

12 Audio Coding Reference

12.1 Introduction to Audio Coding Technology

Introduction

Audio takes up a lot of data. Just a regular phone call uses 64,000 bits per second.

Without data reduction, CD-quality quality audio — 16 bits at 44.1kHz sample rate — requires a transmission capacity of about 706 thousand bits per second (kbps) for each audio channel. But, the wires we use for remote broadcasting are on a telephone system designed for voice-grade communications: 8 bits at 8kHz sample rate, or 64 thousand bits per second (kbps) per channel. That's 11% of what we need.



CURIOSITY NOTE!

*You can arrive at these same numbers with nothing more complicated than grade-school math. Just multiply the sample rate by the sample depth: 44,100 samples per second * 16 bits per sample = 705,600 bits per second for CD-quality mono audio. Multiply by 2 for stereo.*

You can reduce the data requirements by lowering the quality somewhat. 13 bits would yield a respectable 78 dB dynamic range, certainly adequate for casual home listening. And a 32kHz sample rate — with careful equipment design — will give you flat response to 15kHz, the practical limit for analog FM broadcasting in North America. Unfortunately, that still leaves us with telephone data channels about 93% too small to do the job. Besides, 13 bits is an awkward bit depth (resolution) for computers to deal with, and the audio it produces isn't clean enough to survive today's transmitter processors.



CURIOSITY NOTE!

Bit depth and sample rate translate easily into audio specifications. Digital audio must have a sample rate of at least twice the desired bandwidth, so 15kHz audio requires (after a safety margin) 32kHz sampling.

Each bit of sample depth represents slightly more than 6dB of dynamic range.

The first practical coding methods used a principle called ADPCM, Adaptive Delta Pulse Code Modulation. This takes advantage of the fact that it takes fewer bits to code the difference, or delta, between successive audio samples compared to using the individual

values. Further efficiency is had by adaptively varying the difference comparator according to the nature of the program material. G.722 and APT-X are examples of ADPCM schemes. They achieve around a factor of 4:1 reduction in bitrate.

G.722 achieves additional efficiency by allocating its bits to match the patterns in the human voice, and it's considered adequate for news and talk programming over ISDN. But, for high-fidelity transmission, algorithms with more power are required. These are based on psychoacoustics, where the coding process is adapted to the way we hear sounds. There are several algorithms available, with varying complexity and performance levels.

Some years ago, the international standards group ISO/IEC established the ISO/MPEG (*Moving Pictures Expert Group*), to develop a universal standard for encoding moving pictures and sound for digital storage and transmission media. The standard was finalized in November 1992 with three related algorithms, called Layers, defined to take advantage of psychoacoustic effects when coding audio. Layer 1 and 2 are intended for compression factors of about 4:1 and 6 or 8:1 respectively, and these algorithms have become popular in satellite and hard-disk systems. Layer-3 achieves compression up to 12.5:1 — 8% of the original size — making it ideal for ISDN.

Basic Principles of Perceptual Coding

With perceptual coding, only information that can be perceived by the human auditory system is retained.

Lossless — which, for audio, translates to noiseless — coding with perfect reconstruction would be an optimum system, since no information would be lost or altered. It might seem that lossless, redundancy-reducing methods (such as PKZIP, Stuffit, Stacker, and others used for computer hard-disk compression) would be applicable to audio. Unfortunately, no constant compression rate is possible due to signal-dependent variations in redundancy: There are highly redundant signals like constant sine tones (where the only information necessary is the frequency, phase, amplitude, and duration of the tone), while other signals, such as those which approach broadband noise, may be completely unpredictable and contain no redundancy at all. Furthermore, looking for redundancy can take time: while a popular song *might* have three choruses with identical audio data that would need to be coded only once, you'd have to store and analyze the entire song in order to find them. Any system intended for a real-time use over telephone channels must have a consistent output rate and be able to accommodate the worst case, so effective audio compression is impossible with redundancy reduction alone.

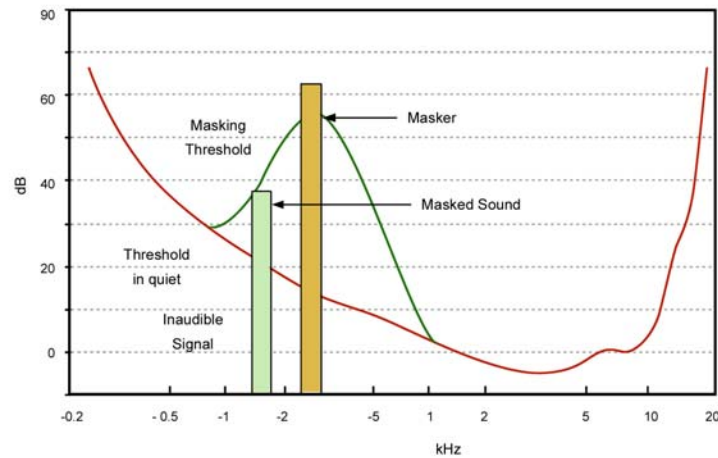
Fortunately, psychoacoustics permits a clever solution! Effects called “masking” have been discovered in the human auditory system. These masking effects (which merely prove that our brain is also doing something similar to bit rate reduction) have been found to occur in both the frequency and time domains and can be exploited for audio data reduction.

Most important for audio coding are the effects in the frequency domain. Research into perception has revealed that a tone or narrow-band noise at a certain frequency inhibits the audibility of other signals that fall below a threshold curve centered on a masking signal.

The figure below shows two thresholds of audibility curves. The lower one is the typical frequency sensitivity of the human ear when presented with a single swept tone. When a single, constant tone is added, the threshold of audibility changes, as shown in the upper curve. The ear's sensitivity to signals near the constant tone is greatly reduced. Tones that

were previously audible become “masked” in the presence of “masking tones,” in this case, the one at 300 Hz.

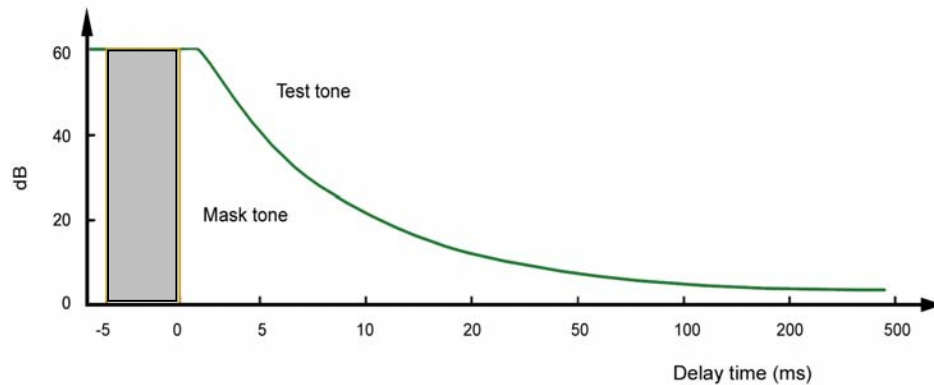
All signals below the upper “threshold of audibility” curve, or Masking Threshold are not audible, so we can drop them out or quantize them crudely with the least number of bits. Any noise which results from crude quantization will not be audible if it occurs below the threshold of masking. The masking depends upon the frequency, the level, and the spectral distribution of both the masker and the masked sounds.



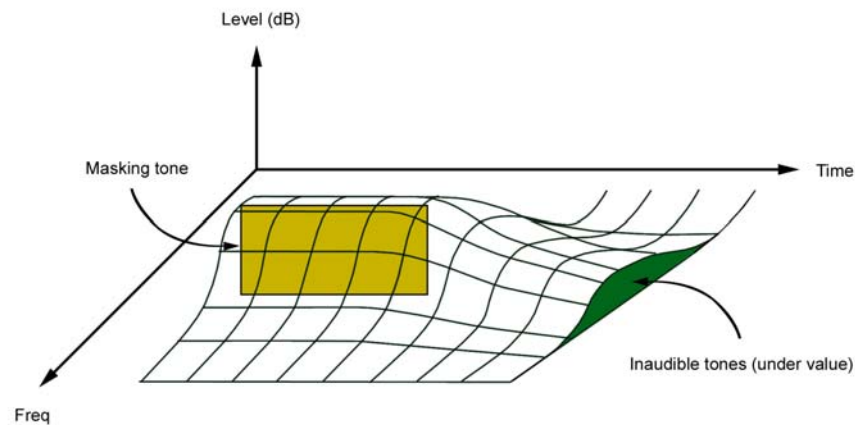
Masking effects in the frequency domain. A masking signal inhibits audibility of signals adjacent in frequency and below the threshold.

To benefit from the masking effects, perceptual coders use a filter bank to divide the input audio into multiple bands for analysis and processing. The maximum masked noise level is calculated depending upon the spectral content, and the available bits are allocated so as to keep the quantization noise below the masking threshold at every point in the spectrum.

While coding efficiency increases with more bands and better frequency resolution, the time domain resolution decreases simultaneously owing to an inevitable side-effect of the band filtering process: higher frequency resolution requires a longer time window – which limits the time resolution. Happily, masking works also in the time domain. A short time before and a longer time after a tone is switched on and off, other signals below a threshold amplitude level are not noticeable. Filter banks with higher frequency resolution naturally exploit the ear’s time-masking properties.



Masking effects in the time domain. Masking occurs both before and after the masking signal.



The combined results of time and frequency vs. masking. Signals under the curve are inaudible.



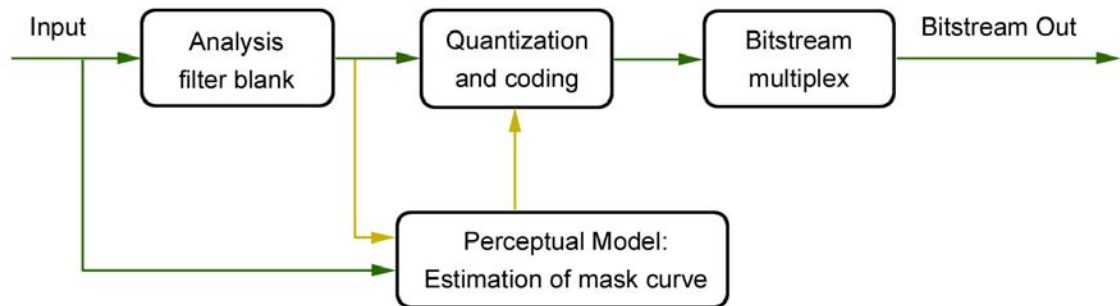
IMPORTANT!

Due to a perceptual coder's reliance on precisely modeling principles of human perception, audio to be coded should not be processed with any non-linear dynamics-processing such as clipping, multi-band compression, or hard limiting. Wideband compression, or AGC, is acceptable, and may be desirable if a consistent level cannot otherwise be achieved.

The same is true of audio that has been decoded, after passing through a perceptual coding cycle, but to much lesser degree.

*For more information on this follow the link to our Omnia Audio website that has a paper delivered at the AES by Frank Foti on this topic.
<http://www.omniaaudio.com/techinfo/default.htm>*

The steps involved in the perceptual coding process are shown below:



The components work as follows:

- The analysis filter bank divides the audio into spectral components. Sufficient frequency resolution must be used in order to exceed the width of the ear's critical bands, which is 100 Hz below 500 Hz and 20% of the center frequency at higher frequencies.
- The estimation of masked threshold section is where the human ear/brain system is modeled. This determines the masking curve, under which noise must fall.
- The audio is reduced to a lower bit rate in the quantization and coding section. On the one hand, the quantization must be sufficiently coarse in order not to exceed the target bit rate. On the other hand, the error must be shaped to be under the limits set by the masking curve.
- The quantized values are joined in the bit stream multiplex, along with any side information.

Doing audio coding effectively means managing several tradeoffs. Most important is the number of samples coded together in one frame. Long frames have high delay, but are more efficient because the header and side information is transmitted less frequently. Longer frames offer the possibility to use filter banks with better frequency resolution. A fundamental principle in signal processing is that spectral splitting filters may have either good time resolution, or good frequency resolution, but not both. This makes sense when you consider that a longer time window means that the analyzer has more complete information, more full audio cycles, to work with⁴.

In the case of rapidly changing input signals (transients) long frames are poorer than short ones because the time spread will lead to so called *pre-echoes*. For such signals, the size of the frame should correspond to the temporal resolution of the human ear. This can be achieved by using short frames or by changing the frame length according to the immediate characteristics of the signal.

⁴ Perhaps this is the DSP designer's equivalent to the economist's TANSTAAFL: There ain't no such thing as a free lunch.

MPEG

By far, the most popular high fidelity audio coders rely upon techniques developed under the MPEG umbrella. MPEG stands for Motion Pictures Expert Group, a Joint Committee of the International Standards Organization (ISO) and the International Electrotechnical Commission. Over a decade ago, when the CD had just been introduced, the first proposals for audio coding were greeted with suspicion and disbelief. There was widespread agreement that it would simply not be possible to satisfy golden ears while deleting 80% or more of the digital audio data. But the audio coding pioneers were persistent and the MPEG audio group was formed. Since 1988, they have been working on the standardization of high quality audio coding. Today, almost all agree not only that audio bit rate reduction is effective and useful, but that the MPEG process has been successful at picking the best technology and encouraging compatibility across a wide variety of equipment.

The MPEG process is open and competitive. A committee of industry representatives and researchers meet to determine goals for target bit rate, quality levels, application areas, testing procedures, etc. Interested organizations that have something to contribute are invited to submit their best work. A careful double blind listening test series is then conducted to determine which of the entrant's technologies delivers the highest performance. The subjective listening evaluations are done at various volunteer organizations around the world that have access to both experienced and inexperienced test subjects. Broadcasters are the most common participants with many of the important test series conducted at the BBC in England, the CBC in Canada, and NHK in Japan. Finally, results are tabulated, a report is drafted and ultimately a standard is issued.

In 1992, under MPEG1 (the first of the MPEG standards), this process resulted in the selection of three related audio coding methods, each targeted to different bit rates and applications. These are the famous layers: 1, 2 and 3. As the layer number goes up, so does performance and implementation complexity. Layer 1 is not much used. Layer-2 is widely used for DAB in Europe, audio for video, and broadcast playout systems. Layer-3 – which Telos was the first to use in the Zephyr – is widely used in broadcast codecs and has gone on to significant Internet and consumer electronics fame under the moniker derived from the file extension: *MP3*. MPEG2 opened the door for new work, and some minor improvements were added to both Layers 2 and 3. In 1997, the first in the AAC family was added to the MPEG2 standard. MPEG4 is ongoing now, but it has already been decided that AAC will be the “general audio” coder under this umbrella. (MPEG3 was skipped for reasons unknown.)

MPEG4 AAC (Advanced Audio Coding)

The MPEG4 AAC system is the newest audio coding method selected by MPEG. It is a fully state-of-the-art audio compression tool that provides performance superior to any known approach at bit rates greater than 64 kbps and excellent performance relative to the alternatives at bit rates reaching as low as 16 kbps.

The idea that led to AAC was not only to start fresh, but also to combine the best work from the world's leading audio coding laboratories: Fraunhofer, Dolby, Sony, and AT&T were the primary collaborators that offered components for AAC. The hoped for result was ITU (International Tele-communications Union) “indistinguishable quality” at 64 kbps per mono channel. That is, quality indistinguishable from the original, with no audio test item falling below the “perceptible, but not annoying” threshold in controlled listening tests.

The MPEG test items include the most difficult audio known to codec researchers, so this was a daunting challenge. The thinking was that if a codec could pass this test, it would surely be transparent for normal program material like voice and pop music, which are much more easy to encode. AAC designers chose to use a new modular approach for the project, with components being plugged-in to a general framework in order to match specific application requirements and the always-present performance/complexity/delay tradeoffs.

Compared to the previous layers, AAC takes advantage of such new tools as temporal noise shaping, backward adaptive linear prediction and enhanced joint stereo coding techniques. AAC supports a wide range of sampling rates (8–96 kHz), bit rates (16–576 kbps) and from one to 48 audio channels.

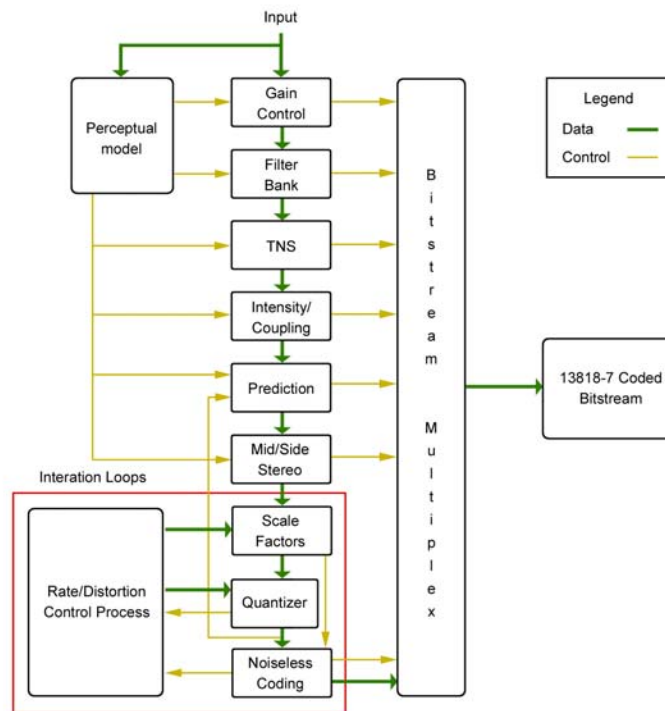
AAC is a lot more sophisticated than the previous MPEG layers 2 & 3, providing significantly more coding power. Because it uses a modular approach, new ideas can be developed and plugged in to the basic structure. This had the additional advantage that it was possible to combine the various components from different developers, taking the best pieces from each. AAC was built on a similar structure to Layer-3, and thus retains most of its features. Nevertheless, compared to the previous MPEG layers, AAC benefits from some important new additions to the coding toolkit:

- An improved filter bank with a frequency resolution of 2048 spectral components, nearly four times the number used by Layer-3.
- Temporal Noise Shaping, a new and powerful element that minimizes the effect of temporal spread. This benefits voice signals, in particular.
- A Prediction module guides the quantizer to very effective coding when there is a noticeable signal pattern, like high tonality.
- Perceptual Noise Shaping allows a finer control of quantization resolution, so bits can be used more efficiently.

Because it uses this modular approach, an implementer may pick and choose among the component tools to make a product with appropriate performance/complexity ratios. Or, new modules can be developed later and "plugged in" to its basic structure. Three default profiles have been defined, using different combinations of the available tools:

- **Main Profile.** Uses all tools except the gain control module. Provides the highest quality for applications where the amount of random accessory memory (RAM) needed is not constrained.
- **Low-complexity Profile.** Deletes the prediction tool and reduces the temporal noise-shaping tool in complexity.
- **Sample-rate Scalable (SRS) Profile.** Adds the gain control tool to the low complexity profile. Allows the least complex decoder.

The block diagram of the AAC encoder is shown below. It is considerably more sophisticated than the previous MPEG Layer-2 and Layer-3 systems, and therefore offers more coding power.



Because AAC was built on a similar structure to Layer-3, it therefore retains some of its powerful features:

- **Redundancy Reduction.** A Huffman encoding process causes values that appear more frequently to be coded with shorter words, while values that appear only rarely are coded with longer words. This results in an overall increase in coding efficiency – with no degradation, since it is a completely lossless process.
- **Bit Reservoir buffering.** Often, there are some critical parts in a piece of music that cannot be encoded at a given data rate without audible noise. These sequences require a higher data rate to avoid artifacts. On the other hand, some signals are easy to code. If a frame is easy, then the unused bits are put into a reservoir buffer. When a frame comes along that needs more than the average amount of bits, the reservoir is tapped for extra capacity.
- **Ancillary Data.** The bit reservoir buffer offers an effective solution for the inclusion of such ancillary data as text or control signaling. The data is held in a separate buffer and gated onto the output bit stream using some of the bits allocated for the reservoir buffer when they are not required for audio.
- **The Joint Stereo mode** takes advantage of the redundancy in stereo program material. The encoder switches from discrete L/ R to a matrixed L+R/ L-R mode dynamically, depending upon the program material.

The result of all this is that the researchers succeeded: AAC provides performance superior to any known codec at bitrates greater than 64kbps, and excellent performance relative to the alternatives at bitrates reaching as low as 16 kbps.

And the researchers succeeded in achieving the ITU goal: AAC is the first codec system to fulfill the ITU requirements for indistinguishable quality at 128 kbps/stereo⁵. It has approximately 100% more coding power than Layer-2 and 30% more power than the former MPEG performance leader, Layer-3. For more information on AAC, and the tests of it, see our web site for a paper on the subject: www.zephyr.com.

It offers:

- 20 kHz mono or stereo audio bandwidth.
- Significantly less delay than Layers 2 or 3.
- Full-fidelity mono on a single 56kbps channel.
- Affordable, transparent, audio transmission for AM/FM radio or television audio.

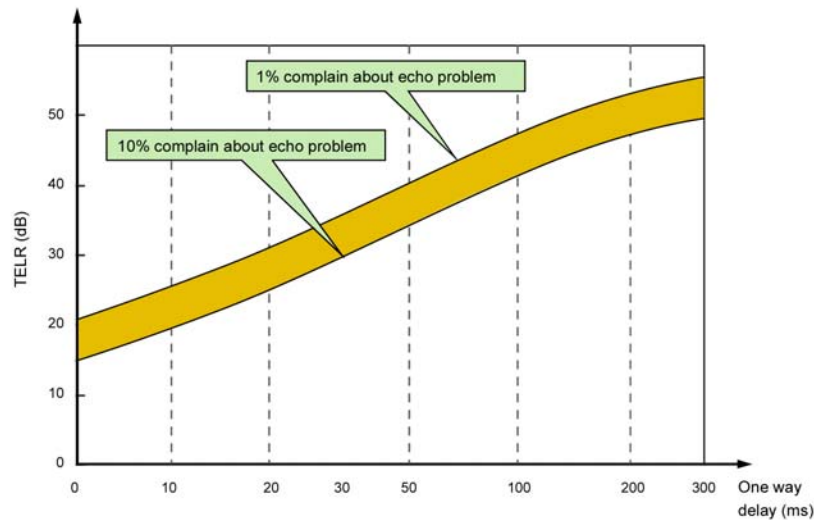
AAC-LD (AAC Low Delay)

When announcers use codecs for broadcast remote applications, they often need to have natural two-way interaction with other program participants located back at the studio, or callers. Because it is a hot topic for engineers working in the field of Internet telephony, a number of studies have been conducted to determine user's reactions to delays in telephone conversations. The data apply directly to the application of professional codecs to remotes, so it is interesting to take a peak over the shoulder of the telecom guys to see what they have learned.

For broadcast remotes, we try to arrange our system so that there is no path for the field announcer's voice to return to his/her headphones. Nevertheless, sometimes echo is unavoidable. For example, this can happen when a telephone hybrid has leakage or when a studio announcer has open-air headphones turned-up loud and the audio makes its way into the studio microphone.

When there is no echo, it has been discovered that anything less than 100 ms one-way delay permits normal interaction between participants. Between 100 and 250 ms is considered "acceptable." ITU-T standard G.114 recommends 150 ms as the maximum for "good" interactivity. Echo introduces a different case. As you might expect, echo is more or less annoying depending upon both the length of time it is delayed and its level. Telephone researchers have measured and quantified reactions, and ITU-T G.131 reports the findings and makes recommendations.

⁵ Subjective Evaluation of State-of-the-Art Two-Channel Audio Codecs, G. A. Soulodre, T. Grusec, et al. Journal of the Audio Engineering Society; Vol. 46, #3; March 1998, Pg 164 - 177.



Summary of ITU-T G.13, with recommendations for designers of telephone systems that must cope with echo. This shows Talker Echo Loudness Rating vs. delay.

There are codecs using other than perceptual technologies that have lower delay, but they are not as powerful. That is, for a given bitrate, they do not achieve fidelity as good as the MPEG systems we have been examining. The common G.722 is an example. It uses ADPCM (Adaptive Delta Pulse Code Modulation), which can have delay as low as 10 ms, but with much poorer quality. So the question arises: Is it possible to have high quality and low delay in the same codec? Until recently, the answer was no. But new developments in codecs have changed the picture. One of the main objectives in audio coding is to provide the best tradeoff between quality and bit rate. In general, this goal can only be achieved at the cost of a certain coding delay. Codecs for voice telephone applications have used ADPCM and CELP because they have much lower delay than perceptual codecs. These are optimized for voice and can have reasonably good performance.

Zephyr users have known for years that Layer-3 offers all the fidelity needed for most broadcast situations. However, they also know that the delay of Layer-3 can be frustrating, particularly if high fidelity is needed in both directions and parties at the two ends must carry on a conversation.

The folks at Fraunhofer were aware of these factors, and have developed an extension to AAC called "AAC Low Delay," or "AAC-LD" for short. AAC-LD offers quality equivalent to Layer-3 with less than 25% of the delay!

AAC-LD combines the advantages of perceptual coders (such as Layer-3) with certain principles of speech coders. Compared to speech coders, AAC-LD handles both speech and music with good quality. Unlike speech coders, however, audio quality scales up with bit rate, and transparent quality can be achieved. AAC-LD's coding power is roughly the same as Layer 3, meaning that mono high fidelity 15 kHz audio may be sent via one ISDN channel. With ISDN's two channels, you achieve near CD quality stereo.

Delay in perceptual codecs is dependent on several parameters:

- Frame length. Time is required to collect all the samples for a frame. The longer the frame, the more the delay.
- Filter bank delay. This causes an additional delay equivalent in time to the frame delay.
- Look-ahead delay for block switching. Layer-3 and AAC use filter banks with high frequency resolution. For signals with high tonality, efficiency is high. But when there are transients, a dynamic switching process changes to a filter bank with lower frequency resolution and better time resolution. In order to correctly decide when to make this change, a look ahead process is required, adding delay.
- Bit reservoir. The length of this buffer determines how much delay this process contributes.

The overall delay is a combination of all of these components, divided by the sampling rate. The delay scales linearly and inversely with the sampling frequency.

How AAC-LD Gets its Low Delay

AAC-LD is based on the core AAC work, so much is similar, but each of the contributors to the delay had to be addressed and modified:

- The frame length is reduced to 512 or 480 samples, with the same number of spectral components at the filter bank output.
- No dynamic block switching is used because the required look-ahead delay is too big. The temporal problem that causes pre-echoes is handled by the Temporal Noise Shaping module.
- The “window shape” of the spectral filter is enhanced to be adaptive. Normally, the shape is a simple sine curve, but AAC-LD can use a shape that has a lower overlap between the bands. This significantly improves performance with transients, without adding any delay.

MPEG-4 High Efficiency AAC (HE-AAC)

HE-AAC is an extension of AAC with the addition of Spectral Band Replication, a technique (developed by Coding Technologies) of synthesizing high frequency audio content based on the lower frequency data and side-chain information. SBR dramatically increases the efficiency of coding when using low bitrates. HE-AAC v1 is AAC plus SBR, while HE-AAC v2 adds Parametric Stereo to further increase the efficiency of coding stereo signals.

MPEG-4 Enhanced Low Delay AAC

High audio quality, low coding delay and very low data rates: AAC Enhanced Low Delay (AAC-ELD) is the perfect choice for any delay critical application that demands full audio bandwidth at data rates down to 24 kbit/s.

AAC-ELD combines the strengths of its two main components, MPEG-4 AAC Low Delay, and Spectral Band Replication (SBR). Whereas MPEG-4 AAC-LD features low encoding/decoding latency, SBR provides high quality audio at very low bit rates. SBR is also used in MPEG-4 HE-AAC, one of today’s most efficient audio codecs.

AAC-ELD is currently under standardization in MPEG. The finalization of the standard is expected for the end of 2007.

ISO/MPEG LAYER-2

MPEG Layer-2 was an extremely popular early perceptual coding method, primarily because it's easier and less expensive to implement — particularly at the encoder — and practical devices using it were available earlier than Layer-3. It's a preferred choice for applications where very large data capacity is available, such as satellite links, high-capacity Primary ISDN or T1 circuits, and hard disk storage systems using Ethernet for signal distribution.

We include it in Zephyr/IP to offer compatibility with the widest variety of codecs, and for use at high bit rates.

Layer-2 J-Stereo

The Layer-2 joint stereo mode uses an “intensity coding” method. This method has high coding power and is quite effective, but hurts stereo separation on some program material. Audio above 3 kHz or so is combined to mono and panned to one of seven positions across the stereo stage.

G.722

This technology pre-dates perceptual coding. It is much simpler than the transform methods, but suffers from poorer audio performance. It has the benefit of low cost and the unique advantage of low delay. It has been around as an international standard the longest, and is probably the most widely used system. In our view, this technology is acceptable for mono voice where high fidelity is not necessary. It is good also for cueing and intercom channels.

We have included G.722 in Zephyr/IP because:

- It had been the most popular coding method early on, so there are many of these codecs in use. Because it is a standard, codecs from various manufacturers have a good probability of being able to interwork with one-another. (We've tested with many units and have found no problems so far.)
- G.722 has the lowest delay of all popular coding methods.

This method was invented in the late 70s and adopted as a standard in 1984 by the CCITT, the *Consultative Committee for International Telephony and Telegraphy* (renamed as ITU-T in 1993). The technique used is Sub-Band ADPCM (Adaptive Delta Pulse Code Modulation), which achieves data reduction by transmitting only the difference between successive samples. G.722 does this in two audio frequency sub-bands: 50 Hz to 4 kHz and 4 to 7 kHz.



DEEP TECH NOTE!

Only two bits are allocated per sample for audio frequencies above 4 kHz – sufficient for conveying the sibilance in voice signals, but not very good for intricate musical sounds. Also, the “predictor model” used to determine the step size in the adaptive function is designed only for speech. This is why music transmitted via G.722 has a distinct ‘fuzzy’ quality.

G.722 has a frequency response extending to 7.5 kHz with fairly poor fidelity. Unless there is no alternative, it should not be used for music.