Perceptual Audio Encoding

By

Marty Fries

Imagimedia.net

Author of "The MP3 and Internet Audio Handbook"

Introduction

Perceptual audio encoding, a form of *lossy compression*, is the technology that makes MP3 possible. Without compression, CD audio takes up about 10mb per minute, at a *bitrate* of 1411kbps (The term *bitrate* means how much *bandwidth* is required to playback a *stream* of audio in *realtime)*. Most people using 56k modems can only move data at 40-50kbps (or worse). At this speed, downloading a 40mb *WAV file* of song would take over two hours. High levels of compression (on the order of 10 to 1) are required to deliver CD quality over a typical internet connection in a reasonable amount of time.

Many people are familiar with *lossless compression* in the form of *ZIP* files. Most kinds of data contain redundant information. If you replace the most common patterns (such as the word "the") in Word document with a short numeric code, the file becomes much smaller. The more frequently appearing patterns in a file are assigned the shorter codes. This is the basis for *Huffman Coding*. Repetitive sequences of the same value, like part of an image where several adjacent pixels are the same color can be replaced by a single code, followed by how many times to repeat it. This is called *Run Length Encoding (RLE)*. When you *decompress* a losslessly compressed file, you get exactly what you put in. Most music does not contain much redundant information. For example, compressing a WAV file using the popular WinZip program results in a file only about 10% smaller.

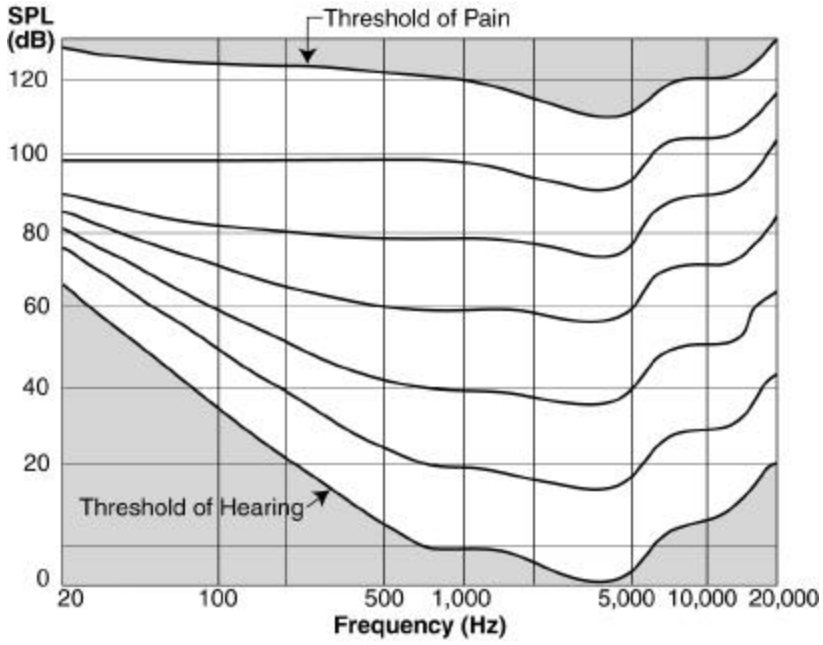
Psychoacoustics

Your senses are *non-linear* and *selective*. Lossy compression works by discarding data that is less significant to the human perception of a picture or a sound. In the case of pictures, details in large areas of similar color (like the sky) are less important than edges in our overall perception. Discarding or averaging some of the variations of blue below a certain threshold may be unnoticeable.

The science of *Psychoacoustics* is the study how our ears and brain perceive and interpret the rapid, periodic air pressure fluctuations that we call sound. The quietest sound you can hear is called the *threshold of hearing*. The perceived loudness of a single tone (at a constant sound pressure level) varies with frequency and sound level. Your ears are most sensitive to midrange frequencies between 1 kHz and 4 kHz (the loudness button on some older stereos boosts bass and treble at low levels in an attempt to compensate for this). A well know illustration of this effect is the Fletcher-Munsen Curve, shown below:

The threshold of hearing only measures your ear's sensitivity to a single tone. Most music

Fletcher-Munsen Curve



consists of much more complex information. What happens when multiple frequencies are present?

Frequency Masking is when softer sounds are "masked" by louder sounds that are close to each other pitch. When sounds are close together in pitch, the louder sounds drowns out the softer. *Temporal Masking* occurs when sounds that follow a short loud sound (like a snare drum hit) can't be heard while your ear recovers from the shock. This is most significant in the 20ms following the loud sound. Since you can't hear these masked sounds, they can be removed from the signal without affecting the perceived quality.

Perceptual audio encoders can compress audio files roughly 10 to 1 while still retaining very high quality. Each person's idea of "CD Quality" can be very different. The choice of encoder and it's settings directly affect the quality and size of the compressed music. Knowledge of the lossy compression process and the effects of encoding software settings allow each user to weigh the inherent tradeoffs between file size and sound quality. Someone has an 80GB hard drive and a fast cable modem may be willing to go for higher quality at the expense of larger file sizes.

The quality of encoded audio is judged by how close is sounds to the original, uncompressed music. The major factors affecting this are the type of encoder, the bit-rate, the type of music, and the sensitivity of the listener's hearing. Critical factors that vary from song-to-song include frequency range, frequency energy distribution, sample-to-sample variation, and signal differences between channels. Virtually all encoded audio will exhibit varying degrees of "artifacts," especially at lower bit-rates. Common artifacts are chirps, pre-echo, pumping and loss of stereo imaging. Some encoded files may also exhibit defects, such as clicks, gaps and phasing, due to errors introduced process of digital audio extraction from a CD.

Perceptual Encoding

Encoding is the process of converting a stream of uncompressed digital audio to a compressed format. The mathematical process used to encode and decode software is referred to as a codec—as in coder/decoder. Perceptual encoders must analyze the music signal to decide what goes and what stays. Mathematical functions, such as the *FFT (Fast Fourier Transform)* and *Modified Discreet Cosine Transform (MDCT)* are ways of converting a stream of audio data from the *time domain* (a stream of bits: 010101) to the *frequency domain* (like a graph showing the distribution of lows, mids and highs).

Perceptual encoders have several ways of handling stereo information. *Simple Stereo* separately encodes each channel. Because much information is common to both channels, this is not very efficient. *Joint Stereo* takes advantage of this fact by encoding the differences between the left and right channels into one stream of data, and the common information into another. The decoder can then reconstruct the stereo signal perfectly, with no loss. This is also called *MS (middle/side)* stereo, and is the same method used in FM stereo broadcasts. *Intensity Stereo* encodes only the information that is important to perceiving a stereo image. Intensity Stereo is the most efficient, but the stereo image can suffer at lower bit-rates.

Certain musical passages need more bits to maintain quality. Many encoders create a *reservoir of bits* by borrowing from simpler passages and applying them to more complex ones. The overall bit-rate doesn't change, because the available bits are simply shifted around as needed.

The bit-rate for digital audio is expressed in thousands of bits per second (kbps) and has a direct relation to the file size. Other than the specific perceptual encoder used (MP3, AAC, etc.), the bit-rate is the factor that will have the most impact on sound quality. File sizes for encoded audio can be calculated by multiplying the bit-rate by the length in seconds. Files compressed at the same bit-rate will be roughly equal in size, regardless of the codec used.

Constant bit-rate (CBR) encoding uses the same number of bits each second to record a section of silence as it does to record a complex passage of music. An advantage of constant bit-rate encoding is that it will always produce a predictable file size. A CBR encoder setting of 128kbps is what is commonly referred to as "CD Quality" and can yield a compression ratio of over 10 to 1. Many users choose even higher settings of 160 or 192kbps. Variable bit-rate (VBR) is an option in most MP3 encoders that varies the number of bits depending on the complexity of the music. For example, a simple passage with just a vocalist and acoustic guitar needs fewer bits than a passage with a full symphony. In general, VBR will produce better sound quality than CBR at a similar file sizes. VBR also helps maintain a more constant signal-to-noise ratio. A disadvantage of VBR is that some older portable players will not report song lengths and elapsed times properly. The encoder settings for VBR represent an arbitrary scale from lower to higher quality. For example, in the popular Musicmatch Jukebox program, the VBR setting ranges from 0-100. A VBR setting of 76% yields MP3 files only slightly larger than CBR 128k (4.1mb vs. 3.9mb from one 43mb pop WAV file). Some players like Winamp show you the bit-rate as they're playing a file, and you can watch the bit-rate jump from as low as 64k in simple passages to more then 256k in more complex ones.

Formats and Codecs

Encoded audio comes in many formats, and multiple formats will be a fact of life for the foreseeable future. Even formats based on open standards like MPEG are sometimes not compatible with each other because of proprietary components. The format of a digital audio file refers to the structure of audio data (PCM, MP3, etc.) within a file. Often a "wrapper" of additional data is used to add features, such as license management information or streaming capability. Most encoded audio formats are frame-based, which allows them to support the insertion of additional program information in the form of text, graphics and other types of data. Audio files can include almost any type of data, such as copyright information, lyrics, album artwork, and links to the artist's website.

The MPEG (Moving Picture Experts Group) committee is part of the International Standards Organization (ISO) and develops standards for encoding audio and video. MP3 is short for MPEG Audio Layer-III. Other MPEG layers include Layer-I, which was designed for the Digital Compact Cassette (DCC) and Layer-II (MP2), which is widely used within the broadcasting industry. Each layer uses the same basic structure and includes the features of the layers below it. Higher layers offer progressively better sound quality at comparable bit-rates and require more processing power for decoding the audio. AAC (Advanced Audio Coding) was developed under MPEG-2 and is sometimes referred to as MPEG NBC, because it is not backwards compatible with MPEG layers I through III.

The MP3 codec works by dividing the stream of audio data into small blocks of samples. The samples are analyzed using a 1024 point *FFT* to determine the frequency content of each chunk of music. This information is used by the perceptual model to calculate masking thresholds. The block of samples are then filtered into 32 *subbands*, and processed by a MDCT into 576 spectral lines. Each component of the sound is then allocated a number of bits based on its importance (as determined by the psycho-acoustic model) and the target bit-rate. MP3 can sound every bit as good as newer, less open formats (like AAC and WMA), at the expense of using more disk space or bandwidth. The decreasing cost of disk storage and bandwidth, and the wide variety of free or inexpensive MP3 software, hardware, and music will ensure it's continued popularity.

AAC stands for *Advanced Audio Coding.* The AAC codec gets rid of the 32 subbands and uses a higher resolution MDCT to allow finer control over what is discarded. It adds techniques like *prediction* (guessing what comes next) and uses a more flexible bit-stream structure to increase the amount of compression without sacrificing quality. *Quantization noise* is created when fewer bits are used to encode less important parts of the music. *Temporal noise shaping* predicts this, and helps the encoder spread it to parts of the signal that are masked by the music.

Real Audio was the first widely used system for listening to music over the Internet. It's used by many Internet radio stations and online music stores for streaming music in realtime over slow modem connections, but is not a great format for high quality downloadable music.

Windows Media Audio (WMA) is a proprietary Microsoft format and codec. Microsoft claims that WMA sounds more like the original CD than MP3, and uses only half the bandwidth. To support this claim, they paid the National Software Testing Lab (NTSL) to conduct a test using comparing WMA at 64k with a now-outdated version of Music Match Jukebox at 128k CBR, using ordinary PC speakers in an office environment.

AT&T's a2b music, Global Music's MP4 and Liquid Audio are all systems for music delivery that are based on AAC. They both include schemes for copyright enforcement and royalty tracking. This is referred to as *rights management*.

Ogg Vorbis is a high-quality, non-proprietary, patent-free, open source, compressed audio format and streaming technology. It's still under development but is expected to offer similar quality to AAC. Many people in the on-line music community are actively working to add Ogg Vorbis support to widely used programs like Winamp.

Summary

Regardless of codecs or formats, perceptual audio encoding is transforming the landscape of the music world today. Future refinements in encoding algorithms, combined with increases in bandwidth and storage densities will someday allow all recorded music to be accessed by anyone, anywhere, anytime. This exciting future is the direct result of the applied sciences of psychoacoustics and perceptual audio encoding.