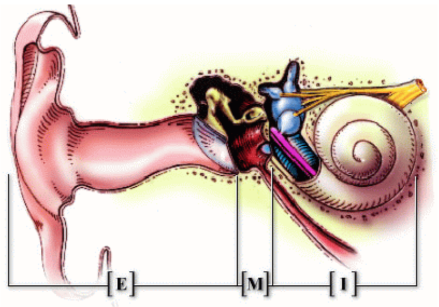
5 Sound and Audio

Multimedia - Department of Computing, Imperial College
Professor GZ Yang ! **http://www.doc.ic.ac.uk/~gzy**

Sound and Human Hearing

Thus far, the media we have been studying are primarily visual, being perceived through our eyes. Sound is perceived through the sense of hearing, and our ears detect vibrations in the air in a completely different manner than that of our eyes detect light, and our brains respond differently to the resulting nerve impulses. Although sound is a familiar phenomenon of our daily life, it is a complex mixture of physical and psychological factors. In a nutshell, sound waves are longitudinal mechanical waves generated by a vibrating surface. To understand how we perceive sound, however, we must first explore how human hearing works.



The figure shown above illustrates the major structures and processes that comprise the human ear. The outer ear is composed of two parts, the visible flap of skin and cartilage attached to the side of the head, and the ear canal, a tube about 0.5 cm in diameter extending about 3 cm into the head. These structures direct environmental sounds to the sensitive middle and inner ear organs located safely inside of the skull bones. Stretched across the end of the ear canal is a thin sheet of tissue called the tympanic membrane or ear drum. Sound waves striking the tympanic membrane cause it to vibrate. The middle ear is a set of small bones that transfer this vibration to the cochlea (inner ear) where it is converted to neural impulses. The cochlea is a liquid filled tube roughly 2 mm in diameter and 3 cm in length. The cochlea is curled up and looks like a small snail shell. In fact, cochlea is derived from the Greek word for snail.

When a sound wave tries to pass from air into liquid, only a small fraction of the sound is transmitted through the interface, while the remainder of the energy is reflected. This is because air has a low mechanical impedance (low acoustic pressure and high particle velocity resulting from low density and high compressibility), while liquid has a high mechanical impedance. In less technical terms, it requires more effort to wave your hand in water than it does to wave it in air. This difference in mechanical impedance results in most of the sound being reflected at an air/liquid interface. The middle ear is an impedance matching network that increases the fraction of sound energy entering the liquid of the inner ear. For example, fish do not have an ear drum or middle ear, because they have no need to hear in air. Most of the impedance conversion results from the difference in area between the ear drum (receiving sound from the air) and the oval window (transmitting sound into the liquid). The ear drum has an area of about 60 $(mm)^2$, while the oval window has an area of roughly 4 $(mm)^2$. Since pressure is equal to force divided by area, this difference in area increases the sound wave pressure by about 15 times. Contained within the cochlea is the basilar membrane, the supporting structure for about 12,000 sensory cells forming the cochlear nerve.

The basilar membrane is stiffest near the oval window, and becomes more flexible toward the opposite end, allowing it to act as a frequency spectrum analyser. When exposed to a high frequency signal, the basilar membrane resonates where it is stiff, resulting in the excitation of nerve cells close to the oval window. Likewise, low frequency sounds excite nerve cells at the far end of the basilar membrane. This makes specific fibres in the cochlear nerve respond to specific frequencies. This organization is called the place principle, and is preserved throughout the auditory pathway into the brain. It must be noted that human hearing is very complex and there are many subtle effects and poorly understood phenomena related to human hearing.

Measurement of Sound

The intensity of sound falls away from a sound source by following the inverse square law**.** As the intensity of the loudest sound we can tolerate is 10^{12} larger than the weakest detectable sound, we express sound in terms of how many powers of ten one intensity is greater than another. That is, if one intensity is 10^6 times another, we say it is 6 Bells louder (60 tenths of a Bell, or 60dB). If one intensity is 1/1000 of another, it is 30 dB less. Mathematically, it is represented as

$$dB = 10 \log \frac{Intensity1}{Intensity2}$$

From the above equation, it is not difficult to see that by doubling the ratio of signal intensities, we add 3 dB to the sound, and halving the ratio of signal intensities we subtracts 3 dB from the sound.

Physically, sound is measured as Sound Pressure Level (SPL), and is expressed as

$$Intensity = \frac{PressureAmplitude^2}{2rc}(wm^{-2})$$

where ñ is the density of the medium through which sound is travelling, and c is the speed of sound. The weakest sound we can hear has a variation in air pressure of $0.6x10^{-12}$ wm^{-2}**.** Normally, we use dB SPL to represent sound intensity, i.e., it represents sound in dB relative to the minimum detectable sound in wm^{-2}**.** The loudest tolerable sound is about 1.4 wm^{-2}, therefore, represented in dB SPL, it is about 124 dB. Throughout our hearing range, normal conversation measures about 60 dB, whereas Jack hammers and rock concerts generate sound at about 110 dB SPL.

Perception of Sound

It is important to note that the intensity of sound does not equal to the loudness how we human perceive it. In fact, our perceived loudness is proportional to Intensity$^{1/3}$. Therefore, if the physical intensity of sound is increased by 10 times, listeners will experience the loudness being increased by a factor about 2 (=10 $^{1/3}$). The same principle follows if we want to reduce the perceived loudness of sound, which explains why it is so difficult to produce a room that is completely sound proof.

The perception of a continuous sound, such as a note from a musical instrument, is often divided into three parts: loudness, pitch, and timbre (pronounced "timber"). Loudness is a measure of sound wave intensity, as previously described. Pitch is the frequency of the fundamental component in the sound, that is, the frequency with which the waveform repeats itself. While there are subtle effects in both these perceptions, they are a straightforward match with easily characterized physical quantities. Timbre is more complicated, being determined by the harmonic content of the signal. This means that when you have a sound which is made of two sine waves of different frequencies, the perceived sound will be independent of their relative phases. In the time domain, the phase difference can result in completely different waveforms. Despite this, these signals will sound identical to us as our hearing is based on the amplitude of the frequencies, and is very insensitive to their phases. The shape of the time domain waveform is only indirectly related to hearing, and usually not considered in audio systems.

The ear's insensitivity to phase can be understood by examining how sound propagates through the environment. Suppose you are listening to a person speaking across a small room. Much of the sound reaching your ears is reflected from the walls, ceiling and floor. Since sound propagation depends on frequency (such as: attenuation, reflection, and resonance), different frequencies will reach your ear through different paths. This means that the relative phase of each frequency will change as you move about the room. Since the ear disregards these phase variations, you perceive the voice as unchanging as you move position. From a physics standpoint, the phase of an audio signal becomes randomized as it propagates through a complex environment. Put it in another way, the ear is insensitive to phase because it contains little useful information.

The aforementioned human hearing characteristics permits the use of signal compression to audio data. The complex and unpredictable nature of sound waveforms makes them difficult to compress using lossless methods. The compression of sound obeys quite different rules than those we have used for image compression, while rapid changes of colour in an image can, in most cases, be safely discarded, the high frequencies associated with rapid changes of sound are highly significant. Due to the non-linear nature of how we perceive sound, telephone companies have used a technique called “companding” to reduce the bandwidth required for transmitting digital audio over the telephone lines. The idea is to use non-linear quantisation levels, with higher levels spaced further apart than the low ones, so that quiet sounds are represented in greater detail than louder ones. Two nearly identical standards are used for companding curves: the µ law used in North America and the A law, used in Europe. Both use a logarithmic non-linearity, since this is what converts the spacing detectable by the human ear into a linear spacing. In equation form, the curve used by the µ-aw is given by

$$y = \frac{\log(1 + mx)}{\log(1 + x)} \quad for \quad x \geq 0$$

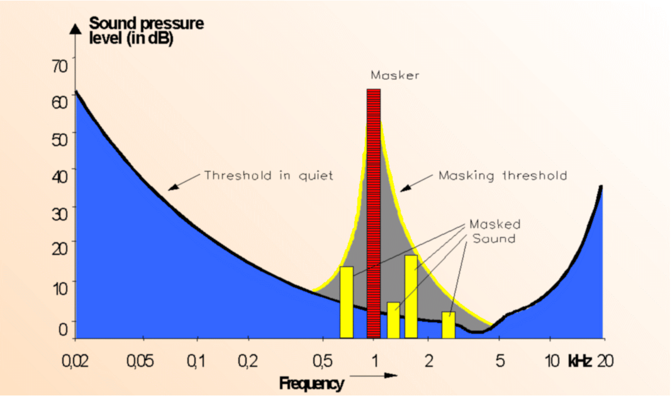
where µ is a parameter that determines the amount of companding. For telephony, µ=255. For the "A" law, the equation becomes:

$$y = \begin{cases} \dfrac{Ax}{1+\log A}, & 0 \le |x| \le 1/A \\ \dfrac{1+\log Ax}{1+\log A}, & 1/A \le |x| \le 1 \end{cases}$$

You may want to try to plot these two mappings in the range of [0,1] to see the difference between the two.

Perceptual Based Audio Coding and MPEG

The secret of effective lossy audio compression is to identify data that is not perceivable by human hearing. When an audio signal is digitized, both audible and inaudible sounds may be included in the digitized version, as the signal records all the physical variations in air pressure that make the sound. There are two perceptual phenomena that can be exploited in audio coding: a sound may be too quiet to be heard or it may be obscured by some other sound. The threshold of hearing is the minimum level at which a sound can be heard. It varies non-linearly with frequency as shown in the following figure.



Very low or very high frequency sound must be much louder than a mid-range tone to be heard. When compressing sound there is little point in retaining sound that fall below the threshold of hearing. Secondly, loud tones can obscure softer tones that occur at the same time. This is not simply a case of loud tone drowning out the softer ones. The effect is more complex and depends on the relative frequencies of the two tones. This is called frequency masking. In the above figure, a loud tone at 1 kHz creates a threshold region within which other tones are masked out. That is, any sound that lies within the masking curve will be inaudible, even though it raises above the unmodified threshold of hearing. In addition to this, another phenomenon called temporal masking can also be exploited for audio compression. This is based on the fact that if we hear a loud sound, then it stops, it takes a little while until we can hear a soft tone nearby. These phenomena are called psycho-acoustics, which are extensively used in MPEG audio encoding.

In MPEG, 32 frequency sub-bands quantized according to a psycho- acoustic model by discarding the following information

- frequencies below hearing threshold
- frequencies masked by other frequencies
- bands with little energy

bits are then allocated between the frequency bands to maximise sound quality per bit, normally with 256 kbit/sec for good quality audios, and use 8ms frames at 48 kHz sample rate. MPEG-1 specifies three layers of audio compression. All three layers are based on the principles outline above and the encoding process increases in complexity from Layer 1 to Layer 3.

- Layer 1: DCT type filter with one frame and equal frequency spread per band. Psychoacoustic model only uses frequency masking.

- Layer 2: Use three frames in filter (before, current, next, a total of 1152 samples). This models a little bit of the temporal masking.

- Layer 3: Better critical band filter is used (non-equal frequencies), psychoacoustic model includes temporal masking effects, takes into account stereo redundancy, and uses Huffman coder.

As a result, the data rate of the compressed audio decreases, from 192 kbps for each channel at Layer 1, to 128 kpbs to Layer 2, and 64 kpbs to Layer 3. The audio part of MPEG-2 standard is essentially identical to that of MPEG-1, except for some extensions to cope with surround sound. MPEG-1 Layer 3 audio, or MP3 as it is usually called, achieves compression ratio of about 10:1, while maintaining high quality. It has become popular in recent years as a means of compressing audio, particularly music for downloading over the Internet, both legally and illegally. Within the MPEG family, MPEG-1 and -2 concentrated almost entirely on compression, MPEG-4, however, moved to a higher level of abstraction in coding objects and using content-specific techniques for coding content. MPEG-4 Audio provides tools for coding of both natural and synthetic audio objects. It permits the representation of natural sounds (such as speech and music) and to synthesize sounds based on structured descriptions. The representation for synthesized sound can be derived from text data or so-called instrument descriptions and by coding parameters to provide effects, such as reverberation and spatialization. The representations provide compression and other functionalities, such as scalability or play-back at different speeds.

MPEG-7 moves to an even higher level of abstraction, a cognitive coding, some might say. In principle, MPEG-1, -2, and -4 are designed to represent the information itself, while MPEG-7 is meant to represent information about the information (although there are areas common between MPEG-4 and -7). Another way of looking at it is that MPEG-1, -2, and -4 made content available. MPEG-7 allows you to describe and thus find the content you need.

Speech Synthesis and Recognition

With the current effort in developing the MPEG-7 standards, research and development in speech synthesis and recognition has taken a new height. Computer generation and recognition of speech are challenging problems and this is an active area of DSP research. Hitherto, most current systems that produce human sounding speech do not synthesize it, but merely play back a digitally recorded segment from a human speaker. This approach has great sound quality, but it is limited to the prerecorded words and phrases. Most human speech sounds can be classified as either voiced or fricative. Voiced sounds occur when air is forced from the lungs, through the vocal cords, and out of the mouth and/or nose. The vocal cords are two thin flaps of tissue stretched across the air flow, just behind the Adam's apple. In response to varying muscle tension, the vocal cords vibrate at frequencies between 50 and 1000 Hz, resulting in periodic puffs of air being injected into the throat. Vowels are an example of voiced sounds. In comparison, fricative sounds originate as random noise, not from vibration of the vocal cords. This occurs when the air flow is nearly blocked by the tongue, lips, and/or teeth, resulting in air turbulence near the constriction. Fricative sounds include: s, f, sh, z, v, and th.

Both these sound sources are modified by the acoustic cavities formed from the tongue, lips, mouth, throat, and nasal passages. Since sound propagation through these structures is a linear process, it can be represented as a linear filter with an appropriately chosen impulse response. In most cases, a recursive filter is used in the model, with the recursion coefficients specifying the filter's characteristics. Because the acoustic cavities have dimensions of several centimetres, the frequency response is primarily a series of resonances in the kilohertz range. Over a short period, say 25 milliseconds, a speech signal can be approximated by specifying three parameters: (1) the selection of either a periodic or random noise excitation, (2) the frequency of the periodic wave (if used), and (3) the coefficients of the digital filter used to mimic the vocal tract response. Continuous speech can then be synthesized by continually updating these three parameters about 40 times a second. This approach was responsible for one the early commercial successes of DSP: the Speak & Spell, a widely marketed electronic learning aid for children. The sound quality of this type of speech synthesis is poor, sounding very mechanical and not quite human. However, it requires a very low data rate, typically only a few kbits/sec.

This is also the basis for the linear predictive coding (LPC) method of speech compression. Digitally recorded human speech is broken into short segments, and each is characterized according to the three parameters of the model. This typically requires about a dozen bytes per segment, or 2 to 6 kbytes/sec. The segment information is transmitted or stored as needed, and then reconstructed with the speech synthesizer. Speech recognition algorithms take this a step further by trying to recognize patterns in the extracted parameters. This typically involves comparing the segment information with templates of previously stored sounds, in an attempt to identify the spoken words. The problem is, this method does not work very well. It is useful for some applications, but is far below the capabilities of human listeners. Most speech recognition algorithms rely only on the sound of the individual words, and not on their context. They attempt to recognize words, but not to understand speech. This places them at a tremendous disadvantage compared to human listeners. This is very much the research area in natural language processing, and is likely to accelerate in the next few years driven by the requirement of implementing the MPEG-7 standards.

Sources

- The Scientist and Engineer's Guide to Digital Signal Processing - S W Smith, California Technical Publishing 1997
- Digital Multimedia - N Chapman and J Chapman, J Wiley, 2000
- MPEG Home page