Xander Arnold
Math 5 Project
Alexander Barnett

MP3 Compression

The MP3 is a compressed file format that uses concepts of psychoacoustics and human perception. Research into file compression that led to the creation of the MP3 began in the late 1970's at Erlangen-Nuremberg University. A research group there was looking into file compression, as computers and transmissions of files back then were extremely slow, and one Brandenberg began using ideas of psychoacoustics to limit the data in files. This group then joined forces with the Fraunhofer Institute to create an alliance, that later was recruited by the International Standardization Organization (ISO) to form the Moving Picture Experts Group (MPEG). Along with the help of other technology development companies, MPEG's job was to make standards for compressed digital audio and video.

As his thesis, Bandenberg developed OCF, which stands for Optimum Coding in the Frequency-Domain. This was then improved, which resulted in ASPEC or Adaptive Spectral Perceptual Entropy Encoding. MPEG decided on providing multiple standards for audio and therefore came out with MPEG-1 layer1, layer 2 and layer audio. Layer 3 used ASPEC and because it was most efficient, it became to use for storing files on hard drives and transferring them over the Internet. These were all published in 1993. Two years later MPEG published MPEG-2, which allowed additional bitrates—different levels of audio quality—, and more channels—so it now could support 5.1-surround sound. That same year the abbreviated name "mp3" was coined.

The reason the mp3 is so great is that it is 1/10-1/12 the size of a CD, depending on bitrate. This allows for very efficient uploads, downloads, and facilitates minimal use of storage space. The type of file compression for mp3 is Lossy, which means data is actually being thrown away during the compression process and cannot be restored. The reason that mp3 compression can get away with this is by using principles or actually more like theories of psychoacoustics. These theories are based on the assumptions that the mind does not process all that the senses do but in fact ignores or represses information that it deems unimportant and also that recording devices are more sensitive than the human auditory system. Therefore audio files potentially contain information that cannot be noticeably perceived by a human being.

The first psychoacoustic concept that mp3 compression takes into account is auditory masking. This is when two notes of very similar frequencies are played at the same time -one quieter than the other- and the human ear cannot distinguish the two discrete pitches. This is also called simultaneous masking. (There is an example attached) Auditory masking also occurs when a loud low frequency is played at the same time as a soft high frequency. This occurrence is due to the way sound waves are sensed by the little hairs of the cochlea in the inner ear: Lower frequencies are sensed further down the tube than high frequencies and they create a tail that can engulf the stimulation of hairs earlier on and therefore mask higher frequencies.

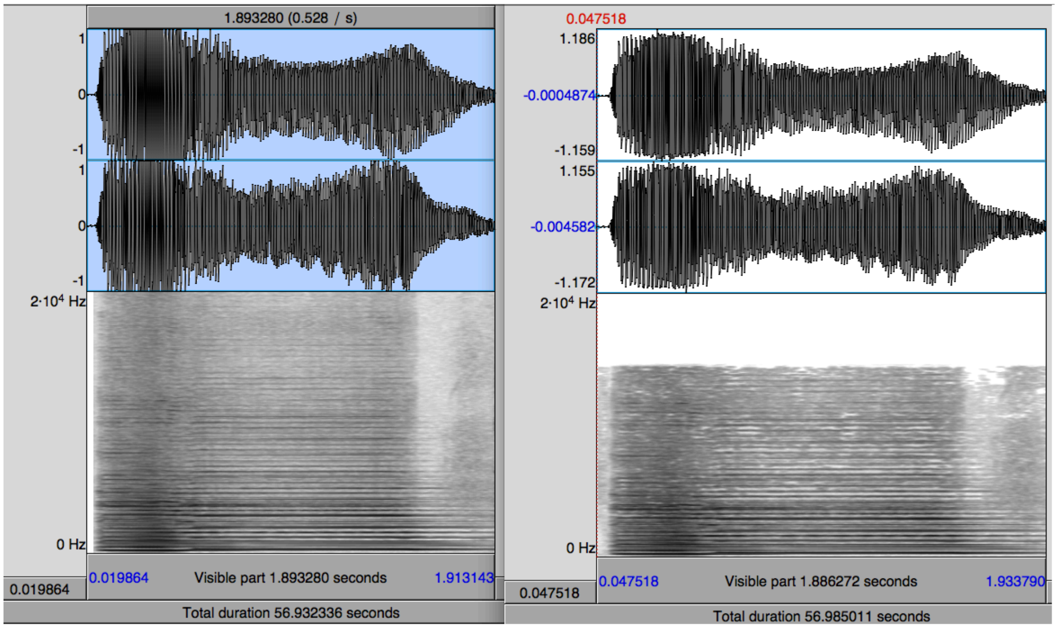
Another major type of masking in the psychoacoustic world is temporal masking. Where auditory masking involves the relationship between frequencies and their relative volumes, temporal masking involves the relationship between time placement of pieces of audio and their relative volumes. More specifically, if you have a quiet sound and a loud sound and you play them at the same time, you will most likely only hear the loud sound. What is more interesting though is that if you play the sounds really close in time to one another, without overlapping, you still probably will not here the quieter sound. (I have also attached an example of this) The average threshold time between sounds where you will be able to here both is on average 5 ms for pure tones. Other psychoacoustic concepts that mp3 compression takes into account are that the average human adult cannot hear over 16-18 kHz and also the human ear is really insensitive to the direction really high and really low frequencies are coming from. All of these concepts are put into effect during the encoding process of the audio.

An mp3 encoder takes a Fast Fourier Transform of the audio with a time window that is a fraction of a second. Each frame of the FFT is then broken into sub-bands based on frequency content because different variations of the applied algorithms work more accurately on specific frequency ranges. Then the bit rate, or the maximum data you can allocate to each frame, is calculated based on the bit rate setting you chose. Then each frame is compared to psychoacoustic models/algorithms. These algorithms account for the temporal and auditory masking; whatever data fits the models is kept and whatever does not is discarded. After all of this, Huffman coding is applied which is the typical sort of compression you would see when creating a .zip file. This type of coding is lossless, meaning no information is thrown out; it is just packaged I a way that takes up less space. Finally the encoder assembles the frames in an organized, readable fashion. Other options are also given to make the file size even smaller and these relate to the other psychoacoustic concepts mentioned previously. This includes being able to remove high frequencies that the human ear cannot hear as well as outputting the really high frequencies (that you have not cut out) only out of one channel rather than two, if it is a stereo file because chances are you will not be able to tell which channel it is coming from any way. After all of this, the result is much smaller file that ideally does not sound any different than the original audio file, ideally.

Example 1: Shows auditory masking exactly like in the picture in the power point but with a different frequency. A loud pitch stays the same frequency the whole time while the quieter one slowly and constantly increases although the human ear cannot quite distinguish the sound as two frequencies until the end.

Example 2: Shows temporal masking. There is a loud plastic bag and a quiet snap. First they are played isolated, or far enough apart temporally that the human auditory system can distinguish the sounds. Then the snap is played directly after (within 5ms of) the plastic bag sound. Afterwards the snap is played directly before the plastic bag sound.

Example 3: Below are the waveforms and corresponding spectrograms I analyzed in Praat of the beginning of a recording of a violin in an electro-acoustic piece. The name of the file is chaconneClipped (ignore the clipping as it was part of the recording); the left is the wav file and the right is the mp3 file. You can very obviously notice that all of the higher frequencies of the mp3 have been cut off. Then if you look closer you will also notice that within the main block of sound, the mp3 has many streaks and splotches of white. This is where information has been thrown out. Notice that the spectrogram of the wav file has absolutely no white (besides the beginning where there is no audio yet) spots; there are places which are very light but because they are not completely white suggests there is data there even if the amplitude is so low that you cannot here. I have attached a larger screen shot in the email.


1.893280 (0.528 / s)
1
0
-1
1
0
-1
2·10⁴ Hz
0 Hz
0.019864
0.019864
Visible part 1.893280 seconds
1.913143
Total duration 56.932336 seconds
0.047518
1.186
-0.0004874
-1.159
1.155
-0.004582
-1.172
2·10⁴ Hz
0 Hz
0.047518
0.047518
Visible part 1.886272 seconds
1.933790
Total duration 56.985011 seconds