IMPROVED ISO AAC CODER

Ivan Dimkovic[1]

[1]PsyTEL Research, Belgrade, Yugoslavia
dim@psytel-research.co.yu

AAC, the part of ISO/IEC MPEG-2 and MPEG-4 standards is the most powerful coding algorithm in terms of perceived audio quality for single channel and multichannel audio content. At bit rate of 64 kbits/s per channel MPEG-4 AAC provides transparent quality (indistinguishable from original) according to EBU measures. This renders AAC most suitable algorithm for PC Multimedia and Internet broadcast usage. This paper discusses the implementation effort of optimized AAC encoder based on our independent research and development.

Introduction

AAC was standardized in April 1997. According to listening tests conforming to ITU Recommendation BS.1116 and carried out by BBC, NHK and CRC, AAC delivers “indistinguishable quality” at bit rates of 64 kBits/s per single audio channel. These results showed that AAC is the current state-of-the-art algorithm for high quality audio compression.

At the time when our work was in planning phase there were only few "successful" AAC implementations available (where term "successful" is implying better performance than optimized MP3 encoders at equal bitrate). The reason for this is that standard is purely normative. MPEG committee does not specify encoding algorithm as the normative part of the standard. So-called “informative” annex of MPEG document describes way to design simple AAC codec, but with limited quality because core algorithms are very simplified. Therefore, lot of know-how and research in audio compression field is required to meet audio quality results published in official MPEG tests. For the purpose of listening tests MPEG used high quality software simulation that was developed by few research labs that took part in standardization process. Informative annex was actually derived by simplifying optimized algorithms found in simulations used in listening tests.

In addition to standard itself, MPEG published so-called “reference software”. However, this software is very low speed and quality, and it should be used only to understand normative parts of the standard better. ISO/IEC MPEG Reference software was designed by large group of contributors and it could be said that some of the companies involved in standardization process don’t want public available source code.

Starting Point

The first project goal was to design encoder implementation according to ISO/IEC 11496-3 Annex A (MPEG-4 Audio / Informative part). Such implementation has very simple psychoacoustic model, slow quantizer and basic functionality of all AAC coding tools. After first implementation we have carried out listening tests and found that such design was still unable to match "reference quality AAC" but with noticeable increase of audio quality when compared to MPEG reference software. However, first implementation was very slow and it was impossible to perform real-time encoding at any platform.

After debugging first implementation we began quality and speed optimizations on the following way:

- Optimization of psychoacoustic model
- Optimization of quantizer/bit allocator
- Optimization of other AAC coding tools
- Speed optimization and final debugging/bug fixing

AAC is defined as International Standard (IS). Our goal was to improve the quality of the compression process while maintaining full compatibility with the standard. Doing so will allow any AAC compatible decoder to decode bit streams from the improved encoder.

Fortunately, this is possible. ISO standard defines so-called "normative" part, which consists of bit stream syntax only. This means that we can modify each process of ISO AAC unless the modification yields changing the bit stream syntax. This paper will describe some of the steps carried out at PsyTEL Research that improved overall coding efficiency of the AAC codec.

Modifications to the Psychoacoustic Model

Psychoacoustic model represents core part of one perceptual coder. The purpose of psychoacoustic model is to estimate maximum allowed distortion, represented as SMR or ISMR (Signal-Mask-Ratio or Inverse-Signal-Mask-Ratio). Psychoacoustic model runs in *frequency domain.* It is possible to use output from codec filterbank as input for psychoacoustic model, or to perform separate transform-filtering for the purpose of psychoacoustic analysis.

ISO 13818-7 Annex-A psychoacoustic model is based on "Psychoacoustic Model II" that was also used in older coding algorithms (MPEG 1 Layers II and III, ISO/IEC 11172-3). This model uses separate FFT filterbank of two sizes for two block length (2048 samples for long blocks and 8*256 samples for short blocks). This model is rather simple, it is using hard coded psychoacoustic values (read from table) and simple convolution with cochlear, masker loudness independent, spreading function. Fixed threshold in quiet (also read from table) is used as minimum threshold of audibility.

While this approach will give good results compared to old MPEG Layer III (MP3) software it is still far from best possible quality expected from optimized AAC codec. We have carried out several important algorithmic modifications to the psychoacoustic model that were giving biggest quality improvement.

Modifications to the Input Filterbank

We have noticed that FFT filterbank used in basic model cannot always simulate energy values from codec MDCT filterbank. Because of this energy estimation might be incorrect and therefore psychoacoustic output would be inaccurate. Our improved coder is using MDCT for the psychoacoustic analysis, too. This way we are performing accurate estimate of the energy and also we are reducing complexity because MDCT output could also be used for the codec thus avoiding one transform. However, we also need complex values (imaginary part) of the transform in order to estimate tonality.

One of the possible solutions is using of CMDCT filterbank (Complex Modified Discrete Cosine Transform) that outputs MDCT as real part and MDST as imaginary part. CMDCT is also very good choice for LSI/DSP solutions because of reduced complexity when compared to ISO Model II approach (MDCT + FFT in parallel).

Other solution is found by using of FFT filterbank with modified window characteristics that are matching aliasing components of MDCT better.

Modification to the Absolute Threshold of Hearing

“Absolute Threshold in Quiet” values from the tables found in psychoacoustic model II description are inaccurate and fixed to specified loudness. We have noticed two problems when using this approach:

- Model is not sensitive to average loudness of one frame which leads to heavy “cut-off” effect
- Signal in HF range (>14.5 kHz) is aggressively masked and found annoying for some listeners

Our model uses fine-tuned formula with loudness approximation. Loudness approximation prevents undercoding in quiet conditions. Experiments carried out showed significant performance increase, especially with younger persons with good hearing abilities. Fine-tuned formula reduced distortion in high frequency (HF) range known as "flanging" artifact - very common to old MP3 technology. In addition to loudness approximation model also employs ATH formula correction. Correction is based on pre-defined ATH tables measured on subject with extraordinary listening abilities in higher frequency range.

Modifications to the Hard-coded Psychoacoustic Parameters

Psychoacoustic parameters found in ISO model are fixed and very simplified. By using these values several problems occurred in encoder design. When target bit rate was set to very small values (high compression ratio) psychoacoustic criteria cannot be met in any way. Bit allocator was not able to allocate bits in proper way. Similar effect occurs when high bit rate was required - the effect was "undercoding" - some frequency bands had distortion much less than allowed.

Perceptually best results would be met if average distortion (positive or negative) is equal for each frequency band. This is achieved by tuning each psychoacoustic parameter for each bit rate / coding profile and improving quantizer (described later). For this purpose we have designed graphic simulation tool and large base of people with good hearing abilities were involved in the listening tests.

We have fed simulation tool with large base of test samples (featuring "critical" samples like castanets, speech, applause and simulation samples like pure tones, pure noise, mixed tones with noise, etc…) and performed listening tests. Best combination of psychoacoustic parameters (found in listening tests and analyzer distortion results) is chosen for each coding profile.

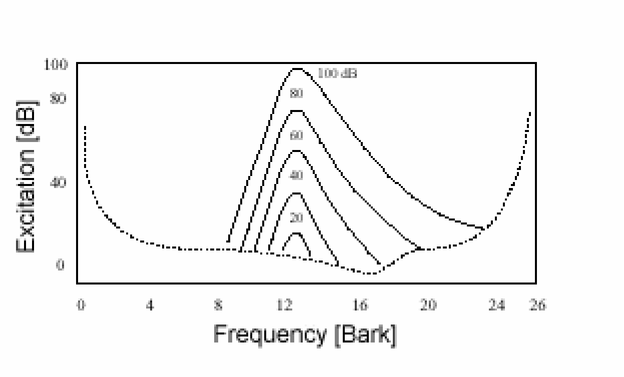
Modifications to the Tonality Detection Algorithm

Pure noise and pure tone have different masking abilities. The difference might be more than 20 dB as measured by [10]. Good tonality approximation is essential in order to simulate human psychoacoustic. ISO model uses prediction based tonality estimation. This model is based on prediction of the magnitude and phase of the FFT spectra. Experiments carried out proved some pitfalls of this approach, especially on speech signals. AAC filterbank is long and some deviations inside frames are omitted. We have improved tonality detection model by using *intra-frame* tonality estimation for long blocks. Intra-frame estimation uses SFM (Spectral Flatness Measure), or "peak detection" for the purpose of tonality detection. This detection is faster than measuring of unpredictability described in ISO model. Another safeguard measure was introduced by limiting unpredictability for some quality profiles.

For highest-quality encoding profiles we have designed “Improved Human Speech Coding” (IHSC) tonality detection algorithm with very conservative tonality estimation for human speech range. This module helps solving most of the problems found in human speech encoding in MPEG algorithms.

Modifications to the spreading function

Spreading function described in ISO model is rather simple. However, psychoacoustic experiments showed that sloppiness of the spreading function is dependent on the masker loudness. ISO model assumes constant loudness, which is incorrect. Using fixed spreading function might lead to over or under masking in some cases which reduces coder performance. This picture shows spreading function as measured in psychoacoustic experiments:



Our model takes masker loudness and mid-frequency into account and defines spreading function in the following way:

$$S_l = 27dB / Bark$$

$$S_u = \left[22 + \min\left(\frac{230}{f}, 10 \right) - 0.2 \cdot L \right]$$

Where Sl is the lower part, Su is the upper part, f is the mid-frequency of the masker and L is masker loudness in dB. This spreading function requires larger memory use, but experiments showed that sound quality improved.

Addition of Temporal Masking Estimation

ISO Psychoacoustic Model II does not take into account time-domain masking effects of human perception. Several experiments proved that time-domain psychoacoustic is very important, especially at high compression ratios.

Our model employs so-called “post-masking” effect since “pre-masking” phenomenon is very small and would lead to pre-echo distortion problems because of filterbank window length. Inclusion of post-masking estimation algorithm required small modifications to psychoacoustic model look-ahead and history buffers.

We have carried experiments that proved importance of time-domain masking, especially on tracks that contained “attacks” in energy (i.e. castanets or *fatboy* samples). On such samples bit rate demands were significantly reduced without any audible distortion. However, time domain masking didn’t have effect on steady state samples, as expected before experiments.

Modifications to the Block Switching Decision

In order to avoid "pre-echo" artifact, common to all transform domain compression algorithms, AAC standard defines two window sizes. Long window size (2048 samples) is used on steady-state signal conditions and provides high coding gain. For "attack" phases of the input signal shorter windows are used (256 samples). Decision when to switch to "short" mode is made in psychoacoustic model by analyzing signal properties from current and last block.

ISO psychoacoustic model defines simple method of switching by comparing the values of *perceptual entropy* (empirical bit allocation) from two consecutive long blocks. If significant change in perceptual entropy is detected, short block mode is triggered.

Experiments proved that ISO method completely fails on some critical samples and even triggers short block mode where not necessary (thus reducing codec performance). Because of these problems we have designed better method of attack detection:

Our model uses shorter length than one AAC block. Also, it works both in time and frequency domain. First, signal is high-passed because low frequency attacks could be coded with long window. Local perceptual entropies are measured in each subblock along with linear prediction in the time domain. Also, so-called "pre-masking" and "post-masking" effects are also taken into account before final decision is made. If the increase of local perceptual entropy is above certain threshold k *and* linear prediction error also exceeds some threshold *l* short block mode is triggered. Our model is successfully detecting attacks where ISO model clearly failed. Fine-tuning of decision parameters is the key issue of model success.

Modifications to the M/S stereo coding tool

M/S coding is the essential tool for medium and low bit rate coding. Stereo correlation is important issue in successful coding at popular compression ratios, like 11:1. ISO model suggests that M/S coding should be triggered if there is significant coding gain in each scalefactor band. However, this approach would lead to stereo-imaging problems as noticed in our experiments. Our model carries threshold difference detection prior to coding-gain estimation. If left and right masking thresholds differ too much, M/S is not used for that scalefactor band.

Modifications to the bit allocation / quantization loops

Quantizer is the tool equally important as perceptual model. Quantizer is responsible for allocating noise while maintaining psychoacoustic demands set by perceptual model. Quantizer in AAC is designed as two loop iteration where inner loop (rate loop) is maintaining bit rate and outer loop (distortion loop) is maintaining perceptual performance by readjusting bit-allocation in each scalefactor band.

$$ix(i) = \operatorname{sgn}(xr(i)) \cdot NINT\left[\left(\frac{|xr(i)|}{\sqrt[4]{2^{quantizer\_stepsize}}}\right)^{0.75} - 0.0946\right]$$

"quantizer_stepsize" is break apart on two values, one is global for entire block (global_gain) and one is local - for each scalefactor band. This way it is possible to readjust quantizer step-size for each scalefactor band.

Idea behind this approach is quite straightforward, but it is working only in general case. There are conditions where bit rate-distortion demands cannot be met (for example, at high compression ratios). In some extreme cases all scalefactor bands will be distorted. Default model of amplification, described in ISO model is amplification of all distorted scalefactors until either perceptual condition is met (distortion lower than allowed) or exit condition is met (all scalefactors amplified, or too many amplifications have been done). Unfortunately, experiments proved that in most cases ISO model fails with "exit case" instead of "perceptual case". Doing so lefts several problems:

- Non equal spreading of quantization noise (leads to annoying artifacts)
- Some frequency bands are non distorted while other have significant amount of distortion
- Poor performance and poor tandem coding ability

To improve quantizer loops we have carried out several modifications and improvements to the quantizer loop strategy. First, we have designed sophisticated *analysis-by-synthesis* method for estimating bit allocation prior to quantization loops. This method is based on NMR analysis and signal energy. By using this model we can estimate distortion after the quantization without performing loops. This model will adjust amplification strategy to one of three possible options:

- Amplification of all scalefactor bands (if model detects successful termination in perceptual terms)
- Amplification of worst distorted band per one loop (if perceptual conditions can't be met)
- Amplification of 50% of distorted bands in medium case

Along with tuning of perceptual parameters for each bit rate (as described earlier) this model allows perceptually best quantization even if perceptual demands can't be met. Final result is lower NMR (noise-mask-ratio) than non optimized, ISO model.

Adaptive bit reservoir

Bit rate in most applications is required to be constant. However signal properties are not constant over time. It is possible to expect accidental "peaks" in signal properties that require higher more bits. Also there are parts of the signal that don't require bit rate specified by user demands. For that purpose AAC standard defines so-called "bit reservoir" for storing or spending extra bits.

ISO model uses very rough method for allocating bits from reservoir, and experimental results show that bit reservoir is most of the time drained. To solve this issue, we have designed bit reservoir management module for allocating bits between frames and even channels. This model takes care of NMR history preventing accidental changes as much as possible. Our reservoir management uses look-ahead PCM buffers for estimation of the bit usage in the actual frames near future.

Variable Bit rate

As the final step towards "transparent" coding we have designed variable bit rate coding module. In this coding mode coder is free to allocate as much bits as required by psychoacoustic model. The result could be described as "constant quality - variable rate" instead of default "constant rate - variable quality" as defined in default coding mode proposed by ISO. For broad acceptance we had to take care of various quality profiles. For some uses it is sufficient to provide "near-transparent" quality grade while some other applications require "higher than transparent" quality grade like tandem-coding applications, for example. Our AAC codec has 8 variable bit rate coding modes.

Performance Optimizations

All described quality optimizations were followed with significant complexity increase. Final encoder had to be optimized in order to be useful on today's PC Multimedia platforms. We have carefully optimized each AAC coding tool without sacrificing quality.

Most demanding functions (quantization and huffman bit counter) were separately optimized with significant algorithmic changes. Quantizer was optimized in a way that it now requires only one multiplication per outer loop. Huffman bit counter was rewritten and now it performs full greedy-merge sectioning only after the last outer-loop. Unused bits are saved in reservoir so they can be available for the next frames.

In addition to algorithmic optimizations we have ported all vector algebra and signal processing functions to Intel(R) Signal Processing Library. This way encoder uses fastest possible CPU instructions available today. Codec is capable of reaching real-time encoding even on low-end hardware.

Conclusions

By testing Improved ISO AAC coder with critical worst case signals we have showed significant performance increase without altering compatibility with ISO 13818-7 bit stream. Each improvement in coding tools was rewarded with higher coding performance with lower NMR. Experiments also showed that further quality improvements are to be expected by additional work on perceptual model and other AAC coding tools.

AAC proved to be state-of-the-art audio codec capable of providing "transparent" CD Quality at lowest possible data rate as well as good quality at high compression ratios. When compared to other MPEG codecs AAC has highest flexibility available today.

Acknowledgements

Author would like to thank Andree Buschmann for the very helpful suggestions on AAC and psychoacoustics, Juha Laaksonheimo, Adam Fox and Aleksandar Dovnikovic for helping in listening tests and tuning AAC algorithm parameters and all colleagues at Mathematics Faculty of Belgrade for support. Special thanks go to Menno Bakker for developing high performance stable ISO AAC decoder.

References

[1] M. Bosi, K. Brandenburg, Sch. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Yoshiaki Oikawa. ISO/IEC MPEG-2 Advanced Audio Coding. In *Proc. of the 101st AES-Convention*, 1996. Preprint 4382.

[2] ISO/IEC JTC1/SC29/WG11 MPEG. International Standard IS 13818-7 Information Technology - Generic Coding of Moving Pictures and Associated Audio, Part 7: Advanced Audio Coding, 1997.

[3] J. Herre and J. D. Johnston, “Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)”, presented ”, presented at the 101st AES Convention, Los Angeles 1996, preprint 4384

[4] J. D. Johnston, J. Herre, M. Davis, U. Gbur, “MPEG-2 NBC Audio - Stereo and Multichannel Coding Methods”, presented at the 101st AES Convention, Los Angeles 1996, preprint 4383

[5] K. Brandenburg, “MP3 and AAC Explained”

[6] J. Herre, K. Brandenburg, E. Eberlein and B. Grill, “Second Generation ISO/MPEG-Audio Layer III Coding ”, presented at the 98th AES Convention, Paris 1995, preprint 3939

[7] J. D. Johnston. ”Estimation of Perceptual Entropy Using Noise Masking Criteria”. ICASSP 1988, pp. 2524–2527.

[8] F. Baumgarte, "Application of a physiological ear model to irrelevance reduction in audio coding," Proc. AES 17th Int. Conf. on High Quality Audio Coding, Signa, Italien, Sept. 1999.

[9] D. Knuth: “The Art of Computer Programming, 2nd cd., vol. 2, Addison-Weseley, Reading, MA, 1981

[10] R. P. Hellman. Asymmetry of masking between noise and tone. Perception and Psy-choacoustics, 11(3):241–246, 1972.

[11] A. Ferreira. Tonality Detection in Perceptual Coding of Audio. 98th Convention of the Audio Engineering Society, February 1995. Preprint n. 3947.

[12] J. D. Johnston and A. Ferreira, Sum-Difference Stereo Transform Coding, Proc. IEEE ICASSP (1992) p 569-571.