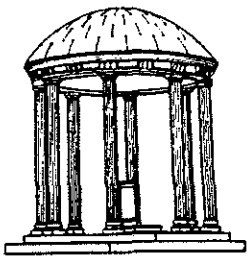
The Use of Sound in Virtual Worlds: A Primer

TR91-050

November, 1991

William Brown



The University of North Carolina at Chapel Hill
Department of Computer Science
CB#3175, Sitterson Hall
Chapel Hill, NC 27599-3175

The Use of Sound in Virtual Worlds: A Primer

William Brown
Department of Computer Science
University of North Carolina
May 7, 1990

1.0 Introduction

In 1965 Ivan Sutherland first proposed the *Ultimate Display*, a display in which computer generated images would behave just as their real counterparts. Virtual apple pie would really smell and taste like apple pie and virtual bullets would be fatal. In the past few years, the University of North Carolina's Department of Computer Science has experimented with virtual reality, creating computer graphic representations of buildings and protein molecules, using force feedback mechanisms, and using several generations of Head Mounted Displays to provide a sense of computer generated reality. One area that has not been explored in much depth, however, is the use of sound to enhance the illusion of reality.

This report will provide an introduction to non-speech auditory feedback in a virtual environment. It is intended to be an introduction for the researcher interested in pursuing the use of sound feedback in virtual worlds applications. The reader is expected to have some knowledge of introductory physics and computer science. This report is a primer; it will sacrifice depth in favor of breadth, providing pointers to information for those interested in pursuing the topic further.

The report is divided into six main sections. The first will introduce the physics of sound on a basic level. The second will discuss the anatomy of the ear and how one actually hears. The third will discuss sound perception. The fourth section will review some of the work that has been done in human factors research studying the potential of sound interfaces. The fifth will explore application areas that could potentially benefit from sound feedback. The discussion of human factors and applications will be limited to subjects relevant to virtual worlds research. The final section will list some ways sound can be added to a virtual environment using existing or readily available software and hardware.

2.0 Fundamentals

Before describing how sound is perceived and what sorts of computational procedures can be used to create sounds for virtual environments, this section will define terms and concepts that will be used in later sections. It will describe what sound is, introduce the reader to sampling, and describe the dimensions of timbre, envelope, and reverberation.

2.1 What is Sound?

Sound can be defined as the wave phenomenon resulting from changes in air pressure (amplitude) occurring at frequencies in the audible range[1]. Air density alternates between a compression phase and a rarefaction phase about a normal air pressure of, roughly, 100,000 pascals[2].

The sound wave propagates outward in all directions from a vibrating body, but the individual molecules of air vibrate about an average resting place. Sound waves are longitudinal waves in which air molecules move in the same direction as the wave as opposed to transverse waves, in which points of the medium move perpendicular to the medium as waves in a rope. As it moves outward, the sound wave weakens according to the inverse square law [3]and is subject to reflections and refractions as other wave phenomena.

2.1.1 Waveforms and Spectra

Sound can be represented as a *waveform*, a graph of air pressure (about the average) vs. time. Thus, an abstract waveform can be associated with a characteristic sound such as a concert "A" with the waveform of a sine wave with a frequency of 440 cycles/second. More complex sounds are associated with more complex waveforms. While these are the conditions that exist in nature, Jean Baptiste Fourier demonstrated that any complex, "natural" sound could be constructed by combining multiple simple waveforms to form a complex waveform.[4]

Fourier's analysis takes advantage of an assumption of linearity in the auditory system to conclude that multiple waves can be combined to form any possible waveform[5]. In actuality, parts of the

---

[1] roughly 20 Hz to 20 kHz

[2] a pascal (PA) is a measure of pressure (SI) -- newtons/m^2, where a newton is a measure of force -- kg m/sec^2

[3]the sound intensity is proportional to $1/r^2$, where r is the distance from the source of the sound, i.e. the radius of a sphere

[4]regardless of whether it's waveform is periodic, i.e. repeating at regular intervals

[5]For a system to be linear, two conditions must hold:

auditory system behave as though they were approximately linear, while others behave in a grossly non-linear way.

Fourier also demonstrated that sound can be described either by its waveform or by its distribution of energy vs. frequency (its *spectrum*). Thus, a signal such as a sound can be described in either the time domain or the frequency domain.

Unfortunately, waveforms do not map well to perceived sounds. Two identical waveforms of different phase[1] will appear different and adding multiple waveforms of differing phase can create complex waveforms that appear very different. Yet the differences in phase that make two complex waveforms look different do not produce sounds that are perceived as different. For this reason, it is argued that waveforms are not useful in analyzing sounds and instead, spectra (which do not have a time component and do not vary with phase) should be used. This report will continue to describe sounds using waveforms because they have come into common use and because most people tend to think in the time domain more easily than the frequency domain. Much of the analysis described below can be done using spectra as well as waveforms.

In generating a spectrum, or map of the variation of energy as frequency varies, it is necessary to measure the *sound energy* of a signal. This is done by calculating the root mean square (rms) of the air pressure. The instantaneous values of sound pressure for each point along the wave are squared, then these squared values are added together, averaging over time, and the square root of the sum is taken (this avoids the problem of having the mean pressure zero since the negative numbers become positive when squared).

Common waveforms and spectra will be discussed next. Each form has a characteristic sound quality. A knowledge of the differences can be useful when attempting to synthesize sounds.[2]

Superposition -- the output of the system (in response to a number of independent inputs presented simultaneously) should equal the sum of the outputs that would be obtained from each input taken independently.
Homogeneity -- if the input is changed in magnitude by a factor of k, then the output should also change by a factor of k.

[1] *Phase* is the measure of how far through the cycle a wave has advanced relative to some fixed point in time. For example, for readers familiar with high school trigonometry, sine and cosine waves can be thought of as the same waveform just out of phase by $\pi/2$ radians.

[2] The Acoustical Society of America distributes a compact disk of auditory demonstrations that complements this material. It costs $20 and can be purchased by writing the A.S.A at 500 Sunnyside Blvd., Woodbury, N.Y. 11797.

2.1.1.1 Sine Wave

Three parameters need to be defined to describe a sine wave: *frequency,* or number of wave cycles per second; *amplitude* , the amount of pressure variation about the mean or normal air pressure; and phase , the measure of how far through the cycle the wave has advanced.

Frequency -- measured in Hz.

20 -- lowest audible frequency
31 -- lowest note of a piano
250 -- middle C
440 -- concert A
4k -- highest note on a piano
16-20k -- highest audible frequency

A *complex* waveform is one that is composed of a number of separate waveforms each superimposed on the other. The frequency components of the complex tone are called *harmonics* or partials. These additional harmonics add "richness" to a sound and are an important element of the complex dimension of timbre to be discussed in a later section. The simplest complex waveform is one that is periodic. It is composed of tones that are each an integral multiple of some common fundamental component. The *fundamental* component, then, has the lowest frequency of any component of the complex tone and this frequency is equal to the repetition rate over time of the complex waveform as a whole.

A sine wave can be used to represent a "pure" tone, a sound with no overtones or harmonics. Choosing time as the horizontal axis, our curve would start at the origin (time 0, normal air pressure) and move upward and to the right as the air pressure increases with time, reaching a peak, declining to normal air pressure at the horizontal axis, decreasing still further as the wave enters the rarefaction phase where air pressure is less than normal, reaching a minimum and returning to normal to complete one cycle. The spectrum of a sine wave is a single peak energy at the frequency of the sine wave.

2.1.1.2 Triangular Wave

A triangular wave is constructed from only the odd numbered harmonics. The amplitude of the harmonics fall off with frequency in proportion to the square of the harmonic number. For example, the amplitude of the 5th harmonic is 1/25th the amplitude of the fundamental.

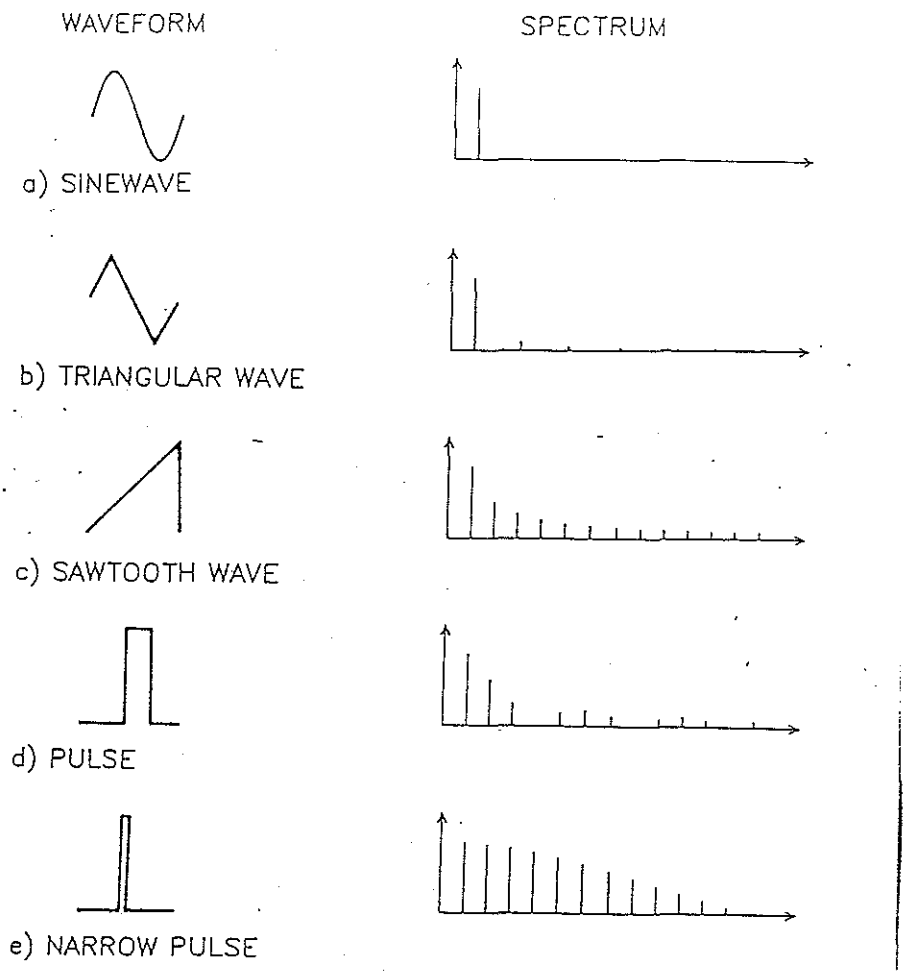


FIGURE 2.18 The spectra of some simple waveforms.

2.1.1.3 Sawtooth (ramp)

In the case of the sawtooth wave, all harmonics are present. They diminish in direct proportion to the harmonic number.

2.1.1.4 Pulse (Square Wave)

A square wave can (theoretically) be constructed by adding together harmonics, using the principle of superposition. The components of such a wave are odd harmonics whose amplitudes diminish with increasing harmonic number in proportion to the harmonic number. For example, the amplitude of the 7th harmonic is 1/7th the amplitude of the fundamental, the first harmonic. The more harmonics

added, the squarer the wave becomes but a perfect square wave requires an infinite number of harmonics.

2.1.1.5 White Noise

White noise in the frequency domain represents a signal of all frequencies at a single energy level. In the time domain, white noise is an example of a waveform that is not periodic in time. There are no repeating cycles. Not all noise is white, for instance a noise might be *band limited*, containing energy only in a certain band of frequencies. *Bandwidth* refers to the range of frequencies involved in a tone or noise.

This section has introduced the concept of waveform and spectrum. It has briefly discussed the relationship between these two ways of representing sounds graphically in either the time domain, on a time varying plot of the waveform, or the frequency domain, on a frequency varying plot of the spectrum. The next section describes sound measures that define what it means to say that one sound is louder than another.

2.1.2 Metrics

The instruments used to measure sound levels, such as microphones, respond to changes in air pressure. Their output is, therefore, proportional to the amplitude of the sound. It is more useful to specify sound levels in terms of intensity, the *sound power* transmitted through a given area in a sound field. In air, there is a simple relationship between the amplitude of a plane (flat-fronted) sound wave in a free field (in the absence of reflected sound) and the acoustic intensity; intensity is the square of the amplitude.

The range between the smallest and largest sound pressures that the ear can sense as sound is approximately a million to one. This extreme range is one reason a logarithmic scale is used to describe the perception of loudness.

Before discussing measures of loudness, it's important to discuss measures of objective sound pressure and power to have a concrete basis with which to compare the perception of loudness

> Decibels (dB) -- a logarithmic ratio of two quantities, one of which is a reference quantity, in the case of sound level 10^{-12} W/m^2. The bel, (10 decibels), was named for Alexander Graham Bell. The reference quantity was chosen because it is near the lower limit of audibility.

Sound Pressure Level (SPL) -- amplitude --10 log (p^2_1/p^2_0) dB, re p_0, where p_0 is a reference sound pressure, 20×10^{-6} PA.

Sound Power Level (PWL) -- intensity --10 log (w_1/w_0) dB, re w_0, where w_0 is a reference sound power, 10^{-12} watt. As stated earlier, sound power is roughly proportional to the square of sound pressure.

Sound travels through the air in a free progressive spherical wave (assuming a non-directional noise source) at approximately 335 m/sec. The total energy of the wavefront remains constant. This implies that, as the sphere expands, the sound pressure diminishes per unit area (explaining the inverse square law described earlier).

Sound *pressure* diminishes inversely with distance as follows:

$p = A/r \cos k(r - ct)$ where r -- radius of the spherical wavefront
t -- time of the propagation of the wave
c -- a constant
A -- area of the sphere

The *power* of the sound wave decreases as a square of the distance from the source:

$I = w/4\pi r^2$ = w/(the area of a sphere)
where I -- sound power in watts/m^2
w -- the sound power of the source in watts

Objective measures of sound pressure and sound power are related in complex ways to actual perceived loudness. This relationship will be discussed in more depth in the later section on sound perception. This section has introduced the difference between representing sound in the time domain and the frequency domain and introduced several measures of sound levels. The next section will briefly introduce the areas of sampling and filtering.

2.2 Sampling

Since sound is a continuous (as opposed to a discrete) signal, in order for a computer to handle sound information, the continuous signal must be broken up into discrete pieces to be encoded digitally. The

process of measuring a continuous signal at discrete intervals of equal duration to generate a sequence of numbers is known as *sampling*. A sequence of samples x[n] is obtained from a continuous signal $x_C(t)$ according to...

$$x[n] = x_C(nT), \quad -\infty < n < \infty,$$

where T is the sampling period[1], and $f_S = 1/T$, is the sampling frequency, in samples per second

Sampling is usually done by an analog-to-digital (A/D) converter (analog referring to the continuous input signal and digital referring to the discrete output signal).

In order to reconstruct a signal from samples, there must be more than two samples taken every cycle of the original signal. To illustrate, if exactly two samples were taken per cycle, it could be that both samples are taken at the point where the signal has zero amplitude. If that were the case, it would not be possible to recognize any signal at all. With less than two samples per cycle, it is possible to identify a signal, though it may not be the signal originally sampled (see section 2.2.2 Aliasing).

Mathematical analysis of the sampling process reveals that, as long as the sampling rate is more than twice the fastest rate of change of the signal being processed, no error will be introduced by discrete sampling. Or, more practically, the highest frequency reproducible without error is less than half the sampling rate. This maximum, $f_S/2$ (where f_S is the sampling frequency), is termed the *Nyquist frequency* and is the theoretically highest frequency that can be represented by a digital audio system.

2.2.1 Aliasing

Given a sampling rate of 40 cycle/second, the Nyquist frequency for that sampling rate is 20 Hz. A signal input at 30 Hz and sampled at 40 Hz is indistinguishable from a signal input at 10 Hz (see the figure). When the signal is converted back from digital to analog form, the output frequency will be 10 Hz. This alteration of frequency caused by sampling signals at greater than the Nyquist rate is known as *aliasing*. In practice, a low pass filter is used to cut out signals above the Nyquist frequency (Filtering will be discussed in more detail in the next section.) Aliasing can also occur in computer generated sounds if the desired frequency is greater than half the Nyquist rate of the sampling.

[1]period -- the length of time between samples (sec/sample), it's the inverse of frequency (samples/sec)

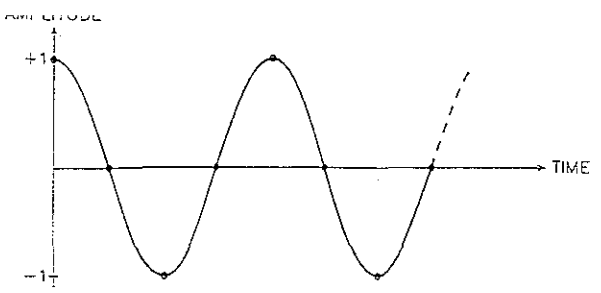


FIGURE 1.14 Sampling a 10-kHz sinusiodal tone at a 40-kHz rate.

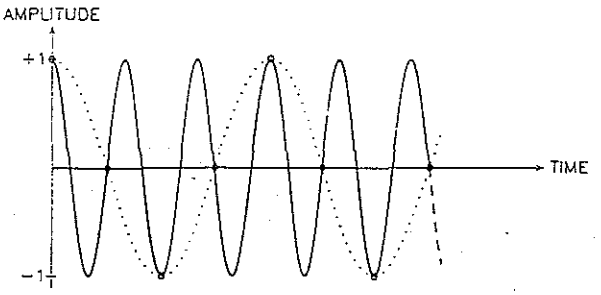


FIGURE 1.15 Sampling a 30-kHz sinusoidal tone at a 40-kHz rate. The samples also describe a 10-kHz sinusoid as shown by the dotted line.

2.2.2 Filtering

To ensure that all signals input to a system are below the Nyquist frequency, a low-pass filter can be applied to the analog signal before the A/D conversion. A filter's function is to separate signals based on their frequencies, passing signals of some frequencies while cutting off, or drastically reducing the amplitude of, signals of other frequencies. A low-pass filter would allow frequencies below the Nyquist rate to pass through unchanged while cutting out signals of higher frequencies. Unfortunately, low pass filters are not perfect and generally only pass a frequency range of 40% of the sampling rate rather than the full 50%. As a result, a 40 kHz sampling rate will only result in reproducible frequencies of below 16 kHz (40% of 40kHz). The upper limit of human hearing is near 20 kHz, which could be reached with a sampling rate of 50%. For economic reasons, many systems use lower rates.

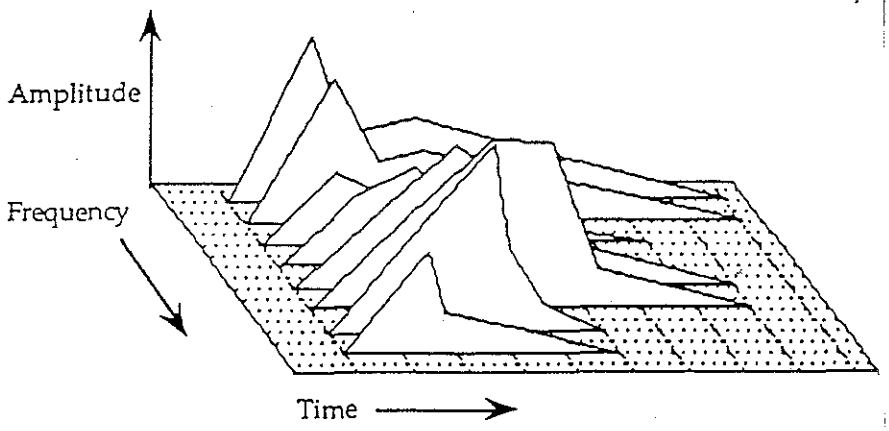


Figure 8. A time varying spectral plot of a complex sound.

2.3 Timbre

Timbre is the characteristic tone quality of a particular class of sounds or, according to the American Standards Association (1960), "that attribute of auditory sensation in terms of which a listener can

judge that two sounds similarly presented and having the same loudness and pitch are dissimilar". In other words, the differences in sound that allow us to distinguish between a middle "C" played on a piano and a saxophone. Timbre is determined primarily by the frequency spectrum of the combination of harmonics for steady tones. Recently, it has been suggested [Buxton] that the relative changes over time of the harmonics are as important or more important than the actual frequencies of the harmonics.

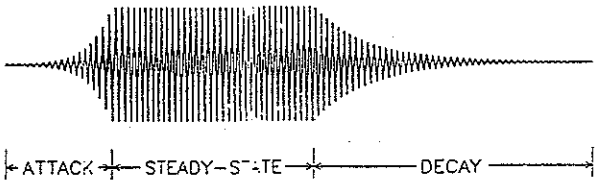


FIGURE 2.14 The three principal segments of a tone which takes the form of a simplified Helmholtz model.

2.4 Envelope

The variation of amplitude over time beyond the cyclic variations of the simple waveform is another important aspect of sound. A musical note can be made to sound from a piano or a woodwind simply be changing the envelope. In his book, *On the Sensations of Tone*, Hermann von Helmholtz characterized tones as consisting of a waveform enclosed in an amplitude envelope consisting of three parts: the rise time or *attack* , the steady-state or *sustain* , and the decay or *release* . During the attack, the amplitude grows from zero to its peak, during the sustain, the amplitude is constant, and during the release, the sound dies away.

2.5 Reverberation

Reverberation, or echo, is actually a combination of effects. Consider a listener sitting in a large concert hall. Only a small part of the sound created by the orchestra reaches the listener directly. Sounds have also reflected off walls, ceiling and the floor. This lengthens duration of the sound. In addition, the amplitude is inversely proportional to the amount of distance it travels so later reverberations will have a lower amplitude and the sound will seem to decay.

The reverberation time is the time required for a sound to die away to 1/1000 (-60 dB) of its amplitude after its source is shut off. Reverberation time varies with frequency, a well designed concert hall, for example, will be constructed to accentuate the lower frequencies so that they will fade more slowly.

The amount of time between hearing the direct sound and hearing the first reflection gives a sense of size to the room. A delay of over 50 ms will result in distinct echoes (and the illusion of a large space) and a short delay of less than 5 ms will result in the perception of a small space. The delay in most good concert halls is between 10 and 20 ms[1].(For a description of how reverberation can be synthesized, see Appendix I.)

This section has provided a brief introduction to the physics of sound and described several objective dimensions of sound. The discussion now shifts to perceptual aspects of sound.

[1]ms -- one millisecond = 10^{-3} seconds

3.0 The Anatomy of the Ear

While not all perception of sound is done with the ears (it is also possible to perceive low frequencies and high frequencies through skull bones without actually hearing them), the most important perceptual organ for sound is the ear. The ear is functionally and geographically broken into three suborgans, the outer, middle, and inner ear.

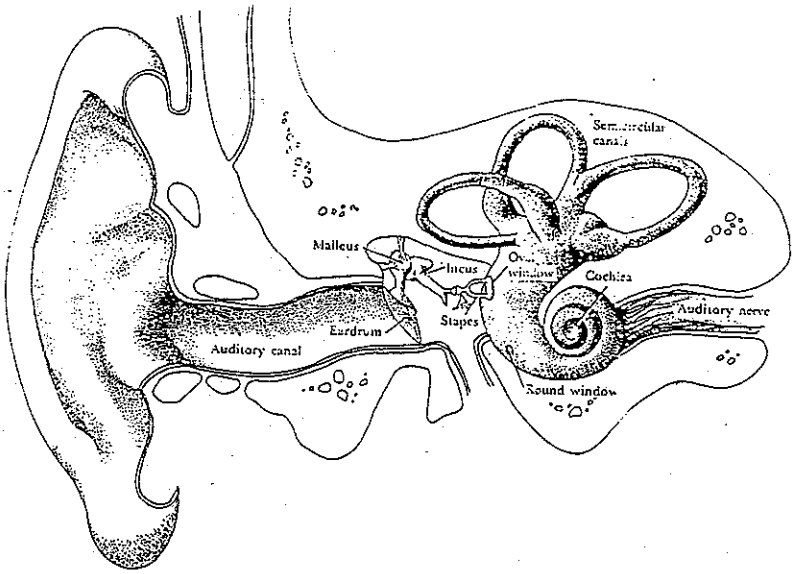


FIG. 1.6 Illustration of the structure of the peripheral auditory system showing the outer, middle and inner ear. From *Human Information Processing*, by Lindsey, P. H. and Norman, D. A. (1972), by permission of the authors.

3.1 The Outer Ear

The outer ear consists of the pinna (the part actually visible) and the auditory canal . The pinna is important as its shape modifies incoming sounds, particularly at high frequencies, and is important in the ear's ability to localize sounds.

3.2 The Middle Ear

Once in the ear, sound waves travel down the auditory canal to the tympanic membrane (the eardrum), where the propagating pressure differentials cause the membrane to vibrate. These vibrations are then transmitted through the middle ear by three small bones or ossicles, the malleus (hammer), incus (anvil), and stapes (stirrup), to another membrane, the *oval window* .

The middle ear acts as a transformer to improve sound transmission between the outer and inner ears. Acting as an impedance matching device, the middle ear uses the difference in the effective areas of the tympanic membrane and the oval window and the lever action of the ossicles to transform sound information from the air outside the tympanic membrane to the fluid in the cochlea beyond the oval window.

In addition to acting as a transformer, the middle ear has the ability to cut off sound transfer to the inner ear. The ossicles have tiny muscles attached which are able to contract when exposed to intense sounds. This reflex is not fast enough, however, to prevent ear damage from impulsive sounds such as gun shots or hammer blows. In addition, the middle ear cuts down the audibility of self-generated sounds such as speech.

3.3 The Inner Ear

The most important part of the ear for perception is the *cochlea,* or inner ear . The cochlea has rigid, bony walls and is filled with incompressible fluids (the endolymph). It is divided lengthwise by two membranes, Reissner's membrane and the *basilar membrane* . The base, or basilar, portion of the cochlea resides at the oval window (the big end) and the apex lies at the inner tip. At the apex, there is a small opening, the helicotrema , that allows fluid to pass between the two chambers of the cochlea.

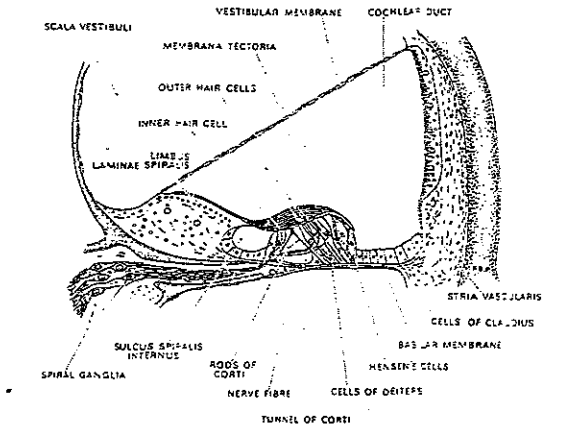


FIG. 1.12 Cross-section of the cochlea, showing the organ of Corti. The actual receptors are the hair cells lying on either side of the tunnel of Corti. From Hamilton, *Textbook of Human Anatomy*, Macmillan (1976).

When the oval window is pushed inward, the fluid in the cochlea flows around the helicotrema and causes a corresponding outward push in a second opening, the *round window.* Sounds coming in set the oval window in motion which causes a ripple through the fluid in turn setting the basilar membrane in

motion. This pattern of motion in the basilar membrane can also result from vibrations in the bones of the head, resulting from speech.

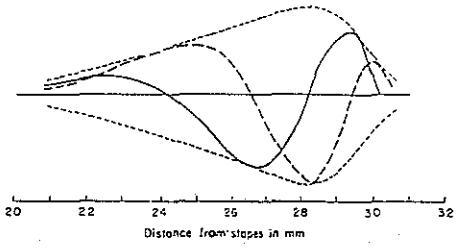


FIG. 1.7 The instantaneous displacement of the cochlear partition at two successive instants in time, derived from a cochlear model. The pattern moves from left to right, building up gradually with distance, and decaying rapidly beyond the point of maximal displacement. The dotted line represents the envelope traced out by the amplitude peaks in the waveform. From von Békésy (1947), by permission of *J. Acoust. Soc. Am.*

Vibrations at the oval window result in a travelling wave which moves along the basilar membrane toward the apex . The amplitude of this wave increases slowly at first, then decreases abruptly. Mechanical properties in the basilar membrane vary from the base to the apex. At the base the membrane is narrow and stiff, while at the apex it is wider and less stiff. Because of these properties, the position of the peak in the vibration differs according to the frequency of the vibration. High frequencies result in a maximum displacement near the oval window and less activity along the rest of the membrane, while low frequencies produce a vibration along the length of the basilar membrane with a maximum before the end of the membrane. In this way, different frequencies will produce different levels of activity at different places along the basilar membrane and the ear behaves as a crude Fourier analyzer .

The position along the basilar membrane excited most by a given frequency varies approximately with the log of the frequency (for freq. above 500 Hz). In addition, the relative bandwidths of the patterns of vibration on the basilar membrane are relatively constant throughout. Together, these imply that the patterns of vibrations for higher frequency harmonics will overlap much more than for lower harmonics.

Attached to the basilar membrane are hair cells which form part of the *organ of Corti* and above those hairs lies a gelatinous structure called the tectorial membrane. This membrane is hinged on one side so that, as the basilar membrane moves, the tectorial membrane rubs against the hairs. It is believed that this rubbing of the hairs on the organ of Corti is what causes the neurons of the auditory nerve to fire, creating a sensation perceived as sound. When the basilar membrane moves back down, the hairs no longer rub against the tectorial membrane and no nerve fibers will fire. For this reason, nerve firings tend to occur on the rarefaction phase of a sound wave.

3.4 The Auditory Nerve

Nerve fibers exhibit phase locking. A given fiber may not fire on every cycle of the waveform, but when they do fire, they tend to fire at the same phase along the waveform. Information, then, about a sound is carried not only by which fiber fires, but also by the period at which it fires.

Individual auditory nerve fibers respond better to some frequencies than to others. The threshold of sensitivity of the nerve fiber tends to be lower at some given frequency (called the *Characteristic Frequency*, CF, of the nerve fiber) that corresponds to the fiber's position along the basilar membrane . These fibers are combined into the auditory nerve in an orderly way so that fibers with high CFs are found in the periphery of the nerve bundle and fibers with low CFs are found in the center.

Within the cerebral cortex of the brain not as much is known about how sound information is processed, certain nerve cells seem to respond to more complex types of stimuli such as clicks, bursts of noise, "kissing" sounds (in cats), or interaural (between ear) time and intensity differences (in monkeys).

This section has briefly described the anatomy of the ear. The next section will discuss the way the ear perceives sounds and how this perception relates to the reality.

4.0 Perception of Sound

In considering ways to use sound in computer systems, it is worthwhile to consider how perceived sound differs from actual sound. For example, how does perceived loudness differ from actual air pressure differentials? How does perceived pitch differ from actual frequency? This section will address the perception of loudness, pitch, duration, and masking.

4.1 Loudness

The *loudness* of a sound is a measure of the subjective response to its amplitude. Loudness depends not only on amplitude but on frequency, bandwidth, spectral composition, and duration of the sound. The Just Noticeable Difference (JND) is the minimum detectable difference in the amplitude of a sound and depends on both the frequency and the amplitude of a sound. At usual levels of frequency and amplitude, the JND in a sine tone is between 0.2 and 0.4 dB. The ear is most sensitive to frequencies in the range 1000-5000 Hz. In general, a tone at a given amplitude with a frequency between 1000 and 5000 Hz is perceived as louder than one of higher or lower frequency. Loudness also depends on bandwidth. Sounds with a large bandwidth are louder than those with a narrow bandwidth, even if they have the same energy. Further, loudness depends on duration. For sounds shorter than a second, loudness increases with duration. For sounds longer than a second, loudness remains constant.

Measures of *perceived* loudness have been developed to separate what is perceived as amplitude from the actual amplitude of a sound. How intense must a 1000 Hz tone be in order to sound equally loud to some reference level? A *phon* is defined as a unit of loudness level, 1 dB at 1000 Hz, and a *sone* is a measure of an individual's perception of loudness at 40 dB above the audible threshold at 1000 Hz. A subject is asked to adjust the level of a 1000 Hz tone until it sounds half as loud as it did, this is assigned a value of 0.5 sones. Similarly, a level twice as loud is assigned 2 sones, and so on. It has been found using this method that a 10 dB step in sound level results in an approximate doubling of perceived loudness.

The normal ear's response to sounds of varying amplitude depends on the frequency of the sound. Below is a graph of the normal ear's response to sounds of varying amplitude and frequency .

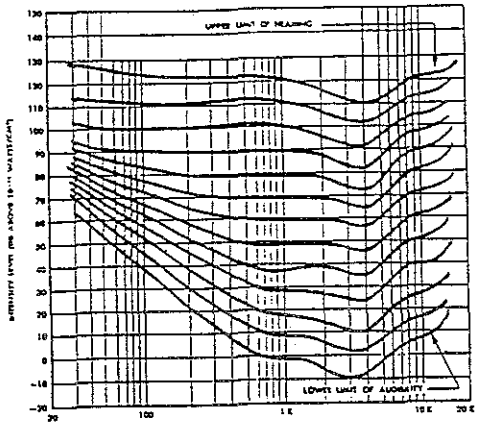


Figure 11: The equal loudness contours for sine waves, after Fletcher and Munsen (1933

Sensitivity -- measured in dB (above 20×10^{-6} PA)

140 -- pain or discomfort

40 - 100 -- area of audible tones

0 - 20 -- threshold zone

4.2 Pitch

Pitch is defined as, "a quality or dimension of tonal *perception* that corresponds most closely to the physical dimension of frequency within the range of 20 to 20,000 Hz."

There are two prevalent theories of how pitch is detected. One theory says that the information is encoded across *different* auditory neurons (the place theory). The other says that the *temporal patterns* of firing within and across neurons determines pitch perception (the time theory).

The amount of frequency change necessary for detecting a change in pitch varies systematically as a function of: duration, intensity, frequency, and acoustic background. At a frequency of 1000 Hz and at a moderate intensity, a change of about 3 Hz can be detected, a Just Noticeable Difference. It takes roughly 25 msec at 1000 Hz to detect a noticeable change in pitch.

At frequencies below 1000 Hz, the pitch goes down as the sound gets louder and above 1000 Hz, the pitch goes up as the sound gets louder. Bright sounds, sound with a relatively high amount of high frequency energy, sound higher than dull ones.

There is a nonlinear relationship between pitch perception and frequency. In addition to pitch height, there seems to be a circular component to pitch perception. Sounds an octave apart are perceived as very "similar". So, it seems, pitch perception could be modelled, not as a vertical continuum from low to high pitch, but as a spiral from low to high circling with every octave.

Pitch is nearly logarithmic when compared to frequency. The stimulation on the basilar membrane occurs at points almost exactly proportional to the logarithm of the frequency. At higher frequencies, when the frequency doubles, the distance between points of stimulation on the basilar membrane changes by an approximately constant distance (roughly 3.4 mm).

4.3 Duration

Perception of sound apparently requires the buildup of a certain amount of energy before it is noticed. The perceived onset of a sound depends on the slope of its attack .

As with vision, there is a rate at which it is no longer possible to perceive discrete sounds as being discrete, the perception of interrupted sound becomes the perception of continuous sound. This phenomenon is known as *Auditory Flutter Fusion* and the rate at which it is no longer possible to distinguish between separate sounds is the *Flutter Fusion Threshold*. The flutter fusion threshold varies widely between subjects and between studies but the range seems to be from 45 to 120 bursts per second. Part of the reason for the disparity between subjects seems to be the ability to detect some sort of interruption at much higher rates than it is possible to identify the interruption, the subject knows that something has happened but doesn't know what.

4.4 Masking

The perceived loudness of a sound can vary depending on the context in which it is perceived. Sounds may *mask* each other, as when conversation becomes difficult in a noisy car. In general, a loud sound will mask a soft sound but masking also depends on frequency. Sounds tend to mask higher frequencies more than lower ones. In addition, sounds can mask other sounds that come after them and, strangely, sounds can even mask other sounds that went before them. Masking, then, has been shown to be dependent on loudness, frequency, and time.

Within a band of frequencies, there are regions (roughly a third of an octave) known as *critical bands* within which, sound energies interact. This phenomenon seems to be explained by the way the basilar membrane responds to vibrations. A pure tone can be masked by surprisingly quiet bandpass noise[1] if the noise frequency is centered around that of the pure tone and within the critical band.

The perception of sound is not as simple as the physical quantities of amplitude, frequency, and duration might suggest. The human perceptual system does not perfectly record the physical dimensions

[1] noise passed through a bandpass filter. A bandpass filter allows a band of frequencies to pass.

of a sound. Each dimension affects the others in ways that are not yet fully understood. This section has given some indication of the complexity of those interactions and points up the need to understand not just the physics of what is being sent but the psychoacoustics of what is being perceived in designing an effective user interface.

5.0 Human Factors Studies

This section will review the human factors literature focusing on three issues that relate to human responses to, and cognition of, sound. First, the section on sound localization will summarize studies suggesting that users are much better able to locate objects using hearing and vision than by using vision alone. Second, the section on gaining attention will review studies arguing whether or not sound is better at gaining a user's attention in a visually rich environment. Third, the multi-dimensional information section briefly summarizes some work by Sara Bly hypothesizing that data that can be difficult to represent visually may be better represented using sound than vision.

5.1 Localization

Under optimal conditions, human visual spatial acuity (the ability to identify where objects are in 3 space) is between 30 and 600 times better than the ability to locate objects based on sound. Using sight alone to locate objects has its disadvantages though. It is only possible to see in front, and focusing on one object makes it difficult to switch to find another. It is necessary to search the entire perceptual sphere and refocus the eyes from near to far to find an object. Worse yet is the ability to adjust to varying degrees of light and dark. Going from a bright room to a dark room takes from several seconds to minutes to adjust to the light intensity. Even under ideal conditions with the object in sight, anyone who has lost car keys can attest to the difficulty in picking out one object of interest from a rich background of colorful images. Commercial products such as key finders, that emit beeps to help people find their keys, point up the potential value of using sound to make localization easier.

Three recent papers in the Human Factors literature explore the capability of humans to use sound cues to locate stationary and moving objects. Perrott [Perrott] studied subjects' ability to locate targets with and without sound cues, Strybel [Strybel] described subjects' ability to determine the velocity of objects based on sound, and Wenzel and others [Wenzel2] attempted to determine what factors effect a subject's ability to locate sounds.

By combining acoustic spatial information with visual information, Perrott found that a subject is able to locate targets 500 to 1000 ms faster in a well lit test field than he could when using visual information alone. The low numbers resulted when the object was directly in front of the subject and the more the azimuth varied, the longer the search time and the greater the disparity between visual-only and visual-with-sound localization.

In another paper from the same symposium, Strybel reviewed current research in auditory motion perception and described studies that found that, while a subject's ability to determine the position of sounds was much worse than visual resolution[1], it appeared that his ability to determine the *velocity* of an object by sound is equal to the visual system. Strybel went on to describe findings that suggested that the ability to locate objects by sound is greatest when the sounds come from straight ahead (and interaural differences can be used to triangulate most effectively). It also appears that changes in frequency of sounds produced by the pinnae (the visible part of the ear) contribute to the detection of motion.

Wenzel [Wenzel2], Wightman and Foster calculated Head Related Transfer Functions (HRTF's) that included frequency dependent phase and intensity information by placing microphones near the subjects' eardrums to determine what factors effect the ability to locate sounds. They found the pinnae of the ear to be especially important in explaining differences in subjects' ability to locate sounds. People may have good or bad ear shapes for locating sounds. Moreover, simulating a good ear-shape in someone with a bad ear-shape didn't help that person locate sounds, suggesting that localization may depend on more than just the pinnae (possibly learning). In general, they found that elevation is especially difficult to judge and the differences caused by differently shaped pinnae are especially pronounced when locating sound differences in elevation.[2]

These papers all report that subjects are able to use hearing alone to judge an object's location and velocity. Compared to visual localization, sound localization is less accurate but has the advantage of being omnidirectional. This leads one to suspect that the process of finding things in three space can be greatly accelerated using sound cues in addition to visual cues.

5.2 Gaining Attention

A great deal of work has been done with jet fighter pilots comparing sound input with visual input. While this may seem to be an extreme environment and lacking in general applicability, the findings are interesting and may be useful to someone designing virtual reality applications. The problem jet

---

[1] on the order of 1-5° vs. seconds of arc for visual resolution

[2] One product of this research is commercially available "Convolvotron", digital signal processing hardware that takes an arbitrary sound source and gives the user the ability to move the apparent source of the sound around in 3 space in real time. This board sells for $25,000 and runs on an IBM PC. For more information on the Convolvotron, contact Scott Foster at Crystal River Engineering, Inc. (12350 Wards Ferry Road, Groveland, CA 95321) or call (209) 962-4118. One competitor of the Convolvotron is Gehring Research Corporation's "Focal Point 3-D Audio System". This system is based on DigiDesign's AudioMaster DSP board which costs $1000. The entire system with four channels of audio input is $9000. For more information about the Focal Point system, contact Bo Gehring at Gehring Research Corporation (189 Madison Ave., Toronto, Ontario M5R 2S6) or call (416) 966-3139.

pilots are faced with is that modern technology has presented them with a myriad of useful instruments and displays, all visual, all competing for the pilot's attention. At the same time, the pilot, in an aerial dogfight, is under great stress and forced to keep his attention focussed on the opposing plane, his own squadron, and ground features (including mountains). Similarly, in a visually busy virtual environment, attracting a user's attention can be challenging. Studies suggest that sound adds another mode of input that may be effective in attracting attention when events occur which require immediate action. Two studies exploring advantages and disadvantages of using sound to gain attention are described here.

In 1983 Wickens, Sandry, and Vidulich [Wickens] described some principles for designing interfaces under two situations. In one scenario, there was a single task to be accomplished and the challenge was to couple the input and output modalities to give the fastest and most accurate response. In the second, there were multiple tasks competing for the operator's limited processing capability.

In the single task scenario, Wickens, et al suggested a Stimulus-Central Processing-Response model (S-C-R) , that expanded on the Stimulus-Response model by a) taking into consideration the belief that operators must incorporate a stimulus signal into a *mental model* of the system before they are able to decide on a response, and b) incorporating developments that suggested that there are two fundamentally different modes of representation of information in memory, *spatial* and *verbal*.

In studying single task input/output compatibility, the paper theorized that *auditory input* coupled with *speech output* would be most effective in *verbal* tasks and *visual input* coupled with *manual output* would be most effective in *spatial* tasks.

Wickens et al, conducted two experiments each with two different tasks. The spatial task was to maintain a cursor on a moving reference display by manipulating a joystick. The verbal task was to respond whether each of a series of letters displayed was a member of a memory set of three characters.

Their first experiment drew four conclusions: 1) Tasks compete for visual input channels, 2) Tasks compete for manual output channels, 3) There is an asymmetry of interference effects, increased perceptual competition will disrupt a task whose demands are cognitive (i.e. remembering), while increased response competition will disrupt a task whose demands are motor (i.e. tracking).

The second experiment tested the effectiveness of each type of stimulus and response as a side task concurrent with a manual control task of flying an F/A-18 simulator through a serpentine three dimensional tunnel. In this experiment, they tried to test the compatibility of input modes with output

modes combined with the effects of resource competition. They drew five more conclusions: 1) Resource competition disrupts performance (i.e. doing more than one thing at a time is hard). 2) Single task S-C-R compatibility is generalized to two tasks. 3) Compatibility and resource competition interact, the gains from using compatible modes (auditory input and speech output for verbal tasks) and the gains from avoiding competing for resources (trying to do two things with the right hand, for example) don't add, the total is less than the some of the parts. 4) Increases in workload will amplify compatibility and resource competition effects. 5) They concluded that their theory was correct, matching auditory input with speech output worked best for verbal tasks and matching visual input with manual output worked best for spatial tasks.

Wickens' study indicated that speech input is effective for verbal tasks, but no mention is made of non-speech auditory input. A reasonable question, then, might be "Since non-speech audio doesn't require the cognitive overhead of translating sounds into words, might non-speech audio be more effective for spatial tasks than speech?", or "How effective might *localized* non-speech audio be for spatial tasks?"

A paper by Robinson and Eberts [Robinson] in 1987, refutes some of the value of auditory input by concluding that operators are better able to remember information presented visually and better able to maintain the context of an emergency. By presenting information pictorially and in a consistent and organized way (that is, a graphical representation of a fire directly above the button to be pushed to extinguish a fire), they found that subjects responded faster to visual input and were able to handle multiple visual inputs more easily than multiple auditory inputs, especially when presentation rates were accelerated.

These examples of human factors research examine the advantages of adding sound input to a visually rich environment to gain the user's attention. Tom Furness, of the University of Washington's Human Interface Technology Laboratory, suggested a compelling example of how sound can be used to attract the attention of a fighter pilot in an emergency, the sound of his young daughter's voice.

5.3 Multi-Dimensional Information

Sara Bly of Xerox PARC has studied the advantages of using sound in representing three types of the data that can be difficult to represent visually: multivariate, time-varying, and logarithmic data. [Bly1], [Bly2], and [Buxton].

As an example, of sound representation of multivariate data, Bly encoded six dimensions used in identifying species of lilies as pitch, volume, duration, the fundamental waveform, the attack

envelope, and the addition of the 5th and 9th harmonics. She found that her subjects were better able to discriminate between two data sets using sounds coded in this way than using graphical representations of the same data.

In two less formal studies, Bly explored the use of sound in representing time-varying data (by encoding functions in two dimensions using pitch and volume and representing different functions with different waveforms) and representing logarithmic data using frequency (which are perceived logarithmically).

These three studies point up advantages sound has over visual input in representing data. The next section will discuss applications that use sound accordingly.

6.0 Applications

This section contains a brief discussion of ways in which sound can be used to enhance the user interface of projects in progress at the University of North Carolina's Department of Computer Science. First, it will discuss three methods of using sound feedback in an interface, followed by a more specific discussion of two sample projects, the Head Mounted Display and ARM projects, where sound could be used in the interface.

6.1 Virtual Reality

"The screen is a window through which one sees a virtual world. The challenge is to make that world look real, act real, sound real, feel real."

[Sutherland]

By enhancing the illusion of reality with sound, the application developer adds one more dimension, making it more difficult to distinguish the virtual reality from true, veridical, reality. This section discusses three ways sound could be used in a virtual environment to provide the user with more information; continuous sound feedback, sound effects, and auditory icons.

6.1.1 Continuous Sound Feedback

Changes in one-dimensional parameters such as temperature, flux, or density can be signalled by altering pitch or loudness and more complex parameters can be represented by altering timbre in ways analogous to the process being represented. This requires generating sounds "on the fly" in response to changes in the virtual environment rather than playing back discrete stored sounds.

6.1.2 Sound Effects

In addition to using sounds as disembodied abstract tones, certain images have associated sounds that occur quite naturally. A computer graphic of a fish swimming can be accompanied by the sound of bubbles, a graphic of a bird, by the sound of a bird and wind whipping past the bird's wings as it flies. Sound effects have the advantage in virtual worlds applications of being consistent with the illusion being promoted. Instead of reminding the user that he is in a cartoon world and forcing him to step back and think about where he really is, sound effects foster the illusion, allowing the user to further immerse himself in subjective reality.

6.1.3 Auditory Icons

In a 1986 paper in Human-Computer Interaction, William Gaver [Gaver] suggested using caricatures of naturally occurring sounds, auditory icons, to provide information based on the way people listen to the world in their everyday lives. He argued that, cognitively, sounds are associated with their *sources*. Studies [VanDerveer] have shown that subjects describe sounds in terms of their dimensions (pitch, loudness, duration) only when they are unable to identify the source events. Typically, it is these dimensions of sound have been used to represent dimensions of data with, perhaps, no analog in the everyday world. Gaver suggests using dimensions of the sound's source instead.

"One can imagine how a single sound could be used to give information about a file. Because it is a large message, it makes a rather weighty sound. The crackle of paper comes from the left and is muffled: The mailbox must be in the window behind the one that is currently on the left side of the screen. And the echoes should like a large empty room, so the load on the system must be fairly low. All this information from one sound!"

Additional information on Auditory Icons can be found in the CHI '90 tutorial notes for "The Use of Non-speech Audio at the Interface" [Buxton]. The next section describes two on going projects at UNC that could make use of sound interaction in the form of Continuous Sound, Sound Effects, and/or Auditory Icons: the Head Mounted Display and ARM projects.

6.2 Head Mounted Display

One on-going project at UNC is the Head Mounted Display project (see [Chung]). The Head Mounted Display is a helmet (or set of goggles). Attached to the helmet is a pair of miniature television screens and a magnetic tracking sensor. The tracker reports position (x, y, z), and orientation (azimuth, elevation, roll). Knowing where the head is and in which direction it is facing allows the software to alter the image presented on the screens. This gives the user the illusion that he is able to walk around a stationary object or that he can walk around in a virtual environment, such as, a computer representation of a building.

Currently, the Head Mounted Display project is using sound in a simple way to demonstrate that sound interactions are possible. When a user selects an object to be manipulated, by placing a 3D cursor within the bounding sphere of the object to be selected and pressing a button on the cursor's control device (a pool ball), a squeak sound emanates from the speaker on the nearby Macintosh. This signals the user that he has successfully selected the object.

One future application for sound in the Head Mounted Display environment might be in planning radiation therapy . When planning radiation therapy, a radiation oncologist must be careful to target the tumor or organ to be treated and avoid irradiating nearby vital organs. Currently, Jim Chung is working on using the Head Mounted Display as a tool in planning where these beams will travel by creating a graphic image of the patient from CAT scan, MRI, or Ultrasound data. In the future, the radiation oncologist may be able to view this data superimposed on the real patient. Sound, in the form of continuous feedback or auditory icons, might be used in such an application to inform the radiation oncologist when the target beams are near vital organs.

6.3 ARM

Another on-going project that could make use of sound feedback is the Argonne III Remote Manipulator (ARM) molecular modeling project. A mechanical arm is used to simulate the forces of molecular bonds. The user is able to manipulate virtual molecules and, through the computer-controlled mechanics of the ARM, feel the "molecule" resist in areas of high energy and give way in areas of low energy. This allows biochemists to experiment with molecules that have not yet been created to understand how these hypothetical molecules might bind with known receptor proteins. For a more detailed description of this system, see [Ouh-Young] and [Kilpatrick].

The current system represents bond energies with an "Energy Thermometer ", a vertical bar displayed on the screen to the right of the molecule representation. In order to gauge the overall energy level of the system, the user must constantly glance over to the right to read the thermometer. Since this is a simple one-dimensional read out, the energy level could effectively be represented by pitch, loudness, or timbre, allowing the user to remain visually focussed on the molecules before him.

This section has suggested three types of interactions, continuous sound feedback, sound effects and auditory icons, that can be used in developing sound interfaces and described examples of two projects where sound could be used. The next section will describe some ways that sounds can be synthesized in a computer application.

7.0 Synthesis of Sound

It is reasonable to ask the question, "How does one use a computer to make these sounds?". There are two categories of sound generation this section will consider, *digitized sounds*, basically sound bites, that have been recorded or generated at some previous time, and *real-time sound* which can be generated "on the fly" given parameters such as frequency, amplitude, harmonics, envelope, etc.

7.1 Digitized (Stored) Sound

Within this department (Computer Science Dept. at UNC in 1990), there are three separate tools available for digitized sound reproduction or creation: an application on the Macintosh called "MacRecorder", CSound, a public domain sound generation application, and XSound, an application under development by a project team from the Software Engineering course.

7.1.1 MacRecorder

MacRecorder is a Macintosh application for sound generation. It presents the user with an image of a waveform and allows him to create and tailor that waveform in a number of ways. Functions include changing frequency, amplitude, waveform (sine wave, square wave, sawtooth), envelope, and duration, adding harmonics and reverberation, looping to repeat a sound and cutting and pasting parts of the waveform. The application gives the user the ability to save sounds that have been created and digitize sounds recorded using a microphone. For more information about MacRecorder, see the documentation[1].

7.1.2 CSound

CSound is a software application in the public domain developed at MIT's Media Lab . It allows the user to create two files, an orchestra file, containing parameters for defining the sound parameters of various instruments, and a score file that contains commands detailing what to play and when. The output of CSound is a sound file and local software and hardware must be provided to actually play the sound file. For more information about CSound, talk with Jim Symon and see the documentation he has in ~symon/sound/csound .

---

[1] "MacRecorder User's Guide" Copyright 1987, Farallon Computing, Inc.. 2150 Kittredge St., Berkeley, CA 94704 phone (415) 849-2331.

7.1.3 XSound

There is a project underway in Comp 145 to create a software package that runs on a Sun-4 under UNIX to provide functions similar to those provided by MacRecorder. The actual functionality of the application is still being determined at the time of this writing. More information about this product is available by talking to a member of the team, Chris Baroudi, Jeff Kiel, Jeff Lewis, and Mark Kupper.

7.2 Real-time Sound Synthesis

Generating real-time sound is difficult. This section discusses why it is difficult and suggests solutions to some of the problems that will be applicable in this department at the time of this writing.

7.2.1 Issues

A few reasons why sound creation is difficult both in terms of speed and memory requirements will be discussed. It takes fast computation to be able to generate real-time sound. As discussed earlier, in order to prevent aliasing, frequencies are limited to roughly 40% of the sample rate. In order to create a tone at 16 kHz, it is necessary to sample at 40 kHz, faster than the clock cycle of most of today's computers. Sampling at 40 kHz puts great demands on computer hardware to be able to process the needed amount of information.

The amount of storage required by digital sound samples can also be prohibitive. The resolution of Digital to Analog (D/A) and Analog to Digital (A/D) converters is measured in bits. The dynamic range [1]that can be achieved is 6 dB/bit. With 16 bit data converters, this gives a dynamic range of 96 dB, which is roughly the dynamic range of a typical concert hall (this is the number of bits used for digital recordings on Compact Disks). Unfortunately, the Macintosh only uses 8 bits and the XSound application to run on a Sun-4 will only use 12 bits. This results in a noticeable decrease in sound quality. At 16 bits/sample and sampled at 40,000 samples/sec, a three minute song will take 13.7 M bytes of storage. So there is a clear and difficult tradeoff between sound quality, memory requirements and clock speed.

[1] the ratio of strongest to weakest signal

7.2.2 Approaches

Three methods for creating sound on a computer will be discussed: Additive (Fourier) Synthesis , a simple table lookup, and MIDI, in reverse order of practicality.

7.2.2.1 Additive Synthesis

Given a specific periodic waveform, how does one find a mathematical representation? The unique series for each periodic waveform is called the trigonometric Fourier series of the signal. Using Fourier Analysis techniques, arbitrary waveforms can be generated from a combination of sinusoidal waveforms given an amplitude and frequency.

Generally, a sinusoidal sequence has the form

$$x(t) = a_0 + a_1 \cos \omega_0 t + a_2 \cos 2\omega_0 t + \ldots + b_1 \sin \omega_0 t + b_2 \cos 2\omega_0 t + \ldots \quad (1)$$

where $-\infty < t < \infty$
ω_0 is the frequency in radians/sample

the a's and b's are the coefficients to be found. For more detail on how these coefficients can be solved for, see Appendix II.

Due to the performance requirements described above, this technique is often implemented in hardware. Function generators create amplitude and frequency values (possibly with a predetermined envelope) that are input to sinusoidal oscillators (one for each harmonic). The output of the oscillators are then added to form the final output sound (hence the name, Additive Synthesis). It is possible to implement this technique in software, using the Fast Fourier Transform algorithm and a very fast processor, but, for most applications, there are better ways.

One alternative synthesis technique, subtractive synthesis, does the opposite. It takes a complex source with a broad spectrum, such as white noise , and subtracts out selected portions to create the desired tone.

7.2.2.3 Table Lookup

One alternate approach is to precompute values for 1/4 of a canonical waveform and store the values in a look up table. To create an arbitrary waveform then, it is only necessary to vary the increment used for running through the table[1], multiply the resulting data values by the proper factors to account for amplitude changes and negate to form the bottom half of the wave.

This technique can be implemented in software or hardware. In fact, this forms the basis for a VLSI chip designed and fabricated by Jim Symon in this department.

7.2.2.3 MIDI

The most popular and most practical approach to creating sound is to make use of the MIDI protocol. MIDI stands for "Musical Instrument Digital Interface". It is the accepted standard for communication between electronic musical instruments and consists of electrical specifications (for cables, plugs and voltage levels) and a serial interface protocol that allows communicating processors to send control messages rather than data samples. Thus it is only necessary to send data such as amplitude and frequency to separate hardware that is specially designed to generate sound, usually a synthesizer.

A MIDI message consists of two or more bytes that signal a musical event, such a *note on* or *note off*. The first byte is a status byte identifying the function of the message. The bytes following are data bytes. Messages are of one of two types: *system messages* which are recognized by all devices, and *channel messages* recognized by devices with matching channel numbers.

There are three types of system messages: *system common messages*, *system real time messages*, and *system exclusive messages*. System common messages coordinate song selection and tuning among devices. System real time messages synchronize timing among different sequencing devices. System exclusive messages are manufacturer-specific messages used to send data between specific devices that can't be sent in other ways.

There are two types of channel messages: *channel voice messages* and *channel mode messages*. Channel voice messages send actual performance data between MIDI devices, pitch, amplitude, timbre,

[1] The second and fourth quarters of a sine wave , for example, can be generated by reading the table backwards, i.e. with an increment of -1. Higher frequencies can be generated by increasing the increment value, thereby reading through the table faster

duration, and other sound qualities. Channel mode messages set various modes in the receiving devices to do such things as stop spurious notes from playing and affect local control of the device.

By connecting a computer to a sampling synthesizer such as the Ensoniq Mirage, arbitrary sound samples can be recorded and stored on the sampling synthesizer and only MIDI messages need to pass from computer to synthesizer. This control information can be used to vary the pitch, duration, volume, and timbre of the original sound sample recorded, creating a variety of useful sounds without incurring the expense of storing the sounds in the computer or synthesizing the sounds in real time.

To connect most computers to a MIDI system, it is necessary to have a MIDI interface, usually a small box with a RS-422 interface plug for the computer and MIDI ports for input and output to system components such as synthesizers, sequencers, and amplifiers. The Austin MIDIface adaptor is an example of such a box. It is possible to buy "smart" interfaces that contain processing capability, but "dumb" interfaces, as described above, are usually recommended because of their simplicity, compatibility, and lower cost.

This section has discussed some of the issues that can make sound generation difficult and some ways of generating sounds. By far the most practical solution to creating sounds is to use MIDI, allowing the computer to simply send control messages to hardware that is better suited to store and synthesize the desired sounds.

8.0 Conclusions

This report has attempted to summarize information related to sound from several different perspectives and to point out potential computer applications of sound. Sound has advantages over visual input in that it is omnidirectional, is valuable as an addition to visual input as a means of getting the user's attention quickly, conveying information that is difficult to represent graphically, and enhancing the illusion of reality in a virtual world. The fundamental concepts of value when generating sound have been described. Current and possible applications that can make good use of sound have been suggested. Finally, the practical aspects of getting started generating sound here at UNC. This report is intended as an introduction to the use of sound. For more information, the book by Dodge and Jerse, *Computer Music, Synthesis, Composition and Performance,* listed in the bibliography is a good place to start. It should be clear by now that sounds are more than just bells and whistles .

9.0 Bibliography

[Bly1] Bly, Sara, "Sound and Computer Information Presentation", Doctoral Dissertation (UCRL-53282), Lawrence Livermore National Laboratory and University of California, Davis, CA, 1982.

[Bly2] Bly, Sara, "Presenting Information in Sound", Proceedings of the CHI '82 Conference on Human Factors in Computer Systems, 371-375. New York: ACM, 1982.

[Brooks] Brooks, Frederick P., "Grasping Reality Through Illusion - Interactive Graphics Serving Science", Proceedings of CHI '88, ACM Conf. Hum. Fac. Comp. Sys., Washington, D.C., 1-11.

[Buxton] Buxton, Bill, Gaver, Bill, and Bly, Sara, "The Use of Non-Speech Audio at the Interface", CHI Tutorial Notes, SIGCHI '90, Seattle, WA.

[Chung] Chung, James, Harris, Mark R., Brooks, F.P., Fuchs, Henry, Kelley, Michael T., Hughes, John, Ouh-Young, Ming, Cheung, Clement, Holloway, Richard L., Pique, Michael, "Exploring Virtual Worlds with Head-Mounted Displays", SPIE Proceedings, Vol. 1083, Los Angeles, CA, Jan. 15-20, 1989.

[Dodge] Dodge, Charles and Jerse, Thomas A., *Computer Music, Synthesis, Composition, and Performance*, Schirmer Books, 1985.

[Gaver] Gaver, William W., "Auditory Icons: Using Sound in Computer Interfaces", Human-Computer Interaction, Vol. 2, 167-177, 1986.

[Gerber] Gerber, Sanford, *Introductory Hearing Science, Physical and Psychological Concepts*, W.B. Saunders Co., 1974.

[Kilpatrick] Kilpatrick, Paul Jerome, "The Use of a Kinesthetic Supplement in an Interactive Graphics System", Doctoral Dissertation, Univ. of North Carolina, Dept. of Computer Science, Chapel Hill, 1976.

[Moore] Moore, Brian C.J., *An Introduction to the Psychology of Hearing*, Academic Press, 1982.

[Oppenheim] Oppenheim, Alan V. and Schafer, Ronald W., *Discrete-Time Signal Processing*, Prentice-Hall, Inc., 1989.

[Ouh-Young] Ouh-Young, Ming, "Force Display in Molecular Docking",Doctoral Dissertation, Univ. of North Carolina, Dept. of Computer Science, Chapel Hill, 1990.

[Perrott] Perrott, David R., "Auditory Psychomotor Coordination", Proceedings of the Human Factors Society-32nd Annual Meeting-1988.

[Robinson] Robinson, Christopher P., and Eberts, Ray E., "Comparison of Speech and Pictorial Displays in a Cockpit Environment", Human Factors, 1987, 29(1), 31-44.

[Sorkin] Sorkin, Robert D., Wightman, Frederic L., Kistler, Doris S. and Elvers, Greg C., "An Exploratory Study of the Use of Movement-Correlated Cues in an Auditory Head-Up Display", Human Factors, 31(2), 161-166, April 1989.

[Strybel] Strybel, Thomas Z., "Perception of Real and Simulated Motion in the Auditory Modality", Proceedings of the Human Factors Society-32nd Annual Meeting-

[Sutherland] Sutherland, I.E., "The Ultimate Display", Proceedings of IFIP 65, vol. 2, 506-508, 582-583.

[Symon] Symon, Jim, Private conversations and Departmental Colloquim presentation, 1989-90.

[Vanderveer] Vanderveer, N.J., "Ecological Acoustics: Human perception of environmental sounds." Dissertation Abstracts International. 40/09B, 4543 (University Microfilms No. 8004002), 1979.

[Wenzel1] Wenzel, Elizabeth M., Wightman, Fredric L., Kistler, Doris J., and Foster, Scott H.,"The Convolvotron: Realtime Synthesis of Out-of-Head Localization", Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan, Honolulu, HI, November 14-18, 1988.

[Wenzel2] Wenzel, Elizabeth M., Wightman, Frederic L., and Foster, Scott H., "A Virtual Display System for Conveying Three-Dimensional Acoustic Information",

[Wickens] Wickens, Christopher D., Sandry, Diane L., and Vidulich, Michael. "Compatibility and Resource Competition between Modalities of Input, Central Processing, and Output", Human Factors, 25(2), 227-248, April, 1983.

[Ziemer] Ziemer, Rodger E., Tranter, William H., and Fannin, D. Ronald, *Signals and Systems, Continuous and Discrete*, Macmillan Publishing Co., 1983.

10.0 Appendix I

10.1 Synthesizing Reverberation

One technique for implementing a delay in sound production digitally is a circular queue. A pointer is kept to the oldest data value in the queue and, when a new sound is input, the oldest data value is output and the new value replaces it in the queue. The pointer is then moved to the new oldest data value. When the pointer gets to the end of the data block, it is wrapped back around to the beginning (thus the queue is circular). The data doesn't have to be moved once it's inserted in the queue. In general, a queue of m cells will result in a delay of m/f_S, where f_S is the sampling rate. A problem with this method is the large queue size required before the delay is noticeable at audio sampling rates (~ 40 kHz).

Sometimes it is desirable to have several different delays for the same signal. This can be done with multiple pointers for output. The queue size is set to be that of the longest delay required and the additional pointers can "tap" the queue to obtain the shorter delays. Taken to the extreme, a *convolution* [1]of the signal is implemented by tapping the queue at every data point with each data value multiplied by the impulse response[2] of a given room for that time delay. All these products are then added together to form the complete output signal. The problem with this method is that it is highly compute intensive. For a delay of 1.5 seconds and a sampling rate of 32 kHz, 48,000 multiplications have to be performed on every sample and storage is required to accommodate 48,000 samples.

A superior technique for reverberation has been described by M.R. Schroeder (see [Dodge], page 229). Briefly, Schroeder's technique uses two types of unit reverberators, one a comb filter and the other an all-pass network.

---

[1] in the continuous case, convolution amounts to taking an integral of two functions multiplied together, one function x(t) is the system input, the other function h(t) is the system characterizing function, the impulse response

$$y(t) = \int_{-\infty}^{\infty} x(\lambda)\, h(t-\lambda)\, d\lambda, \qquad -\infty < t < \infty$$

[2] the response of a system to a unit impulse applied with a wavelength of zero.

10.1 Comb Filters

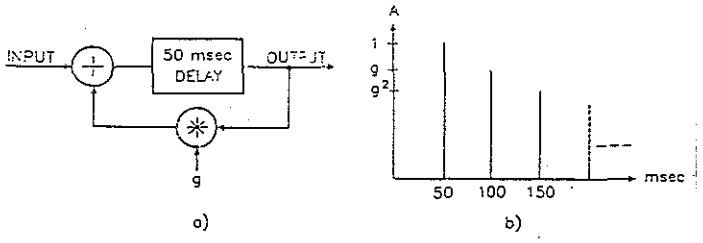


FIGURE 7.6 (a) The internal configuration of a comb filter and (b) its impulse response.

In a comb filter, a signal enters a delay pipeline and cycles through (taking, say, 50 ms) until it reaches the other end where it is multiplied by some amplitude factor (<1) and fed back in the front end again. The time it takes to cycle through the pipeline is the *loop time*. The frequency of such a comb filter is 1/(loop time), the natural frequency of the filter. It is called a comb filter because of the graph of amplitude over time, which shows impulses every cycle (taking loop time), evenly spaced. This graph resembles the teeth of a comb.

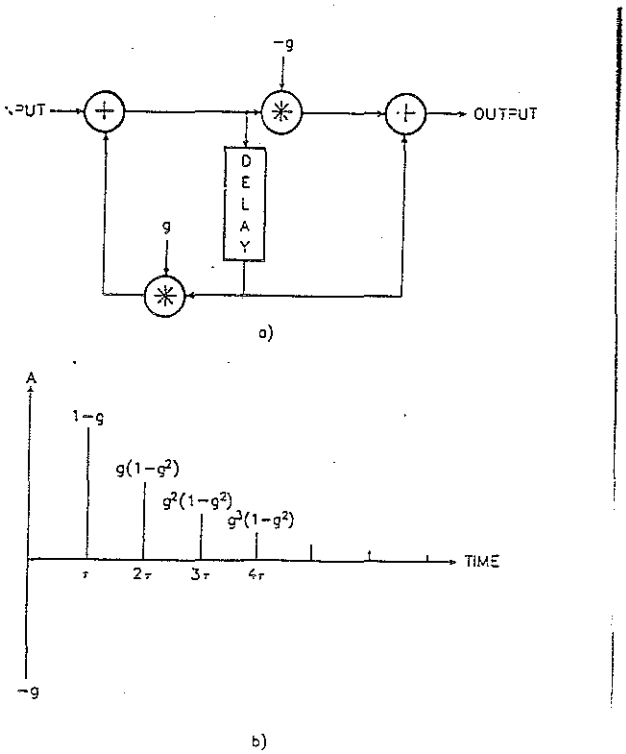


FIGURE 7.9 (a) The internal configuration of an all-pass network and (b) its impulse response.

10.2 All Pass Networks

The all-pass network is similar to a comb filter, signals pass through a delay queue , are multiplied by an amplitude factor and recycled through again. One difference is that the all-pass network outputs the first signal immediately rather than forcing it through the delay queue once. What results is an output signal resembling the input signal except that the network will "ring" with a period equal to the loop time of the delay queue.

11.0 Appendix II

11.1 Calculating Fourier Coefficients

As noted in the section on Additive Synthesis, a sinusoidal sequence has the form

$$x(t) = a_0 + a_1 \cos \omega_0 t + a_2 \cos 2\omega_0 t + \ldots + b_1 \sin \omega_0 t + b_2 \cos 2\omega_0 t + \ldots \quad (1)$$

where $-\infty < t < \infty$
ω_0 is the frequency in radians/sample

the a's and b's are the coefficients to be found.

This may be rewritten as ...

$$x(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega_0 t + \sum_{n=1}^{\infty} b_n \sin n\omega_0 t \quad \text{where } -\infty < t < \infty \quad (2)$$

Integrating, starting with a_0 gives...

$$\int_{T_0} x(t)\,dt = a_0 \int_{T_0} dt + a_1 \int_{T_0} \cos \omega_0 t\,dt + a_2 \int_{T_0} \cos 2\omega_0 t\,dt + \ldots + b_1 \int_{T_0} \sin \omega_0 t\,dt + b_2 \int_{T_0} \sin 2\omega_0 t\,dt + \ldots \quad (3)$$

where $T_0 = 2\pi/\omega_0$ is a period of the fundamental

Since all terms except the first integrate a sine or cosine over an integral number of periods, they are equal to zero (the negative half of the curve cancels the positive half) and the equation simplifies to...

$$a_0 = 1/T_0 \int_{T_0} x(t)\,dt \quad (4)$$

the average value of the waveform.
Going back to equation (2), if both sides are multiplied by cos $m\omega_0 t$ (using m permits choosing any a_n coefficient), and integrate over the period, one gets...

$$\int_{T_0} x(t) \cos m\omega_0 t \, dt = a_0 \int_{T_0} \cos m\omega_0 t \, dt + \int_{T_0} \left(\sum_{n=1}^{\infty} a_n \cos n\omega_0 t\right) \cos m\omega_0 t \, dt + \int_{T_0} \left(\sum_{n=1}^{\infty} b_n \sin n\omega_0 t\right) \cos m\omega_0 t \, dt \quad (5)$$

Again, the first term integrates to zero and by multiplying through gives...

$$\int_{T_0} x(t) \cos m\omega_0 t \, dt = \sum_{n=1}^{\infty} a_n \int_{T_0} \cos n\omega_0 t \cos m\omega_0 t \, dt + \sum_{n=1}^{\infty} b_n \int_{T_0} \sin n\omega_0 t \cos m\omega_0 t \, dt \quad (6)$$

Using the fact that...

$$\int_0^{T_0} \sin m\omega_0 t \cos n\omega_0 t \, dt = 0 \quad \text{for all } m, n \quad (7)$$

All terms of the second series of (6) are zero.

Using the fact that...

$$\int_0^{T_0} \cos m\omega_0 t \cos n\omega_0 t \, dt = \begin{cases} 0, & m \neq n \\ T_0/2, & m = n \neq 0 \end{cases} \quad (8)$$

all terms of the first series of (6) are zero except where n = m. When n = m, the integral gives $T_0/2$, which results in...

$$\int_{T_0} x(t) \cos m\omega_0 t \, dt = a_m (T_0/2) \quad (9)$$

Multiplying both sides by $2/T_0$ gives...

$$a_m = 2/T_0 \int_{T_0} x(t) \cos m\omega_0 t \, dt, \quad m \neq 0 \quad (10)$$

similarly, using the fact that...

$$\int_0^{T_0} \sin m\omega_0 t \sin n\omega_0 t \, dt = \begin{cases} 0, & m \neq n \\ T_0/2, & m = n \neq 0 \end{cases} \quad (11)$$

it is possible to show...

$$b_m = 2/T_0 \int_{T_0} x(t) \sin m\omega_0 t \; dt, \quad m \neq 0 \qquad (12)$$

By making simplifying assumptions about the waveforms, closed form solutions for these integrals can be derived.

12.0 Index