ADVANCED SOFTWARE IMPLEMENTATION OF MPEG-4 AAC AUDIO ENCODER

Danijel Domazet, Mario Kovac

Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia
dand@rasip.fer.hr, mario.kovac@fer.hr

Abstract: *MPEG-4 AAC audio encoder is developed. AAC main and AAC Low Complexity object types, defined as part of Main profile in General Audio part of ISO/IEC 14496-3 (MPEG-4) standard, are implemented. Encoder targets high-quality, wideband, complex audio. Encoder was carefully designed to respect present accomplishments in audio coding (standardized in MPEG-4). At the same time, several implementation novelties were introduced. Encoder achieves 'perceptible but not annoying' 44.1 kHz audio quality at bitrate 64 kb/s/ch, while satisfying quality is accomplished at 48 kb/s/ch.*

Key words: *Audio coding, MPEG-4, AAC*

1. INTRODUCTION

Perceptual audio encoder is implemented according to General Audio part of ISO/IEC 14496-3 (MPEG-4) standard. Usual perceptual audio encoder structure is followed: MDCT filterbank, psychoacoustic module, spectral processing tools, quantization module, lossless compression. MPEG-4 GA *tools* implemented include: Mid/Side coding, Intensity Stereo coding, Perceptual Noise Substitution and Temporal Noise Shaping.

2. ENCODER IMPLEMENTATION

2.1. Psychoacoustic module

Psychoacoustic module is based on the so called Psychoacoustic model II. The model has undergone significant changes.

Tonality detection

Tonality detection algorithm uses intensity and phase prediction to estimates the tonality of each frequency component. Algorithm assumes tonal signal as a clean sinusoid with constant amplitude and linear phase. Accordingly, amplitude of the current block is predicted to be the same as the amplitude of the previous block and phase is calculated as a linear interpolation from two previous blocks. After the predicted values have been calculated, a standard way for tonality calculation is carried out.

Spreading function

Spreading function implemented has the following shape:

$$s_{dB} = \begin{cases} 27 \times (i-l), & i \le l \\ \left[-24 - \dfrac{230}{f_c(l)} + 0.2 \times 10 \log_{10}(E) \right] \times (i-l), & i > l \end{cases} \quad (1)$$

where i and l are Bark values of masker and target bands respectively, E is the energy of frequency components inside masker band, and f is the masker band central frequency. Energy is introduced into the spreading function because of the increase in masking (approx. 0.2 dB/Bark) due to the increase in signal energy [5].

Transient detection

Transient detection algorithm is based on the energy distribution analysis and perceptual entropy (PE) calculation [1]. During transients, high frequency signal components energy is significantly increased. Algorithm calculates signal energy for components above 6 kHz. If current block energy increases significantly compared to last block analyzed, and perceptual entropy of the signal is above some predefined threshold, signal is marked as transient. Algorithm proved to be very effective.

Short window grouping

Since there are 8 short blocks in one standard block, and in order to minimize side information which goes along each block, short block grouping is implemented. Grouped blocks share same scalefactors and other side information, thus saving in an overall block bitcount.
Short block grouping is implemented in the following way: signal energy is calculated for each short block. If two neighboring blocks energies do not defer by more than threshold k (typ. 0.7 – 1.3), blocks are joined to form an individual block group.
Signals encoded without short block grouping, showed that, even 'pre-echo' effect was successfully eliminated, the side information required for each short block was too excessive and directly influenced the final quality of the encoded signal. Short block grouping significantly reduced this effect.

2.2. Filterbank

Block length (long or short) and window shape (Sine or Kaiser-Bessel Derived) completely define filterbank operation.
Block length decision is supplied from psychoacoustic module and is directly controlled by transient detection algorithm – if transient is detected short block is used, otherwise long block is selected.
Window shape choice is based on tonal frequency components spacing. Psychoacoustic module supplies information on each frequency component's tonality. Strong tonal components are isolated and decimated inside 70 Hz range, than their mutual frequency

distance is analyzed. If the distance is higher than 220 Hz, KBD window is used (better stop-band attenuation, compacts more energy into a single component), otherwise, sine window is selected (better pass-band frequency selectivity).

2.3. Quantization

Quantization is implemented in a standard inner-outer loop method. Inner loop adjusts scalefactors in order to keep the quantization noise below masking threshold, while the outer loop checks the overall block bitcount and, if bitcount exceeds required bitrate, rises quantization scalefactors for all quantization bands.
Quantization is the most critical part for encoding speed due to its iterative nature. Time spent on the quantization process is reduced to some extent by increasing the outer loop scalefactor stepsize (outer loop scalefactor stepsize of 4 did not noticeably decrease encoding quality at bitrates 48 kb/s/ch and above). Also, instead of only one, inner loop adjusts scalefactors for a group of bands with worst distortion in a singe loop, thus decreasing overall number of loops.

2.4. MPEG-4 GA tools implementation

Perceptual Noise Substitution

Main challenge of the PNS tool is detection of tonal and noise signal components. Since implemented tonality calculation algorithm does not clearly separate tonal from
noise signals (it only calculates tonality index), it is necessary to determine the limit under which signals would be considered noise, and upon which PNS tool would then be applied.
A very common method is to take a fixed tonality index limit and consider all frequency components under this limit to be noise. This method proved to be very doubtful in many cases, especially for transient signals (Fig. 1.), since nearly all signal components showed low tonality index and were automatically considered noise. In those cases PNS did indeed reduce the quality of the encoded signal.
A novel approach is implemented. Components are sorted according to their tonality, and only certain number of the *noisiest* components (which still lay under some threshold) is sent to PNS tool. This way, encoder is unable to overuse the PNS tool. Also, if signal entropy is low PNS is not used at all.

Mid/Side

A relatively conservative way of switching between Left/Right and Mid/Side (MS) encoding is implemented. Each scalefactor band is analyzed in the following way: intensity of two corresponding scalefactor band components is compared and MS encoding is permitted if the difference is under a defined limit L. Since MS encoding can be used only on a scalefactor band basis, it is necessary to decide about MS encoding for a group of individual components. MS encoding is used only if all components in a scalefactor band are marked as MS. This way MS tool can be used with no concern that it will be switched on for signals that would not benefit from the tool (Fig. 2.).

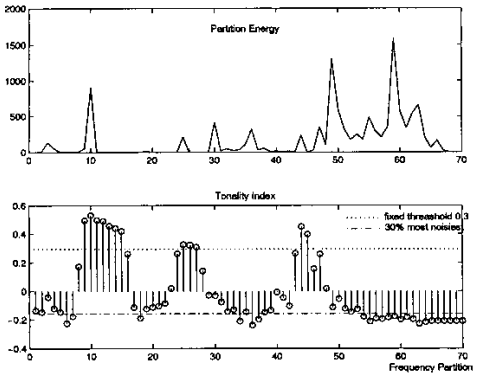

Fig. 1. Limiting noise substitution: Perceptual Noise Substitution tool is applied only to *noisiest* bands. Fixed threshold would affect too many bands and degrade the quality of the encoded signal.

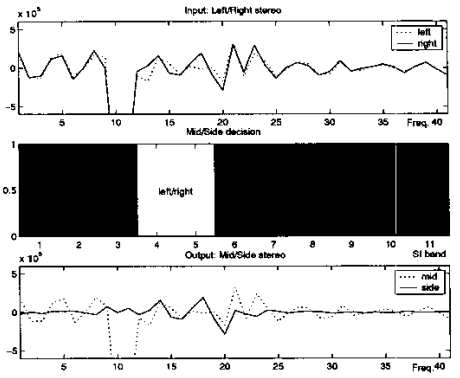

Fig. 2. Mid/Side coding is applied only in bands where left and right channel components are close to one another. Resulting Side channel is around zero, which guarantees coding gain.

Intensity Stereo

Intensity Stereo (IS) tool has been implemented according to [4] and [9]. IS is used only in high frequency range, above 6-7 kHz. However, IS has been disabled for parts of the signal with high tonality index due to highly perceptible distortions for tonal music. Tonality index threshold for IS tool has been experimentally determined to be between 0.25 and 0.35.

3. RESULTS

3.1. Encoded signal quality

Subjective tests have been held with *naive* listeners. Special attention has been placed on bitrates 48 kb/s/ch (48000 bits per second per channel) and 64 kb/s/ch, since main goal of the encoder implementation was achieving same quality of complex, wideband audio, at bitrate 48 kb/s/ch, as present day *mp3* audio encoders do achieve at 64 kb/s/ch.
Primary goal of the test was to identify medium and large impairments, and to differentiate the way each MPEG-4 GA audio *tool* affects signal quality. Accordingly, test sequences used were not critical, but rather ordinary 44.1 kHz stereo material consisting of speech only, music only, and music with speech. Tested bitrates were 64, 96, 128 and 160 kb/s stereo. Results on an ITU-T 1 to 5 scale are given in Table I.

Following encoder tool configuration is found to be the most appropriate for majority of signals tested:

- TNS above 3 kHz, only long blocks;
- MS whole bandwidth;

- PNS on lower bitrates (64 and 96), only when signal entropy is high, starting frequency should be above 4 kHz, and no more than 25 to 35% of noisiest frequency components should be caught;
- IS tool above 7 kHz, exclude strong tonal components.

Table I: Subjective test results summary

Bitrate [kb/s stereo]	64	96	128	160	Comment
Speech	1.9 - 2.5	3.5 - 4.1	4.0 - 4.2	4.5 - 4.8	PNS tool significantly degrades signal quality at high bitrates. Female speech harder to encode.
Music	1.5 - 2.8	3.2 - 4.2	3.8 - 4.4	4.2 - 4.8	PNS should only be used if signal entropy is high, clean harmonic music is degraded by PNS. IS above 7 kHz.
Music with speech	1.5 - 2.8	3.0 - 4.2	3.5 - 4.3	4.0 - 4.8	Channels often uncorrelated: IS destroys stereo image, M/S ineffective.

3.2. Encoding speed

Encoding speed achieved on AMD Athlon™ XP +1600 processor (1400 MHz), with encoder configuration stated above, was 2 (expressed as ratio between signal and encoding duration), 10% more or less depending on signal characteristics.

4. CONCLUSION

Even though encoder development is still in process, encoder manages to achieve 'perceptible but not annoying' quality at bitrate 64 kb/s/ch. Bitrates as low as 48 kb/s/ch still produce audio quality that completely satisfies average *PC-internet-diskman* users and should therefore comfortably replace present day *mp3* encoders.

Audio encoder implementation confirmed that MPEG-4 AAC is the 'top' audio coding technology at the moment. Variety of AAC tools allow fine tuning for different signal types, and leave enough space for each encoder implementation to be unique.

REFERENCES

[1] J.D. Johnston, Estimation of Perceptual Entropy Using Noise Masking Criteria, *Proceedings Int. Conf. Acoust., Speech and Signal Proc,* 1988, 2524-2527
[2] D. Schulz, Improving Audio Codecs by Noise Substitution, *Journal of AES,* Volume 44, Number 7/8, 1996, 593-598

[3] J. Herre, D. Johnston, Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS), *101st AES Convention,* Los Angeles, 1996
[4] J. Herre, K. Brandenburg, D. Lederer, Intensity Stereo Coding, *96th AES Convention,* Amsterdam, 1994
[5] T. Jelakovic, Zvuk, Sluh, Arhitektonska akustika, *Skolska knjiga, Zagreb,* 1978
[6] B.W. Kernighan, D.M. Ritchie, The C Programming Language, *Prentice Hall*, 1988
[7] J. Makhoul, Linear Prediction: A Tutorial Overview, *Proceedings of the IEEE,* Vol. 63, No. 4, 1975
[8] G. Stoll, F. Kozamernik, EBU Listening Tests on Internet Audio Codecs, *EBU Technical Review*, 2000
[9] ISO/IEC JTC1/SC29/WG11, International Standard IS 14496-3 Information Technology – Coding of audio-visual objects, Part 3: Audio, 1999