16.4	—	16.4	—	—	—
12,000	12.0	—	12.0	—	—	—
15,000	24.1	—	24.1	—	—	—

a. Adapted from (Scharf and Buss, 1986), (ANSI, 1969), (Berger, 1981), (ISO, 1961), (ISO,1975), (Robinson and Dadson, 1956), and (Wessler, 1968).
b. These values are for the Western Electric 7050A earphone mounted in an MX041/AR cushion and calibrated on a National Bureau of Standards 9-A coupler. The standard deviations are associated with the original threshold determinations.
c. MAP values of the Telephonics TDH-39 earphone mounted in the MX-41/AR cushion. (On the basis of new measurements and review of the literature, Robinson et al (1981) suggest that the MAP values for the TDH-39 earphone are 2.0 to 2.5 dB too high at the 500 to 4000 Hz levels.

The figures given in Table 5 are averages. In fact, thresholds are greatly dependent on age and sex of the listener. Numerous studies have shown that hearing deteriorates with age, particularly for men (for example, Robinson and Sutton, 1979; Moller, 1983). Table 6 summarizes the results from Hinchcliffe (1959), who randomly sampled 400 persons from a rural population of 9000. After rejecting ears with any otological abnormality, a total of 645 ears were tested. Subjects included both women and men between the ages of 18 and 74. In all these cases the men had significantly higher thresholds than the women. Threshold increases with age, more so at high frequencies than at low, and more rapidly after 45 to 54 years. Above 1,000 Hz, males usually have higher thresholds than women in most age groups. At all frequencies, the threshold increases continuously with age, with the greatest loss at frequencies above 2000 Hz. In addition, Table 6 shows that at frequencies from 3,000 to 6,000 Hz, women of all ages have lower thresholds than do men.

It also has been suggested that women have lower thresholds at all ages and frequencies, with this advantage increasing with both age and frequency (Cors, 1963).

Table 6. Threshold as a Function of Subject's Age and Stimulus Frequency

<table>
<tr><th></th><th colspan="12">Threshold (dB) re Youngest Age Group</th></tr>
<tr><th>Frequency
(Hz)</th><th colspan="2">Age 8-24
(176)</th><th colspan="2">Age 25-34
(104)</th><th colspan="2">Age 35-44
(93)</th><th colspan="2">Age 45-54
(104)</th><th colspan="2">Age 55-64
(74)</th><th colspan="2">Age 65-74
(94)</th></tr>
<tr><td>125</td><td colspan="2">4.3
0.0
-3.9</td><td colspan="2">5.5
1.7
-2.1</td><td colspan="2">6.1
2.6
-1.1</td><td colspan="2">9.5
4.8
1.8</td><td colspan="2">13.1
8.7
4.6</td><td colspan="2">17.1
10.1
6.0</td></tr>
<tr><td>250</td><td colspan="2">2.9
0.0
-3.4</td><td colspan="2">5.3
1.0
-2.8</td><td colspan="2">5.6
1.7
-1.5</td><td colspan="2">7.6
3.2
0.1</td><td colspan="2">11.6
6.5
2.3</td><td colspan="2">16.4
9.6
4.5</td></tr>
<tr><td>500</td><td colspan="2">4.0
0.0
-2.7</td><td colspan="2">4.1
0.7
-2.4</td><td colspan="2">5.7
1.7
-2.0</td><td colspan="2">8.7
3.9
0.4</td><td colspan="2">12.8
7.0
2.4</td><td colspan="2">20.8
9.7
4.8</td></tr>
<tr><td>1000</td><td colspan="2">3.7
0.0
-3.6</td><td colspan="2">4.1
1.0
-2.4</td><td colspan="2">6.4
1.7
-2.1</td><td colspan="2">9.5
4.7
0.8</td><td colspan="2">10.0
5.6
1.3</td><td colspan="2">24.7
12.8
5.2</td></tr>
<tr><td>2000</td><td colspan="2">4.6
0.0
-4.2</td><td colspan="2">4.8
0.4
-3.5</td><td colspan="2">6.9
2.5
-0.5</td><td colspan="2">10.9
5.5
1.2</td><td>14.9
8.7
4.6</td><td>17.9
12.1
5.7</td><td>26.6
14.6
9.4</td><td>41.1
25.1
15.6</td></tr>
<tr><td>3000</td><td>4.6
0.0
-4.0</td><td>8.0
2.1
-0.9</td><td>6.9
1.5
-2.8</td><td>11.1
5.8
1.1</td><td>10.3
5.5
0.4</td><td>16.5
8.6
3.7</td><td>18.2
9.9
4.8</td><td>29.4
18.2
6.3</td><td>20.2
14.8
8.8</td><td>45.3
31.5
18.7</td><td>40.6
19.8
10.1</td><td>53.1
40.9
30.7</td></tr>
<tr><td>4000</td><td>4.3
0.0
-4.3</td><td>10.2
3.5
-1.6</td><td>8.5
3.8
-0.2</td><td>12.9
7.5
2.4</td><td>10.0
5.3
1.7</td><td>19.8
12.6
5.4</td><td>18.9
13.2
6.6</td><td>45.3
22.2
12.4</td><td>26.3
19.4
8.7</td><td>59.6
37.8
30.7</td><td>45.6
22.2
12.1</td><td>59.6
45.5
29.9</td></tr>
<tr><td>6000</td><td>5.8
0.0
-6.6</td><td>9.6
2.5
-2,6</td><td>8.8
3.6
0.5</td><td>12.9
5.3
-1.2</td><td>13.6
6.2
0.6</td><td>20.9
12.4
5.9</td><td>22.4
11.2
3.5</td><td>36.7
23.1
14.0</td><td>28.7
22.3
11.3</td><td>59.7
49.5
30.6</td><td>47.2
33.9
17.4</td><td>66.5
50.9
34.9</td></tr>
<tr><td>8000</td><td colspan="2">5.6
0.0
-6.3</td><td colspan="2">9.6
3.3
-5.4</td><td colspan="2">15.7
7.2
-2.3</td><td>28.4
8.2
2.6</td><td>45.1
20.7
10.7</td><td>39.7
24.7
11.0</td><td>67.8
53.5
30.1</td><td>52.3
42.2
32.2</td><td>68.5
57.2
48.7</td></tr>
<tr><td>12000</td><td colspan="2">9.2
0.0
-6.6</td><td colspan="2">17.6
5.2
-4.0</td><td colspan="2">82.5
14.4
3.9</td><td colspan="2">58.0
41.7
19.0</td><td colspan="2">70.0
64.2
54.2</td><td colspan="2">70.0
70.0
63.1</td></tr>
</table>

Note: The middle value in each triplet is the median value; the bottom value is the 25th percentile and the top value is the 75th percentile. The number of ears within each age group is given in parentheses. Where two sets of triplets appear together, the set on the left is for women and that on the right for men. The table displays the medians and the 25th and 75th percentiles of the hearing levels relative to the thresholds of the youngest age group at the corresponding frequency. Where single scores are given, there were no statistically significant gender differences at that frequency.

Typically, sounds are described using three variables: pitch (frequency), tone color (spectral content), and loudness (intensity). When one is trying to synthesize sounds within a virtual environment, a fourth variable comes into play—spatial location. This variable is dependent on a number of factors. First, interaural intensity and time difference cues are essential to determination of auditory localization. For example, when a sound occurs at 45° to the listener's right at 45° azimuth and 0° elevation, the sound is louder in the right ear than in the left. As the intensity differences vary between the two ears, the listener interprets changes in the sound location. In addition, since the path to the right ear is shorter than that to the left ear in this example, it will reach the right ear fractionally sooner than the left ear.

The original theory of sound localization was based on two types of sound measurements. Since wavelengths smaller than the human head create an intensity loss or head shadow at the ear farthest from the sound source, researchers thought that the brain used interaural intensity differences (IIDs) to localize high frequency sound. Interaural time differences (ITDs), on the other hand, were thought to be important for low frequencies since the interaural delay relationships were clear for wavelengths larger than the human head.

This duplex theory, however, does not account for the ability to localize sounds on a vertical median plane with minimal interaural cues. In addition, the duplex theory does not account for the fact that sounds often appear to be coming from inside the head when heard over earphones, even though the appropriate IIDs and ITDs are present. It is now thought that direction-dependent filtering resulting from sound interacting with the outer ears at least partially explains these deficiencies. Research has shown that the pinnae shape the sound waves in highly direction-dependent ways, and are at least partially responsible for the perception that sounds are occurring outside the head.

The synthesis of a 3-D auditory display typically involves the digital generation of stimuli using location-dependent filters. These filters are constructed from acoustical measurements made using small microphones placed in the ear canals of individual subjects. These ear-dependent filters are usually referred to as head-related transfer functions (HRTFs), and act much like graphic equalizers. The HRTFs capture the IIDs, ITDs, and spectral coloration produced by a sound's interaction with the pinnae that are essential for localizing sounds. An alternative method of measuring HRTFs involves the use of a geometric model of the "average" human head, shoulders, and upper torso. This model purportedly yields results comparable to those obtained from sampled HRTFs as discussed above. The modeling process has also been used to construct artificial heads that can be fitted with miniature internal microphones so that live 3-D sounds can accurately be recorded.

However, 3-D effects are lateralized or externalized rather than truly being localized. This occurs for several reasons: First, there is a lack of input from head motion cues. In trying to localize sounds, humans tend to move their heads toward the hypothesized source, thereby increasing the data they receive and interpret concerning spatial localization. Second, modification of the sound wave by the pinnae acts to emphasize some frequency regions and attenuate others. As discussed by Begault (1992), although two sound sources—one at right 60° azimuth, 0° elevation and another at the mirror image position of 120° azimuth, 0° elevation—have the same overall interaural time and intensity difference, the rearward sound has a relatively "duller" quality. In fact, for a given broadband sound source, each elevation and azimuth position relative to the pinnae has a unique set of spectral modifications. This is due to the complex construction of the outer ears, which impose a set of minute delays that collectively translate into a particular binarual HRTF for each sound source position.

4.2 Commercially Available 3-D Audio Products

A wide continuum of commercial products are available for the development of 3-D sound. These range from low-cost, PC-based, plug-in technologies that provide limited 3-D capabilities to professional quality, service-only technologies that provide true surround audio capabilities. The characteristics of the products identified here are summarized in Table 7, and the products themselves are discussed in alphabetical order below.

In addition to these products, Focal Point 3D Audio license their family of technologies for creating positional sound. The first of the technologies, Focal Point Type 1, uses DSP-based real-time binaural convolution and supports head tracking. Focal Point Type 2 also binaurally positions multiple sounds, but using only software on any standard PC (Focal Point 3D Audio has applied for a patent on this technology). The final technology, Focal Point Type 3, also called the Focal Point Audio Animator, is software-based and automatically creates positional 3-D audio to match objects and motions in 3-D graphic animation.

4.2.1 Acoustetron II

The Acoustetron II, from Crystal River Engineering, Inc., is a stand-alone, turn-key sound system for developing interactive, 3-D sound for real-time graphics workstations. The system is capable of the full spectrum of 3-D sound, including Doppler shifts, spatialization, and acoustic ray tracing of rooms and environments. Accordingly, it can create sounds that originate from exact positions in 3-D space and exhibit Doppler shifts as they travel past a listener. Additionally, sounds exist in a custom acoustic environment—such as a room, a cathedral, or even outdoors—where they bounce off surfaces, travel in the atmosphere, or pass through materials, and are reproduced in real time.

	Specification
Concurrent Wave Files and Spatial Sources per Card	9 at 44,100 Hz sample rate 16 at 22,050 Hz sample rate
Update Rate	44 Hz
Input	64x oversampled, 16-bit A/D converters
Output	8x oversampled, interpolating filters
Stereo Crosstalk	100 Hz-100 dBv, 1 kHz-80dBv 10kHz-60 dBv
Interface	RS-232

Figure 48. Acoustetron II

The Acoustron II is controlled from a central simulation 486DX2-based computer over a communication line. The client sends information such as audio source and listener positions to the server via RS-232. The server continually computes source, listener, and surface relations and velocities, and renders up to 16 separate spatialized sound sources accordingly. The audio output can be presented over headphones, earphones, or speakers. The ANSI C functions of the server allow fast, high-level development of 3-D sound spaces and easy integration into existing

Table 7. Characteristics of Commercially Available Auditory Products

Vendor	Product	Description	Product Type	Typical Applications	Price
Crystal River Engineering, Inc.	Acoustetron II	An 8/16 channel, turn-key AudioReality Server for use with SGI, Sun, HP, or PC client systems. Full spectrum 3-D sound, including Doppler shifts, spatialization, and acoustic ray tracing of rooms and environments	Stand-alone system	Computer simulation applications (e.g., vehicle and industrial training, location-based entertainment, video conferencing, and architectural walk-throughs)	Base System $9,995, Development System $12,495
	Proton	AudioReality plug-in for Digi Design's ProTools TDM system	Plug-in	Creation of spatialized 3-D tracks in the Digidesign ProTools dynamic mixing environment	$995
QSound Labs	QSystem I	Underlying QSound technology	Algorithm	Video games, PC applications, consumer electronics, and multimedia	Licensable
	QSystem II	8 channel real-time audio processor for professional recording environment	Stand-alone system		Digital System $8,500, Digital Plus System $9,750, Digital Plus/Analog System $17,000
	QCreator	Windows-based audio authoring tool for creation of preprocessed and run-time effects	Authoring tool		Included in licensed package
	QExpander	Plug-in for Sound Designer	Plug-in		$295
	QSYS/TDM	4 channel software plug-in version of QSystem for use with Mac-based Pro Tools III	Software plug-in		$995
	QMixer 32 and QMixer 95	32-bit DLLs for Windows 3.1 and Windows 95 for mixing and interactive processing of multiple.WAV files	DLL		Included in licensed package

Table 7. Characteristics of Commercially Available Auditory Products

Vendor	Product	Description	Product Type	Typical Applications	Price
Roland Corporation US	RSS-10	Two channel sound space processor utilizing a new DSP 3-D sound processor. Capable of generating a complete 360° reverb soundscape, including digital processing of reflections, delays, and Doppler effects	Dedicated, single rackmount unit	Professional studios and sophisticated project recording, as well as audio post production, broadcast, and sound design	$9,750
	SDX-330	Dimensional Expander to produce modulation effects processing	Dedicated, single rackmount unit	Professional recording, public address, and sound reinforcement, enhancement of individual instruments, and personal recording	$8,500
	SRV-330	Dimensional space reverb unit that provides reverb sounds with a total of 22 reverb algorithms, including sync, mono, and 3-D	Dedicated, single rackmount digital effects unit	Studio recording, public address, sound reinforcement, home recording, and live performance	$1,295
	SDE-330	Dimensional space delay unit with a total of 19 delay algorithms. Delay effects can have times as long as 2.9 sec and can be combined for a variety of delay effects	Dedicated, single rackmount digital effects unit	Studio recording, public address, sound reinforcement, home recording, and live performance	$1,295
Reality By Design	SoundStorm 3D	System generates from five 3-D up to thirty-two 2-D simultaneous sounds using Focal Point rendering algorithms. It can be networked for integration into simulations and supports both DIS and SIMNET protocols.	Stand-alone system	Military simulations	$20,000
Audio Cybernetics	VAPS	Process of actively or interactively encoding audio in 3-D	Currently offered only as a service	High-end recording and mixing of 3-D sound	Contact vendor

VEs. Specification details for the Acoustetron II are given in Figure 48. The base system is available for $9,995.

The acoustic ray tracing mode (4 concurrent wave files and spatial sources per system at 44,100 Hz sample rate) is supported as an option. Wave file recording and editing studio software, high-speed communication protocols, and quad speaker output are also available as options. In addition, there is a development option to control additional input and output devices from the host computer. Finally, it is worth noting that the Acoustetron II is supported by Coryphaeus' EasyScene, Paradigm Simulation's Vega, Sense8's WorldToolKit, Division's dVs, and Autodesk's CDK world building toolkits.

4.2.2 Protron

Crystal River Engineering, Inc.'s Protron enables sound designers to interactively place and move audio sources in a 3-D custom acoustic space. Using Protron, audio designers can create fully spatialized 3-D audio tracks in the Digidesign Pro Tools dynamic mixing environment.

Protron takes monophonic input and adds the psychoacoustic cues and environmental effects that make it appear to be located at a specific point in space, in a specific environment. Simple mouse operations are used to interactively place the source in 3-D; pop-up menus and sliders are provided to customize the acoustic space. Protron features:

- AudioReality sound field synthesis,
- Complete, interactive 3-D source position control,
- User selectable environment size and materials,
- Continuously adjustable parameters,
- Monophonic compatible output, and
- True RMS level meters.

Protron is fully compatible with Pro Tools III, and Pro Tools II with the TDM option. It has the same minimum Mac hardware and system software requirements as Pro Tools. The maximum number of sources which may be simultaneously rendered is equal to the number of available Pro Tools TDM DSPs. The price for Proton is $995.

4.2.3 Q Products

QSound Labs is an audio technology company that specializes in low cost sound localization and enhancement. Its products range from analog and digital hardware to stand-alone and add-on software. The three major products are QSystem I, a stereo speaker-based sound localization process, QSystem II, a headphone-based sound localization process, and QXpander, a stereo enhancement process for both speakers (QXpander I) and headphones (QXpander II) that can operate on existing stereo material.

The QSystem I process is the fundamental QSound technology. Using multiple monaural inputs, the process produces a stereo output signal that allows each input to be placed anywhere within an 180° arc around and in front of the listener. Although the QXpander is based on the same underlying technology as the QSystem I, it filters existing stereo images to expand the depth and separation of the sound field. QXpander also provides the ability to create 3-D sound in situations where the elements of the audio mix are not individually accessible.

In addition to these basic audio processing technologies, QSound also has available a number of tools. These include:

1. QCreator: Windows-based audio authoring tool for creation of preprocessed and run-time effects.
2. QSystem II: Eight-channel real-time audio processor for the professional recording environment.
3. QSYS/TDM: Four-channel software plug-in version of the QSystem for use with the Mac-based Pro Tools III from Digidesign.
4. QTOOLS/SF: A low-cost plug in for Sound Forger, by Sonic Foundry, that offers QXpander, static QSystem I, and a sample rate conversion tool.
5. QMixer 32 and QMixer 95: QMixer 32 is a 32-bit DLL for Windows 3.1 (with Win32s extensions) for mixing and interactive processing of multiple sound files. QMixer 95 provides the same technologies for Windows 95.

QSystem I is available for licensing and includes QCreator and QMixer, contact QSound Labs for price information. Prices for QSound II start at $8,500. QExpander and QSYS/TDM cost, respectively, $295 and $995.

4.2.4 RSS-10 Sound Space Processor

Preliminary Specification	
Nominal Input Level	XLR +4 dBm (Headroom: 20 dB), Phone -10 dBm (Headroom: 20 dB)
Nominal Output Level	XLR +4 dBm (Headroom: 20 dB), Phone -10 dBm (Headroom: 20 dB)
Source Distance	Max. 81 m (1 cm step), Max. 655 m (8 cm step)
Reverb Time	0.1-40 sec
Room Size	1-100 m
Interface	RS-422A/232C

Figure 49. RSS-10 Sound Space Processor

Roland Corporation US, a wholly-owned subsidiary of Roland Corporation Japan, focuses on high-end recording, sound reinforcement, and broadcasting applications. Products include digital hard disk recorders, 3-D sound processors, and other digital signal processors. The organization uses HRTF sound processing to create 3-D audio for playback via speakers. Provided the listener's position relative to the speakers is fixed in accordance with Roland's instructions, their transaural crosstalk cancellation techniques address the problem of sound designated for one ear entering the other, thereby producing strong 3-D effects.

The RSS-10 Roland Sound Space Processor is a two-channel system that utilizes a fast, new DSP 3-D sound processor. A complete 360° reverb soundscape can be generated, including digital processing of reflections, delays, and Doppler effect. Using the RSS-10, sounds can be placed or moved above, towards, or around the listener using standard stereo playback. As the sound source moves through the 3-D plane, the reverb reflections move accordingly, in real time. This creates 3-D sound with natural room ambiance. Specification details for the RSS-10 are given in Figure 49. Software control is available for both the Windows and Mac platforms. The price for the RSS-10 is $9,750.

4.2.5 SDX-330 Dimensional Expander

Another Roland Corporation US sound processor is the SDX-330 Dimensional Expander. This product is designed to produce high-quality modulation effects. It features Roland's proprietary 3-D spatial simulator sound localization technology, and creates unique 3-D effects with conventional 2-channel playback. The high performance capabilities of the SDX-330 come from the newly developed custom DSP chips that perform over 33 million computations per second. This level of digital processing results in enhanced resolution for precise, smooth, and warm effects, performed in a dedicated 384 kBytes of memory. The SDX-330 has discrete stereo processing, with two independent pairs of inputs and outputs that process the left and right channels independently to maintain true stereo localization of the direct sound within the effected sound.

Specification	
Signal Processing	A/D conversion: 16-bit, Delta-Sigma modulation, D/A Conversion: 16-bit, Delta-Sigma modulation
Sampling Frequency	44.1 kHz
Frequency Response	5 Hz to 70 kHz: -3/+0.3 dB (direct), 20 Hz to 20 kHz: -3/+0.3 dB (effect)
Nominal Input Level	-20/+4 dBm (selectable with Input Level Switch)
Input Impedance	300 kΩ (Input Level Switch: -20 dBm), 10 kΩ (Input Level Switch: +4 dBm)
Nominal Output Level	-20/+4 dB (selectable with Output Level Switch)
Output Impedance	1.5 kΩ (Output Level Switch: -20 dBm), 9 kΩ (Output Level Switch: +4 dBm)
Total Harmonic Distortion	< 0.012% (Direct, 1 kHz at nominal output level)
Dynamic Range	100 dB or greater (Direct), 90 dB or greater (Effect)

Figure 50. SDX-330 Dimensional Expander

The SDX-330 features 16 different algorithms, including stereo chorus, multiband choruses, classic chorus simulations, rotary, stereo 3-D chorus, and 3-D panner. Additionally, the SDX-330 offers extensive MIDI control. Effects parameters can be controlled in real time from a MIDI keyboard by using performance information such as note number, aftertouch, velocity, and control change messages sent from the modulation lever, data entry slider, or (optional) pedals. User patches can be bulk dumped to external MIDI devices, such as a sequencer or personal computer, for storage. Effects patches can be selected and parameters can also be changed in real time from a sequencer, enabling sequencer-controlled effects processing. Specification details for the SSE-330 are given in Figure 50. The cost of this product is $8,500.

4.2.6 SRV-330 Dimensional Space Reverb and SDE-330 Dimensional Space Delay

The SRV-330 Dimensional Space Reverb and SDE-330 Dimensional Space Delay units are two additional Roland Corporation USA products, that also rely on newly developed custom DSP chips for high speed digital processing. They provide high resolution, and permit the creation of a wide range of effects. Their sizable internal audio memory allows original sound quality to be maintained even after effects processing. Moreover, A/D/A conversion is 16-bit, with 30-bit internal signal processing and a sampling rate of 44.1 kHz. These technical characteristics deliver high-quality effects typically found only in professional recording equipment, including a flat frequency response of 20 Hz to 20 kHz, dynamic range of 90 dB or higher, and signal to noise ratios of 78 dB or greater.

Both units have stereo configurations (equipped with two inputs and two outputs), and can accommodate any input source. Since stereo algorithms perform internal signal processing for left and right channels independently, the exact stereo sound image localization of the direct sound is maintained. The inputs and outputs accommodate professional +4 dBm line level signals as well as the multipurpose -20 dBm level to meet a wide range of applications.

The SRV-330 Dimensional Space Reverb provides reverb sounds, with a total of 22 specially-developed reverb algorithms. These include:

- Sync Reverb: A stereo reverberation algorithm that creates basic reverb types like Hall, Room, and Plate, as well as extra-high density reverb that adds chorus. The unit also includes internal signal processing for both left and right channels.
- Mono Reverb: This algorithm combines two separate reverb sections, such as Hall and Room, to recreate sound environments with complex reverberation characteristics.
- 3-D Reverbs: These unique algorithms are based on Roland's proprietary 3-D spatial simulator sound localization technology. For example, the 3-D ambiance algorithm provides 24 early reflections that can be positioned at 12 locations for high-density reverb effects. The 3-D reverb algorithm adds dense rear reverberation to 12 early reflections positioned at six locations. Based on these algorithms, 300 preset patches are available. Any of these presets can be edited as desired and 100 customized patches can be stored to SRV-300 memory for instant recall. A built-in 3-band parametric pre-EQ permits precise tonal shaping to match the original sound texture and reverb type selected.

Conventional stereo reverberation units localize early reflections only within a two-channel stereo sound field. In real life 3-D sound environments, however, early reflections are localized at various points. Accurate early reflection reproduction recreates realistic acoustic spaces under different conditions. Algorithms that use the 3-D spatial simulator allow the SRV-330 to generate some extraordinary effects. For example, the 3-D ambiance algorithm generates up to 24 early reflections and positions them at a maximum of 12 locations in a 3-D sound field. The optimal delay time, phase differences, and filtering for each reflection are automatically calculated. Using the SRV-330, it is possible to get a brand-new palette of creative effects with just a conventional two-channel system, including simulation of high-ceiling rooms, hard-walled rooms, and many other specialized acoustic environments.

Specification

Signal Processing	A/D conversion: 16-bit, Delta-Sigma modulation, D/A converesion:16-bit, Delta-Sigma modulation
Sampling Frequency	44.1 kHz
Frequency Response	5 Hz to 70 kHz: -3/+0.3dB (direct), 20 Hz to 20 kHz: -3/+0.3 dB (effect)
Nominal Input Level	-20/+4 dBm (selectable with Input Level Switch)
Input Impedance	300 kΩ (Input Level Switch: -20 dBm), 10 kΩ (Input Level Switch: +4 dBm)
Nominal Output Level	-20/+4 dB (selectable with Output Level Switch)
Output Impedance	1.5 kΩ (Output Level Switch: -20 dBm), 9 kΩ (Output Level Switch: +4 dBm)
Total Harmonic Distortion	<0.012% (Direct, 1 kHz at nominal output level), <0.02% (Effect, 1 kHz at nominal output level)
Dynamic Range	100 dB or greater (Direct), 90 dB or greater (Effect)

Figure 51. SDE-330 Dimension Space Delay

The SDE-330 Dimensional Space Delay's unique effects algorithms also use Roland's 3-D spatial simulator. For example, the multitap space delay algorithm generates up to eight delay taps that can be positioned at any point in a 3-D sound field. This allows multidelays to echo around a 360° sound field with only conventional 2-channel stereo systems. The SDE-330 has a total of 19 delay algorithms, and includes new algorithms specifically designed for the unit. Delay effects can have times as long as 2.9 seconds and can be combined for an variety of delay effects:

- Stereo Delay: Allows realistic sound effects—from simple delays, all the way up to complex cross feedback—to be easily created through separate processing of the left and right channels.
- Quad Delay: The SDE-330 includes four delay sections connected in series, to produce dense delays.
- Multitap Delay: This algorithm allows delay effects to be maximized, and can use up to eight taps.
- Gate/Duck Delay: The built-in gating function allows the creation of special delay effects. For instance, a delay signal can be switched on and off alternately at the gate threshold setting.
- Pitch Shift Delay: The SDE-330 allows delay effects to be enhanced through the pitch-shifting of the four delay taps by ±1 octaves.
- 3-D Delays: The SDE-330 is distinguished by several algorithms for radical effects. For instance, the multitap space delay allows delay taps to be positioned at different points around a 360° circle of sound. Using the 3-D chorus, richer and fatter textures than those of conventional chorus units can be achieved. This algorithm positions multiple pitch-shifted signals at many different points all around the direct signal.

Based on these algorithms, the SDE-330 incorporates 100 preset effects patches. More creative effects can be obtained by customizing the effects patches and storing up to 200 user patches in SDE-330 internal memory. The unit also includes a built-in 3-band parametric EQ that provides the versatility to make fine tonal shading or radical effects without having to employ a separate EQ unit.

Both the SRV-330 and SDE-330 are available for $1,295 each.

4.2.7 SoundStorm 3D

From Reality By Design, SoundStorm 3D is a stand-alone Pentium PC-based system capable of generating from five 3-D up to thirty-two 2-D simultaneous sounds. The system operates in a Unix environment and can be networked with a variety of computers, such as Sun and Silicon Graphics workstations, for integration into simulations. The system hardware platform is augmented with dual 16-bit sound cards, four magnetically shielded speakers, and Ethernet IP networking. It supports both the Distributed Interactive Simulation (DIS) and SIMNET protocols, and provides a sound effect library suitable for military applications. A set of customization tools allows users to record and manipulate additional sounds. The sound generation is performed independently of the simulation and linked to it by associating particular sounds with specific simulation entities. The actual sound generation employs Focal Point binaural rendering algorithms. Sound Storm 3D is available for $20,000.

4.2.8 Virtual Audio Processing System

The Virtual Audio Processing System (VAPS) of Audio Cybernetics creates sounds containing significant psychoacoustic information to fool human sensory organs into perceiving that the sound is actually occurring in the physical reality of 3-D space. The sounds can be reproduced on a conventional stereo system, surround system, or headphones, without special decoding equipment.

One of the most significant features of VAPS is that it not only requires no special equipment to decode, but also eliminates the sharp “sweet spot” limitations present in many other systems. This results in a wide listening area that allows for more listener mobility than is available with more conventional systems. Since the process is encoded directly onto the recording medium, the need for special processing devices for playback is eliminated. Rather than being based on anatomical simulation (artificial head architecture), the VAPS uses a highly-detailed mathematical model to define the variables of human audition in the creation of 3-D audio effects. The cited results are sounds that reportedly not only are perceived to occur in 3-D space, but that may also have the synesthetic properties that fool the listener into believing that subtle “tactile” perceptions were received.

The VAPS also allows the user to give some spatial qualities to pre-recorded material, or to “move things around in the mix.” Additionally, it allows the creation of sophisticated room simulations with the proximity effects that are necessary for VEs.

The VAPS is part of the on-going research of Audio Cybernetics. Other areas being investigated are the development of a 3-D sound recording chip and the feasibility of producing VLSI components for 3-D audio processing. Currently, the VAPS process is only offered as a service. Audio Cybernetics will provide the virtual audio equipment at a daily rental along with a trained engineer. Expert consultation and full virtual audio production services are also available. Contact Audio Cybernetics for pricing information.

4.3 Current Research and Development

Research and development efforts for the generation of realistic 3-D sound have been conducted since the 1880s. Within the past twenty years, these efforts have escalated, primarily due to impetus received from the entertainment industry. With the relatively recent advances in signal processing technology, acceptable results in 3-D sound are now available for reasonable prices. As a result, increasing numbers of potential applications are being found for this technology. However, although the current state-of-the-art in virtual audio technology is far advanced from where it was even ten years ago, many facets still are not well understood, both in the area of basic research concerning human auditory perception and in the way technology can be improved and applied. The following paragraphs discuss some of the current work in the field of virtual auditory displays that is being performed today. Although other work is also continuing, no information was available at the time of this writing.

4.3.1 NASA Ames

A pioneer in the field of virtual audio, NASA Ames has been working with both the University of Wisconsin - Madison and Crystal River Engineering to develop increasingly sophisticated audio spatial displays, and use these in practical, real-world applications. As a result of their efforts, the Convolvotron, the world's first real-time, 3-D acoustic display, was developed in 1987. Recent work has included investigations of heads-up auditory displays for traffic collision avoidance systems (Begault, 1993), localization in virtual acoustic displays (Wenzel, 1992), multi-channel spatial auditory displays for speech communications (Begault and Erbe, 1993), localization using non-individualized HRTFs (Wenzel et al, 1993), virtual acoustic displays for teleconferencing (Begault, 1995), and headphone localization of speech (Begault and Wenzel, 1993).

Currently, NASA Ames is performing multiple projects relating to the development of virtual acoustic environments, and has recently been awarded a contract to perform both basic and applied research and technology development to implement 3-D auditory displays for improved operative efficiency and safety. As part of this program, NASA Ames will conduct perceptual studies of human sound localization using techniques developed for real-time synthesis of 3-D sound over headphones and apply this knowledge for both enhancing and perceptually validating the advanced acoustic display systems that have been developed as part of their ongoing spatial sound project.

The binaural listening system will enable an astronaut, ground-controller, or other human operator to take advantage of their natural ability to localize sounds in 3-D space, and is intended to be used to enhance situational awareness, improve segregation of multiple audio signals through selective attention, and provide a means of detecting a desired signal against noise for enhanced speech intelligibility. NASA Ames also plans on developing an in-house capability to measure HRTFs and a real-time room modeling technology in the near future.

4.3.2 Naval Postgraduate School

As part of their Naval Postgraduate School Networked Vehicle Simulator (NPSNET) research effort, the Naval Postgraduate School is another institution currently investigating the practical application of virtual audio technology. Overall, the research effort is addressing many issues, including large-scale networking of VEs, representation of the human body in VEs, and the integration of hypermedia into VEs. The NPSNET itself was developed as student-written, real-time, networked software running on commercial, off-the-shelf workstations. Although originally envisioned as a low cost, government-owned, workstation-based visual simulator, it has evolved to include many facets of VEs, including virtual audio.

The NPSNET Polyphonic Audio Spatializer (NPSNET-PAS) can play sounds either in the spatialized sound mode or directly from the computer's built-in sound board. It keeps track of each sound occurrence in the virtual world (for example, a detonation or vehicle) by "listening" to the packets on the network. The program then determines whether the source of the sound is within hearing range of the player and calculates the correct volume and direction of the sound to be played by the Emax II. The sound is delayed to simulate the correct distance between the occurrence and the listener.

The previous version of the MIDI-based sound system for NPSNET could only generate aural cues via free-field format in 2-D. To increase the effectiveness of the auditory channel, a sound system was needed that could generate aural cues via free-field format in 3-D. To do this, hardware limitations of the NPSNET-PAS sound generating equipment were identified and more capable off-the-shelf sound equipment was procured. In software, new algorithms were developed to properly distribute the total volume of a virtual sound source to a cube-linked configuration of eight loudspeakers and to enhance the ability to localize a sound source. Synthetic reverberation using digital signal processors was added to enhance perceptual distance of the generated aural cues.

4.4 Summary and Expectations

Virtual audio has a value far beyond the music industry for which it was originally developed or for gaming where it is commonly used. Many other applications have been found in such fields as medicine, training and simulation, and architecture. Further, as the technology matures, it is likely even more applications will be found and it will become widely used in VEs.

Although high-end virtual audio approximates the real world, the technology is still far from perfect. In the near future, work is expected to continue on improving the realism and full-surround capabilities of the technology. To do this, better algorithms need to be developed, based on a more thorough understanding of how humans perceive sounds. Other research that needs to be accomplished is a determination of what auditory stimuli are necessary to simulate various environmental sounds. This is closely tied to the need to determine how "realistic" an auditory simulation must be in order to result in the desired effect. In addition, generic HRTFs that maximize accurate localization of sounds in space need to be developed and made publicly available. Linked to this is the need to develop refined algorithms for use in recording sounds, and mathematical models to simulate human hearing organs. Other issues that must be addressed include better control of ambient noise that distracts from the reality of the virtual environment, elimination of unwanted reflections, and technology that will allow the listener to move in relation to the sound source without noticeable distortion of the sound quality.

Since the digital synthesizers used in virtual audio were originally developed for the music industry, synthesized speech and sound effects are not well developed. Although there has been much research in this area, this is a difficult problem and further work is required before automatic virtual audio can be produced in real time. This work needs to address the necessary spectrum of sounds and how they are affected by changes in stimuli, and algorithms need to be created. Similarly, much work still needs to be done in the area of speech synthesis. The significant factors in natural speech yet have to be identified, and ways to synthetically reproduce natural speech without significant deterioration in perceptual quality developed.

Currently, the use of high-end virtual audio technology requires specialized technicians. As the technology matures, one of the expected trends will be an increase in the ease of use. This will include more user-friendly interfaces for the technology, such as standardized options that allow inexperienced or non-professional users to easily approximate professional results, and better human-machine interfaces for power users and professionals who need or desire a wider range of options. As part of this trend, generic HRTFs will need to be synthesized, or current HRTFs modified, so that they can be easily applied by inexperienced users.

Finally, validation studies need to be performed to determine the utility of virtual audio in various applications. It is only in this way that the technology can be fine-tuned and properly applied.

In the near future, as digital signal processing technology becomes less expensive, it is expected that virtual audio will become more widely available at a lower cost. This is happening to some extent already with many dedicated game systems, major computer companies, and audio chipset manufactures licensing low-end virtual audio technology. As a result of increasing availability and the lower cost of the technology, virtual audio should become a common component of VE systems within the next five years.