“A Comparison of the ATRAC and MPEG-1 Layer 3 Audio Compression Algorithms”

Christopher Hoult, 18/11/2002
University of Southampton

Abstract

This paper intends to give the reader some insight into the workings of ATRAC and MPEG-1 Layer 3 and to compare the two approaches to audio compression. ATRAC is used in MiniDisc technology, while MPEG-1 Layer 3 is used extensively as an all-purpose software audio compression technique.

1. Introduction

The purpose of this paper is to give the reader some insight into the workings of two of the most popular compression algorithms, ATRAC and MP3, and compare their approaches to the problem of compressing audio whilst retaining its quality. While it is difficult to compare algorithms that solve different problems, an attempt will be made to make a decision as to which is the better format, on several grounds.

ATRAC is the compression technique upon which the MiniDisk format is based, and has recently been used by Sony in its Memory Stick technology to store audio files.

MPEG-1 Layer 3 was developed by Fraunhofer for the Moving Picture Experts Group as an all-purpose audio encoder suitable for streaming data at variable and fixed bit-rates.

2. Background

2.1 ATRAC [Sony]

In the late 1980s, with the advent and rising popularity of the Compact Disc (CD), and the peak of audio cassette sales, Sony and Phillips entered into discussions on the production of a successor to the cassette - Phillips was of the opinion that a digital cassette would be best, whereas Sony had decided on a magneto-optical (MO) solution, and so both went their own ways. Sony went on to introduce the MiniDisc (MD) in 1992, a MO disc of roughly a quarter the size of a CD in a plastic housing, but retaining the capacity of an audio CD. This was only made possible through the development of Adaptive Transform Acoustic Coding (ATRAC) audio encoding. The MD has since gone on to replace the audio cassette as the recordable audio media of choice, although the development of CD-R, CD-RW and MPEG-1 Layer 3 have prevented it from rising to the same popularity enjoyed by the audio cassette. In recent years, Sony have further developed ATRAC, and are currently employing it in their Memory Stick solid-state audio storage technology.

2.2 MPEG-1 Layer 3 [Fraunhofer-IIS, 2001]

MPEG-1 Layer 3 (better known as MP3) began life in 1987 at Fraunhofer Institut Integrerte Schaltungen (Fraunhofer-IIS) as EUREKA project EU147 for Digital Audio Broadcasting (DAB). In January 1988, a sub-committee of the International Standards Organization/International Electrotechnical Commission (ISO/IEC) called

the Moving Picture Experts Group (MPEG) was formed to develop a standard for low-bandwidth video compression. In 1992, Fraunhofer-IIS's audio algorithm was integrated into the MPEG-1 standard, which was published in 1993 as ISO/IEC 11172 (ISO/IEC 11172-3 relating specifically to the audio, and defining MP3). MP3 has now gained massive popularity on the Internet, especially with the advances in broadband technology. File-sharing programs such as Napster, Morpheus and KaZaA have made MP3 a thorn in the side of the recording industry. The MPEG-1 audio compression scheme includes three layers arranged in a hierarchy so that Layer 3 decoders can decode all layers, Layer 2 decoders can handle Layers 1 and 2 and a Layer 1 decoder can handle only Layer 1 compressed audio.

2.3 The Human Ear

Many audio encoding processes perform perceptually loss less compression by exploiting the limits of the human ear, a technique known as psychoacoustics. There are several of these effects that can be used, but the main three that are used in audio compression are covered here.

2.3.1 Masking

Masking is where a quiet sound is drowned out by a louder sound. Parallel masking occurs in the frequency domain, and non-parallel masking occurs in the time domain (see figure 1). In non-parallel masking, a masker will cover a tone not only if it comes at the same time as that tone, but for a short time before and after as well. This is particularly useful in covering up the quantization noise caused by the compression process, and comes as a by-product of MDCT's 50% overlap.

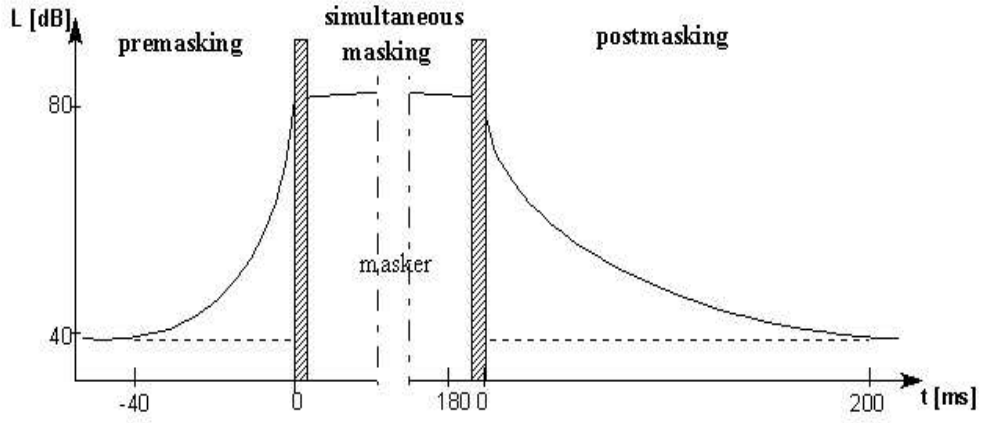

Figure 1 Masking [Gdansk University, 2000]

2.3.2 Audible range

The human ear is limited in its sensitivity to frequencies. The average human can hear tones between 20 Hz and 20 kHz. This means that in any audio, the frequencies outside this range can be removed without problem, as the human ear can' t hear them anyway.

2.3.3 Equi-loudness

Equi-loudness occurs due to the differing sensitivity of the ear according to frequency - two tones at different frequencies, but of the same strength, will not sound as loud as each other. Figure 2 shows equi-loudness curves, and how the ear is most sensitive to frequencies around 4 kHz. The dotted line in the figure is the hearing threshold in quiet. A sone is defined as the loudness of a 40 dB tone at 1kHz [Tsutsui et al, 1992].

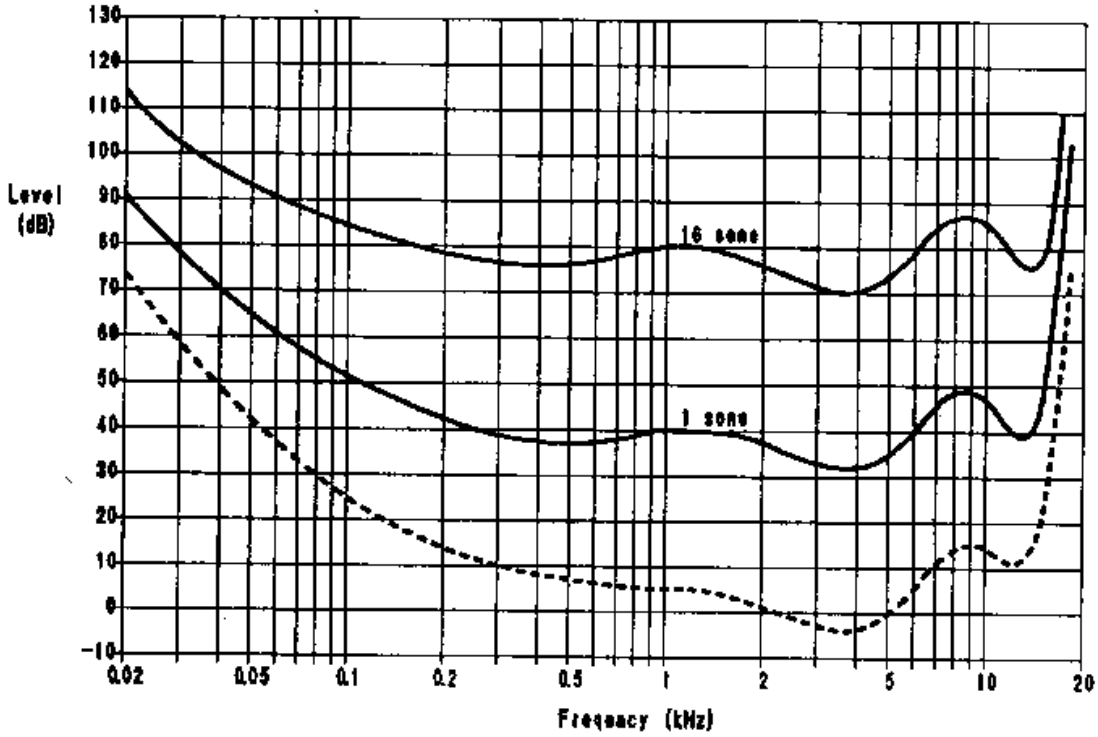

Figure 2 Equi-loudness Curves [Tsutsui et al, 1992]

2.4 Terms Used in This Paper

MDCT - Modified Discrete Cosine Transform. The MDCT is a matrix-based function that converts data from one domain to another. It is similar to the Fourier transform, but deals only with real numbers. In the case of audio compression, it converts audio data from the time domain to the frequency domain. Although one output coefficient does not correspond to one input sample, the overlapping nature of the MDCT allows this discrepancy to be covered up, and is used to great effect in masking quantization noise. [Lincoln, 1998][Wikipedia, 2002]

IMDCT - The inverse of the MDCT.

PCM - Pulse Code Modulation. The format used to store uncompressed audio on CDs and in Wave files on computers.

Quantization - the approximation of an analog signal to digital values using non-overlapping sub-bands.

Time Domain – the representation of sound in terms of amplitude versus time.

Frequency Domain – the representation of sound in terms of amplitude versus frequency.

3. Functionality

3.1 Encoding

The compression of any data must be done in such a way that the decoding process can reproduce the original data exactly. However, where human perception is involved, a lot of data can be removed from the original signal whilst retaining the accuracy required to be acceptable - this can be seen in image compression algorithms such as JPEG and many audio compression algorithms such as the two discussed in this paper.

3.1.1 ATRAC [Tsutsui et al, 1992]

The ATRAC encoding process has three components (see Figure 3). The first of these processes is the time-frequency analysis block (see Figure 4) that decomposes the signal into spectral coefficients grouped into Block Floating Units (BFUs). The signal is first split into three sub-bands (0 - 5.5 kHz, 5.5 - 11 kHz and 11 - 22 kHz) through the use of two Quadrature Mirror Filters (QMFs), which work by splitting the input signal into two bands, High and Low. Chaining two together (with a delay on the first High band output) allows the original signal to be split into these three bands (the delay removes problems caused by the propagation delay associated with the second QMF). These bands are then transformed into the frequency domain through the use of the MDCT with adaptive block sizes, chosen by the block size decision block (see Figure 4). These can be either long (11.6 ms) or short (1.45ms in the High band, 2.9 in the others), allowing for better frequency resolution in stationary regions of the input signal (using long mode), and better backward masking of quantization noise in attack regions caused by the compression (using short mode, due to the short time-frame associated with backward masking). The resulting spectral coefficients are then grouped into BFUs, each containing a fixed number of coefficients - in long mode, a BFU represents 11.6ms of a narrow frequency band; short mode BFUs represent a shorter time, but a wider frequency band. The concentration of BFUs is greater at lower frequencies, due to psychoacoustic effects in the ear.

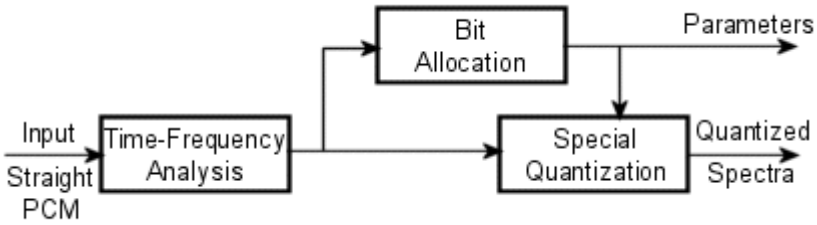

Figure 3 Block Diagram of the ATRAC encoder [Tsutsui et al, 1992]

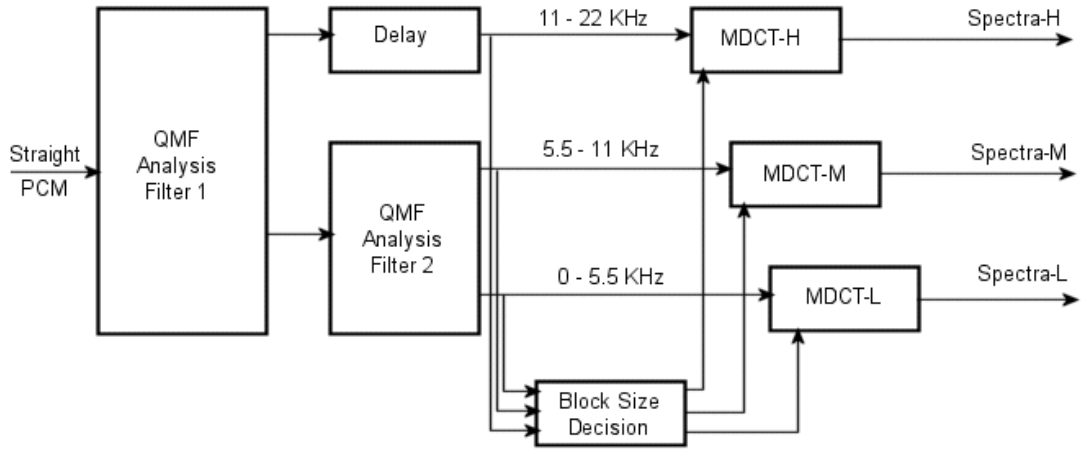

Figure 4 The ATRAC Time-Frequency Analysis Block [Tsutsui et al, 1992]

The next component is the bit allocation block. The ATRAC standard does not define which algorithm to use, to allow the encoding process to evolve without outdating the decoding process, as word length is stored along with the quantized data on the disk. It is important that the chosen algorithm based soundly on psychoacoustic principles, but even with simple allocation algorithms ATRAC retains the quality of its sound. [Tsutsui et al, 1992] suggest the following algorithm:

$b(k) = integer\{b_{tot}(k)-b_{off}\}$

Where:

$b_{tot}(k) = Tb_{var} + (1-T)b_{fix}$

The final component is the spectral quantization block. The spectral values of the audio are quantized using a scale factor defining the range of the quantization, and a word length defining the precision within the scale. Each BFU has the same word length (determined by the bit allocation algorithm) and scale factor, which is chosen from a fixed list of possibilities. For each sound frame, corresponding to 512 input points, the following information is stored:

- MDCT block mode
- Word length
- Scale factor
- Quantized spectra

The first three items of information may be stored redundantly in order to guarantee the accurate reconstruction of the input audio.

3.1.2 MPEG-1 Layer 3 [ISO/IEC, 1995]

Due to the protective nature of the companies that developed the standard, the MP3 specification does not go into detail about their actual encoding process, although it does provide high-level pseudocode so that other parties may produce their own encoders. Figure 5 shows the structure for an MP3 encoder.

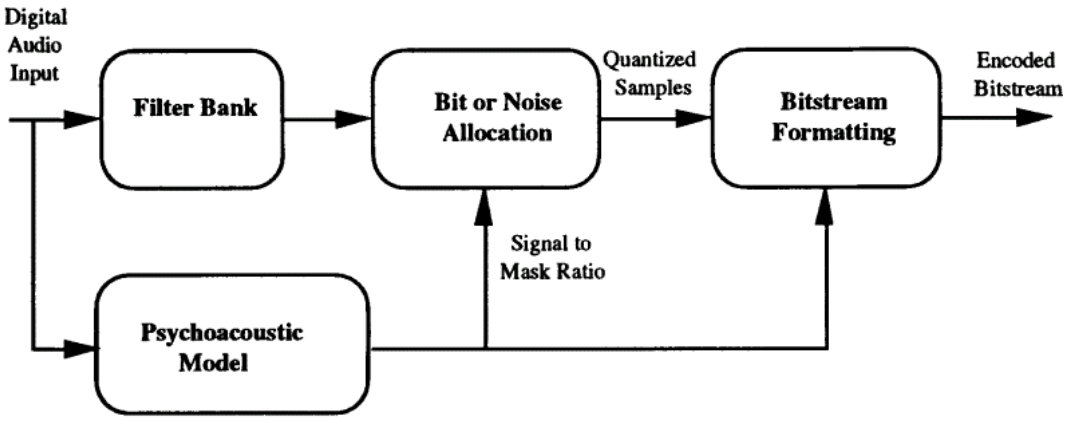

Figure 5 MPEG-1 Layer 3 Encoder [ISO/IEC 1995]

Input audio samples are fed into the encoder and the mapping process creates a filtered and sub-sampled representation of the input audio stream, called transformed sub-band samples. An appropriate psychoacoustic model is used to provide a set of data with which to control the quantization and coding. [ISO/IEC, 1995] suggests using an estimation of the masking threshold to do this control. The quantizer and coding process creates a set of coding symbols from the mapped input samples. The frame packing stage then assembles the encoded bitstream from this data, and adds in any redundant data such as error correction codes.

In MPEG-1, as the Layer number increases, the complexity of the encoder also increases. Layer 3 is built upon the lower Layers, and includes their features. In Layer 1, the input is mapped into 32 sub-bands with fixed segmentation of the data into blocks. The psychoacoustic model determines the adaptive bit allocation, and quantization is performed using block "companding" (compression and expanding) and formatting. Layer 2 uses additional coding of bit allocation, scale factors and samples, as well as a different framing technique. Layer 3 adds increased frequency resolution through a hybrid filterbank, a non-uniform quantizer, adaptive segmentation and entropy coding (such as Huffman Coding) of the quantized data.

3.1.3 Comparison

In general, the ATRAC and MP3 encoding standards share the same structure - both map the input PCM into sub-bands and blocks for easier and more uniform processing; both utilise psychoacoustic models to provide the control data for the quantization process; and both processes use MDCT to transform the data into the frequency domain and to cover up quantization noise through forwards and backwards masking. In fact, comparing Figures 3 and 5, one could say that the processes are identical, with respect to semantics.

However, the algorithms differ in a few key ways. First of all, ATRAC has a fixed data rate, whereas MP3 has a variable one. This is due to the fact that MP3 uses Huffman coding, and so the block size varies with the input data. While this leads to impressive performance and is well-suited to audio streaming, it does have the drawback that track length cannot be determined by size alone - that data must be stored in the header of the file, estimated through the size, or determined by a quick pass through the file itself. ATRAC's fixed-size blocks mean that its length may be calculated by size alone.

The Huffman part of the MP3 algorithm adds another level of compression for which ATRAC has no equivalent. This is not needed, though, as the specifications for ATRAC only require 74 minutes of audio on a MiniDisc, and so compressing even further than necessary would be pointless - indeed, Huffman coding of the output could be counterproductive to ATRAC, as the encoding process would run the risk of producing too much data to fit on the storage medium. With MP3, the benefits of Huffman coding out-weigh the dangers, as the object is to produce maximum compression, and not to hit a specific time/data relationship.

3.2 Decoding

The decoding process is as important as the encoding process, if not more so. If encoded audio cannot be suitably reconstructed from its compressed form, the whole point of the exercise is lost.

3.2.1 ATRAC [Tsutsui et al, 1992]

The ATRAC decoding process consists of two stages (see Figure 6). In the spectral reconstruction stage, the word length and scale factor stored with the compressed data are used to reconstruct the MDCT spectral coefficients from the quantized values. These spectral coefficients are transformed back into the time domain by the IMDCT using the appropriate block length as specified by the BFU mode (long or short). Finally, the output audio signal is synthesised from the three resulting signals by a bank of Quadrature Mirror synthesis filters. This results in a straight PCM signal, returning to an accurate reconstruction of the original PCM input.

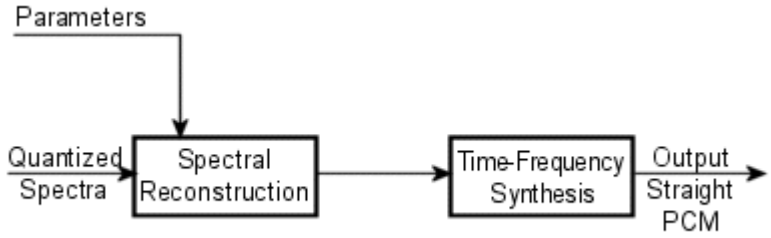

Figure 6 ATRAC Decoder [Tsutsui et al, 1992]

3.2.2 MPEG-1 Layer 3 [ISO/IEC, 1995]

Common to decoders for each Layer of MPEG-1, the decoding and reconstruction process begins with the synchronization of the decoder with the incoming bitstream. This is done by searching the bitstream for a syncword, the position of which can be determined by header information at the beginning of the stream. Other information is also gleaned from this header, to be used as control information later on in the decoding process. This data includes CRC information, scale factors and table select information (for the selection of the appropriate Huffman code tree). The data blocks from the bitstream are then decoded using the Huffman tree, before being fed into a requantizer and then mapped back into the time domain through the use of the IMDCT.

3.2.3 Comparison

As with encoding (see section 3.1), the decoding processes of ATRAC and MP3 are very similar. This is probably due to the fact that the encoding processes are similar, and so their inverses must be. The main difference is the syncwords and frame information, which are not required in ATRAC, as the data rate and block sizes are fixed, and so synchronization need not occur. However, this frame information can still be stored redundantly in ATRAC, to allow the expansion and evolution of the decoding system. MP3 uses the syncwords to facilitate audio streaming, allowing the stream to pick up at any point.

3.3 Quality

It is hard to objectively measure the performance of compression in the realm of perceptual media, due to the variation in human senses, and the qualitative nature of such a process. However, some attempt has been made to do this. [Ruben, 1999] states "The bottom line is this: mp3s sound bad." from the point of view of an "audiophile". The argument is that, by exploiting weaknesses of the human ear above 12 kHz, MP3 "lacks the dynamic range, sound stage imaging and general ' air' " of CDs and audio cassettes. In fact, the author goes on to recommend that Sony license ATRAC technology for software use to replace MP3.

[Churchill, 1999], whilst not comparing ATRAC and MP3 directly, reports that with ATRAC "... small decrease in imaging, ' space' around instruments not as pronounced, sound is more 2 dimensional ..." when compared to CD. In the same tests, MP3 was described as "... highs washed out. Lows still there but lack ' oomph' ...".

While it is difficult to judge conclusively which algorithm gives best quality, the general opinion of similar listening tests is that ATRAC outperforms MP3 when the latter is encoded at a similar bit-rate. When amplifying certain frequencies in the output of both formats, the lost information is more conspicuous, but it is even more so in MP3, resulting in a "bubbly" sound.

3.4 Compression ratios

For ATRAC, [Tsutsui et al, 1992] only specifies the compression ratio as "less than 5:1" although various other sources on the Internet ([MiniDisc.org], [Ruben, 1999]) quote this ratio as 4.83:1.

While MP3 shows better compression performance than ATRAC, it should be noted that, as stated previously in this paper, the original ATRAC algorithm does not need to achieve better results, as its aim was to fit audio from one CD onto one MD, a total of 74 minutes.

[Fraunhofer-IIS, 2001] states that the compression ratios for MPEG-1 Layer 3 are those below.

sound quality	bandwidth	mode	bitrate	reduction ratio
telephone sound	2.5 kHz	mono	8 kbps *	96:1
better than short-wave	4.5 kHz	mono	16 kbps	48:1
better than AM radio	7.5 kHz	mono	32 kbps	24:1
similar to FM radio	11 kHz	stereo	56...64 kbps	26...24:1
near-CD	15 kHz	stereo	96 kbps	16:1
CD	>15 kHz	stereo	112..128kbps	14..12:1
*) Fraunhofer uses a non-ISO extension of MPEG Layer-3 for enhanced performance ("MPEG 2.5")				

4. Conclusions

As the domains of ATRAC and MP3 begin to overlap, they move more and more into direct competition. While ATRAC is currently hardware-based, and MP3 software-based, new advances in audio technology mean that the boundaries between their domains are beginning to blur. If the criterion for pronouncing a "winner" were compression ratio, MP3 would come out on top. If the criterion were quality, ATRAC would come out on top. Both standards are still evolving - MPEG-1 will soon be replaced by the superior MPEG-2 with its wavelet technology; ATRAC comes in several different forms, and is currently up to version 7.

Once the two algorithms' domains overlap completely, consumers will have a tough choice between ATRAC and MP3 – indeed, with so many competing audio formats, and so many in development, the choice gets tougher every day. However, if, today, the applications of both ATRAC and MP3 overlapped, consumers wishing to obtain highest-quality compressed audio would be best advised to choose ATRAC, due to its superior performance and sound reproduction at it's chosen bit-rate. MP3 will still be most useful for computer- and internet-based applications, as its streaming qualities and low file size provide good service to those end-users who do not care too much about the quality of their audio.

5. References

Churchill, 1999: **"***Format Listening Tests: CD, MD (ATRAC 4.5 & 3.0), MP3, VQF, RM*" - http://www.minidisc.org/format_comparison.html - Guy Churchill (last modified 19th March 1999)

Fraunhofer-IIS, 2000: "*MPEG Audio Layer 3*" - http://www.iis.fraunhofer.de/amm/techinf/layer3/index.html - Fraunhofer-IIS

Gdansk University, 2000: "*Fundamentals of Psychoacoustics*" - http://sound.eti.pg.gda.pl/SRS/psychoacoust.html - Gdansk University

ISO/IEC, 1995: "*Information technology— Coding of moving pictures and associated audio for digital storage media at up to about1,5Mbit/s — Part3: Audio*" - EN ISO/IEC11172-3:1995

Lincoln, 1998: "*Modified Discrete Cosine Transform (MDCT)*" – http://ccrma-www.stanford.edu/~bosse/proj/node27.html - Bosse Lincoln (last modified 7th March 1998)

MiniDisc.org: "*MiniDisc FAQ: Audio Topics*" - http://www.minidisc.org/faq_sec_4.html

Ruben, 1999: "125 Hertz, or the Myth of MP3 Hi-Fi" - http://www.macopinion.com/columns/curmudgeon/99/01/26.html - Matthew Ruben (last modified 26th January 1999)

Sony: "MiniDisc History" - http://www.angelfire.com/empire/minidisc/sony_md_history.htm

Tsutsui et al, 1992: "*ATRAC: Adaptive Transform Acoustic Coding for MiniDisc*" - MPEG/AUDIO CA11172-3, 1992 - Kyoya Tsutsui et al.

Wikipedia, 2002: "*Discrete cosine transform*" - http://www.wikipedia.org/wiki/Discrete_cosine_transform - (last modified 20:47 Dec 8, 2002)