



Audio Engineering Society Convention Paper 5476

Presented at the 111th Convention
2001 September 21–24 New York, NY, USA

This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Beyond CD-Quality: Advanced Audio Coding (AAC) for High Resolution Audio with 24 bit Resolution and 96 kHz Sampling Frequency

C. Burgel, R. Bartholomaus, W. Fiesel, J. Hilpert, A. Holzer, K. Linzmeier, M. Weishart
Fraunhofer Institute for Integrated Circuits IIS-A
Erlangen, D-91058, Germany
bgl@iis.fhg.de

Contrary to the MPEG-1 Audio compression schemes, Advanced Audio Coding (AAC) in its MPEG-2 and MPEG-4 flavours has no inherent upper limit for the sampling frequency of the input signal. Furthermore, the bitstream format allows to cover a dynamic range far beyond the range provided by 24 bit linear PCM coding. This makes the AAC coder an ideal candidate for representing audio signals with parameters that are usually associated with high resolution audio systems. This paper discusses the application of this highly efficient compression scheme to digital programme material represented with 24 bit and 96 kHz.

INTRODUCTION

The term "High Resolution Audio" usually specifies digital audio with higher amplitude and time resolution than the ubiquitous Compact Disc audio standard, i.e. with 16 bit and 44.1 kHz. While CD quality has been considered sufficient for home audio use in the last decades, many articles have more recently been published to re-

port audible differences between 44.1 kHz and higher sampled systems. Some of them point out anti-aliasing filters [4, 5] and non-linearities of the equipment used for the performed listening tests [6] as being responsible for these effects, while there are others claiming that the frequency range audible by humans extends over more than 20 kHz. The Acoustic Renaissance for Audio (ARA), a private committee exploring future possibilities of audio

applications, suggests in a proposal that a carrier, intended to convey everything a human listener can hear, should cover a frequency range of 25 kHz as 1 percent of young adults would be able to detect 25 kHz tones [7]. Furthermore, they suggest an amplitude resolution of at least 20 bit. A commonly used combination covering these demands for high quality audio consists of 24 bit word length and 96 kHz sampling frequency.

Apart from these articles related to actually audible differences, there are also other aspects of high resolution audio to be taken into account. One of these is the advantage of archiving audio in best possible quality which means the quality the material is produced in [3]. As audio material is often mixed and digitally processed in its production phase, it is inevitable to use a sufficiently higher quality for production than the one intended for the final release. This is the reason why high resolution audio is quite popular in the professional domain. There are many companies in the music and hardware industry which already offer a variety of different products, i.e.:

- mixing consoles capable of handling 24 bit and 96 kHz signals
- audio recording DSP cards for PCI with 24 bit and 96 kHz
- multi effect audio processors
- high quality 24 bit A/D and D/A converters
- 96 kHz audio transmitters and receivers and forth.

As the future demands on audio material intended for publishing are not yet known, it is advisable to preserve the recorded material in the best possible quality without the degradations introduced by the conversion to the final release format.

In the past few years, the demand for high quality audio systems has been migrating from professional studio applications to the consumer market. With the forthcoming availability of inexpensive media offering the required capacity and corresponding hardware devices, like audio transceivers and D/A converters, the traditional audio CD format is about to lose its status as a reference for home audio equipment. New formats, such as DVD-Audio and the Super Audio CD, are being introduced as the new standards.

While media to distribute high resolution audio already exist, the problem of the significantly increased amount of data required to represent music at this high quality remains, for example, in the fields of transmission and mass storage. In the field of Electronic Music Distribution (EMD) in particular, the interest in a compression scheme that preserves the perceptual quality is growing

steadily. Defined within the ISO/MPEG-2 audio standard [1] and extended within MPEG-4 [2], the Advanced Audio Coding (AAC) scheme includes the capabilities for representing audio at 24 bit/96 kHz and would therefore be a suitable answer to this demand.

COMPARISON OF EXISTING HIGH RESOLUTION CODING METHODS

Most of the available popular schemes for coding high resolution audio are not designed to achieve a maximum compression ratio. Specifically, the DVD technology defines a new media storage capacity margin where an audio encoder should satisfy the consumer's interest in delivering two to eight channels of audio, sometimes together with a high quality video programme on a single disk.

The Super Audio CD (SACD) uses Direct Stream Digital (DSD), a coding scheme that more or less directly represents the input format of sigma-delta digital to analog converters [8]. An optional lossless coding stage using linear prediction and arithmetic coding may reduce the raw data rate of about 2.8 Mbps per channel by a factor of approx. two. Thus a SACD may contain 74 minutes of audio material which is represented multiple times on the disk. The compatibility layer holds 74 minutes of Red Book CD stereo audio. The high density layer includes 74 minutes of both stereo DSD *and* multichannel DSD audio.

Tailored for the DVD-Audio, the lossless compression scheme MLP (Meridian Lossless Packing) is a format for encoding up to 63 audio channels of 24 bit material with up to 192 kHz sampling rate [9]. The maximum data rate of 9.6 Mbps provided by the DVD-Audio forms the upper limit for 6 channel audio, when coded with 24 bit and 96 kHz. A stereo downmix can be derived directly from the multichannel bitstream without having the need to store an additional stereo stream. Since the stereo compatible downmix signal is already prepared in the encoder, the 2 channel data can be stored with only a small data rate overhead inside the 6-channel bitstream.

A popular lossy coding scheme for high resolution audio is the Coherent Acoustics Coder from Digital Theater Systems (DTS) [10]. Designed as a 32 band subband coder, psychoacoustic masking effects are exploited to control the quantization of the subband samples. Up to 10.1 audio-channels can be transmitted with a maximum sampling frequency of 192 kHz and 24 bit input data. The frequency region above 24 kHz can be added to a standard 48 kHz sampling rate coder as an extension bitstream in a scalable way [11, 12]. The typical bitrates are adopted to the storage media. The standard data rate of CD audio (1.411 Mbps) can be used to carry a 24

bit/192 kHz stereo DTS stream or a 44.1 kHz 5.1 DTS-Stream [14]. In [11] a range from 96 to 256 kbps/channel is noted as recommendation for coding 24 bit/96 kHz audio. Nevertheless, for storage on a DVD, the range from 1.5 to 3 Mbps for 5.1 channel audio are the figures that can be commonly found in the publications [11, 12].

The following table depicts the comparison of the relations in bitrate reduction resulted by extrapolating the typical data rates of the above coding schemes to a value for 2-channel coding at 24 bit and 96 kHz. The values for the Super Audio CD are included for comparison, even if the terms “word length” and sampling frequency used for linear PCM get a different meaning when applied to a 1 bit oversampled bitstream.

encoding scheme	data rate (kbps)
SACD	~ 2822...5644
MLP	~ 2600...3000
DTS	512...1024

Of course, there is an ongoing increase of transmission bandwidth on the Internet coming along with high speed access over satellite, cable and DSL modems for the end user. But even if the above mentioned coding schemes are adjusted perfectly for the capacities of the DVD-type storage media, the field of electronic music distribution cannot be covered by such algorithms due to the enormous amounts of data that would have to be transmitted. The following results shall show how AAC can extend the application range of high resolution audio.

HIGH RESOLUTION AAC IN THEORY

AAC is a perceptual transform coder of which frequency and time resolution are determined by the analysis/synthesis filterbank and the quantization module. The following two paragraphs discuss its feasibility to support high resolution audio.

Sample Rate Support

Typically, AAC uses a Modified Discrete Cosine Transform (MDCT) with a fixed frame length of 1024 for long and 128 for short blocks respectively.

An increased sample rate leads to a better resolution in the time domain which in turn may lead to an improved coder efficiency for transient signals as fewer short blocks will be needed for the encoding process. Since short blocks produce more overhead than long blocks, more space for the actual audio data is left. On the other hand, the increased time resolution means a reduced frequency resolution. This may result in a reduction of the coder efficiency as each scalefactor band covers a wider frequency range. The overall quality difference can only be evaluated by listening tests.

AAC has explicit support for the following set of sampling frequencies which are given in the right column of the following table [2]:

frequency range (Hz)	sampling frequency f_s (Hz)
$f \geq 92017$	96000
$92017 > f \geq 75132$	88200
$75132 > f \geq 55426$	64000
$55426 > f \geq 46009$	48000
$46009 > f \geq 37566$	44100
$37566 > f \geq 27713$	32000
$27713 > f \geq 23004$	24000
$23004 > f \geq 18783$	22050
$18783 > f \geq 13856$	16000
$13856 > f \geq 11502$	12000
$11502 > f \geq 9391$	11025
$9391 > f$	8000

For configurations using a value of f_s lying in one of the intervals given in the left column, frequency dependent definitions are used to refer to the closest corresponding match. Thus, there is no upper limit in AAC’s support of sampling frequencies and AAC can be used with PCM material at 96 kHz and beyond.

Dynamic Range Coverage

After Huffman decoding and inverse quantization the spectral coefficients have to be rescaled before they are fed into the Inverse Modified Discrete Cosine Transform (IMDCT). The dynamic resolution is given by the largest possible interval which can be represented.

The smallest absolute quantized value is zero and the largest quantized value is limited to $\max(x_{\text{quant}}) = 8191$ which can be obtained using the escape Huffman codebook. The process of the inverse quantization is then given as :

$$x_{\text{invquant}} = \text{sign}(x_{\text{quant}}) \cdot |x_{\text{quant}}|^{\frac{4}{3}}$$

Afterwards the inverse quantized spectral values are rescaled using the equation

$$x_{\text{rescal}} = x_{\text{invquant}} \cdot 2^{\frac{1}{4}(\text{sf}-100)}$$

where the scalefactors sf are limited to the interval $[0..256]$ [2].

This allows to cover a dynamic range from $[0..9 \cdot 10^{16}]$ which corresponds to 339 dB. This exceeds the required range of 144 dB given by 24 bit PCM material by far.

LISTENING TEST RESULTS

In order to assess the subjective sound quality of signals encoded and decoded at 24 bit resolution and 96 kHz sampling frequency, listening tests in the style of the ITU test specification BS.1116 have been performed. The listener has to mark the original and the coded signal that are given in a random order according to a perceived loss of quality. Any number smaller than 5 will constitute a perceived loss in quality to the supposed original signal that has to be marked with a 5.

The PC based Fraunhofer professional AAC MPEG-2 Low Complexity Profile encoder has been used in stereo mode for generating the test bitstreams.

Three different bandwidths have been used for encoding:

- 20 kHz, the commonly known upper limit of the human perception.
- 25 kHz, the limit according to the ARA recommendation [7].
- 40 kHz as an upper limit for more or less full bandwidth of the input signal.

To exploit the scalefactor bands of AAC at 96 kHz, these bandwidth settings were slightly increased to 21 kHz, 27 kHz and 42 kHz respectively.

The choice of bitrates has been performed to reach almost transparent quality at these bandwidths.

- Setting 1: 160 kbps for 21 kHz
- Setting 2: 192 kbps for 27 kHz
- Setting 3: 256 kbps for 42 kHz
- Setting 4: 160 kbps for 42 kHz.

Setting 4 has been included as a lower anchor. All these bitrates should be read as total bitrates for stereo.

In the listening test nine different test items were presented to the listeners. Six of them contain the sound of single instruments (cymbal, castagnets, guitar, rattle, metal rings and triangle), one contains female Chinese speech, one applause and one a cantle of a track of popular music ("Isn't she lovely" by Livingston Taylor). The latter has been grabbed from an audio DVD, the others were an excerpt of items recorded at Fraunhofer Institute for Integrated Circuits. The equipment used for this purpose was capable of producing high quality 24 bit/96 khz recordings: two SCHOEPS CMC5 capacitor microphones, two RIM preamplifiers and an RME DIGI96/8 PAD digital audio card for A/D conversion and recording.

Figure 1 shows a spectral plot of part of the triangle item from which can be seen that indeed there is relevant

signal energy in the frequency range between 20 kHz and 40 kHz. As can be seen from Figure 2, most parts of the upper frequency spectrum are kept unchanged when the signal is coded at a bitrate of 256 kbps.

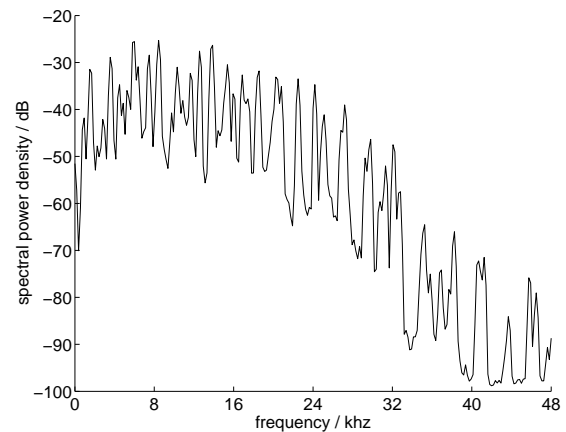


Figure 1: Spectrum of part of the triangle item

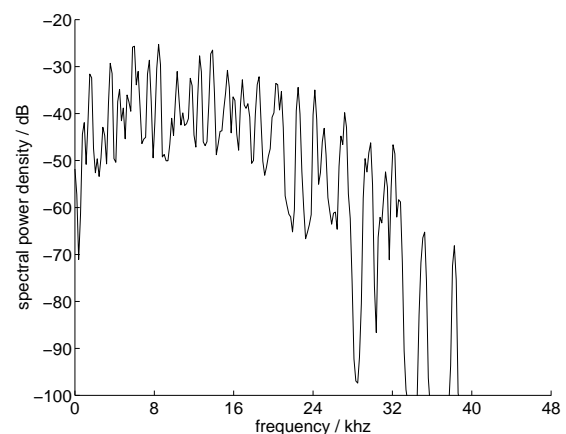


Figure 2: Spectrum of triangle item after coding with 256 kbps

To take the peculiarities into account that come along with the loudspeaker reproduction of high frequency components, the testsignals were presented to one half of the listeners via a pair of loudspeakers by Geithain. The remaining 5 subjects used high quality STAX SR Lambda Pro headphones.

As shown in figure 3 and the corresponding tables at the end of this paragraph, no statistically significant difference could be detected for settings 1, 2 and 3. Only at 160 kbps/ 42 kHz, the items 'cymbal', 'applause', 'guitar' and 'triangle' showed a degradation with a confidence in-

terval not crossing the zero line.

All of these results tend to be consistent for both, head-phone and loudspeaker reproduction.

At least for the available number of test items and listeners, the AAC encoder proved to be transparent for the first 3 combinations of bitrate and bandwidth. A further reduction of the necessary bitrate might be possible when using additional tools for AAC standardized in MPEG-4 Audio.

The following tables summarize the listening test results for the settings 1 to 4:

Setting 1	mean	95% conf.	upper	lower
cymbal	0.095	0.137	0.232	-0.042
speech	0.100	0.143	0.243	-0.043
castagnets	0.010	0.159	0.169	-0.149
rattle	-0.015	0.173	0.158	-0.188
applause	0.125	0.126	0.251	-0.001
livingston	-0.055	0.115	0.060	-0.170
metal rings	0.050	0.194	0.244	-0.144
guitar	0.000	0.050	0.050	-0.050
triangle	0.080	0.261	0.341	-0.181
overall mean:	0.043			

Setting 2	mean	95% conf.	upper	lower
cymbal	-0.045	0.117	0.072	-0.162
speech	0.070	0.123	0.193	-0.053
castagnets	-0.025	0.085	0.060	-0.110
rattle	0.000	0.052	0.052	-0.052
applause	-0.015	0.124	0.109	-0.139
livingston	0.050	0.050	0.100	0.000
metal rings	0.065	0.118	0.183	-0.053
guitar	0.035	0.117	0.152	-0.082
triangle	0.130	0.228	0.358	-0.098
overall mean:	0.029			

Setting 3	mean	95% conf.	upper	lower
cymbal	0.035	0.044	0.079	-0.009
speech	-0.015	0.071	0.056	-0.086
castagnets	-0.090	0.115	0.025	-0.205
rattle	0.060	0.065	0.125	-0.005
applause	0.065	0.085	0.150	-0.020
livingston	-0.030	0.061	0.031	-0.091
metal rings	0.080	0.111	0.191	-0.031
guitar	0.015	0.073	0.088	-0.058
triangle	0.050	0.069	0.119	-0.019
overall mean:	0.019			

Setting 4	mean	95% conf.	upper	lower
cymbal	0.250	0.202	0.452	0.048
speech	0.080	0.132	0.212	-0.052
castagnets	-0.080	0.109	0.029	-0.189
rattle	0.060	0.128	0.188	-0.068
applause	0.430	0.301	0.731	0.129
livingston	0.100	0.139	0.239	-0.039
metal rings	0.040	0.143	0.183	-0.103
guitar	0.075	0.068	0.143	0.007
triangle	0.200	0.180	0.380	0.020
overall mean:	0.128			

IMPLEMENTATION CONSIDERATIONS

Real world implementations of 24 bit/96 kHz systems face two major challenges:

- The 24 bit dynamic range demands for a careful design of the data path in encoder and decoder systems.
- The increased sampling rate of high resolution audio leads to a higher processor load on real time systems.

The following discussion will investigate this situation in greater detail.

Number Representation

It is important to distinguish between dynamic range and precision. In the encoder a high precision is mainly required for the time to frequency (t/f) mapping. If true 24 bit audio resolution is demanded, the filterbank has to fulfill this requirement. The AAC algorithm performs the t/f mapping by means of an MDCT with a maximum resolution of 1024 spectral lines. A typical fast and efficient implementation can be realized by a 10 stage radix 2 butterfly operation. Here each butterfly stage reduces the available precision by 0.5 bit in average. Therefore, this implementation loses about 5 bits precision inside the MDCT transformation. In case of a 24 bit data representation, a final precision of only 19 bit can be achieved after the t/f mapping. For most of the remaining parts of the encoder it is in general sufficient to be able to handle the dynamic range of 24 bits. This is due to the fact that even for worst case signals within one bark band a SNR value of approx. 24 to 30 dB leads to a perceptually unimpaired signal representation. To be on the save side the AAC quantizer can handle up to 13 bits precision within one scale factor band. However, a sufficiently large dynamic range has to be available to encode level differences between the bark bands. On the other hand in the decoder, in order to fulfill the MPEG conformance criteria, a full 24 bit precision has to be maintained throughout the decoding process, if 24 bit conformance is to be claimed.

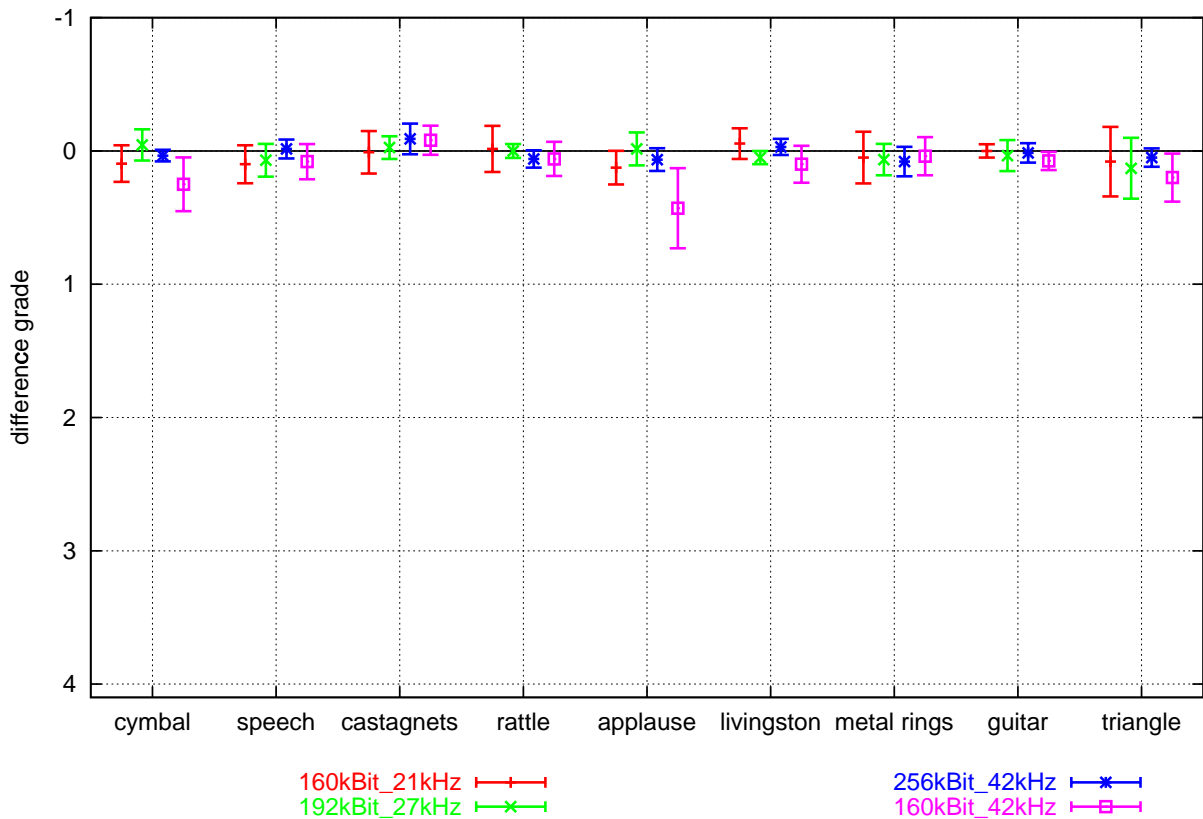


Figure 3: Listening test results

This imposes certain requirements as to which of the usually available number formats are suitable:

- Fixed-point representations

This format interprets the binary number in a linear data format. Typically implemented binary lengths are 16, 24 and 32 bits. Especially for digital audio signal processing, a data length of 24 bit is very popular. In order to avoid data overflow during arithmetic operations and to limit the accuracy loss, a careful scaling of data is necessary. However, in general some precision loss can usually not be avoided, making 24 bit processing unsuitable for 24 bit accuracy. Therefore, a number extension towards more bits must be used. The combination of two 24 bit numbers into one 48 bit integer seems to be a practical implementation solution[13].

- IEEE 752 Floating-point representation

Due to the automatic scaling feature the floating-point format is very suitable for handling a wide

dynamic range. The IEEE 752 floating-point standard offers two different formats, called *single precision* and *double precision*. The 32 bit single precision format allows 24 bits of mantissa (including the sign bit) and an 8 bit exponent for the scaling factor. Most floating-point DSP devices employ the 32 bit floating-point format. Although this format provides sufficient dynamic range, the precision is usually not good enough. Especially for final 24 bit precision, the mantissa does not offer sufficient resolution for the arithmetic operations.

On the other hand the IEEE 64 bit double precision floating point representation with eleven bits exponent and 52 bits for the mantissa provides ample headroom in both dynamic range and precision. However, this format is not generically available on DSPs today.

Neither the 24 bit fixed point nor the 32 bit floating-point format fulfills the demands of a 24 bit AAC decoder

conformance test condition. In this case, the number format inside the audio signal path has to be increased to 64 bit double precision or a suitable integer format.

Real Time Test Setup

In order to show that real time encoding is still possible on standard hardware, a DSP based real time chain was implemented. Due to the different number representation on DSPs special issues had to be addressed.

For testing the discussed ideas, a real time encoding/decoding chain was implemented, processing a digital 24 bit/96 kHz PCM signal. The signal is sent into the digital input of a Texas Instruments evaluation module (TMS320C6711 DSK), running a stereo MPEG-2 AAC encoder. The most important key features of the module are:

- TMS320C6711 DSP with 150MHz CPU Clock and external memory interface at 100 MHz
- 16 Mbytes of 100 MHz synchronous dynamic random access memory (SDRAM)

The encoded digital bitstream output is then transferred to the input of a Motorola DSP56362EVM evaluation module. The features of the Motorola module are:

- 24 bit DSP56362 Digital Signal Processor operating at 80-100MHz
- 128k x 24-bits of external SRAM and 128k x 8-bits of EPROM
- One 20 bit stereo Analog-to-Digital converter (ADC), four 24-bit stereo Digital-to-Analog converters (DACs)
- RCA jacks for all analog audio input and output connections
- Optical and transformer-isolated electrical S/PDIF stereo digital audio inputs and outputs.

Running the implementation of an MPEG-2 AAC decoder on the Motorola 56362 DSP again produces a digital 96 kHz stereo PCM output signal. The encoding and decoding setup is depicted in Figure 4.

For the AAC stereo real time encoder at 96 kHz sampling frequency on the TMS320C6711 DSP, an overall workload of about 80 percent is needed.

The total memory requirements for the encoder implementation are about 220 kbyte of SDRAM.

The recommended memory specification for the MPEG-2 AAC (ISO/IEC IS 13818-7) decoder implementation on the Motorola DSP56362 is (words in 24 bit):

- 13.5k words of ROM (program)
- 8.9k words of working memory
- 8.3k words of coefficient ROM

The computational requirements for a stereo decoder running with 96 kHz are about 30 MIPS (million instructions per second). Therefore, there is still enough computational space to add more channels to decode on a single Motorola DSP56362 and it is possible to decode a 5.1 multi-channel (MC) bitstream. In case of a MC 5.1 bitstream, the memory requirements for an MPEG-2 AAC 5.1 MC decoder implementation with a 96 kHz input signal increases to 18.9 k words and the computational power to 80 MIPS.

CONCLUSION

This paper shows the possibilities of encoding high resolution audio with the MPEG-2/4 Advanced Audio Codec. After a short comparison of existing solutions for high resolution audio encoding the AAC-specific issues coming up with the increased sampling rate and the increased amplitude resolution are discussed. The results of listening tests that were performed to determine the bitrates required for a transparent audio quality are presented.

Furthermore, the problems of implementing such a system on a DSP platform are discussed. A real time chain consisting of a Texas Instruments TMS320C6711 DSP and a Motorola 56362 DSP was built and the necessary processor load for encoding and decoding was examined. The result shows, that one Texas Instruments TMS320C6711 DSP is capable of encoding a high resolution audio stereo signal in real time with about 80% processor load and that a single Motorola 56362 DSP is sufficiently performant for decoding a high resolution 5.1 multi channel configuration.

REFERENCES

- [1] ISO/IEC 13818-7, "Information Technology - Generic coding of moving pictures and associated audio - Part 7: Advanced Audio Coding", 1997
- [2] ISO/IEC 14496-3, "Information Technology - Coding of audio-visual objects - Part 3: Audio", 1999
- [3] J. Robert Stuart, Meridian Audio, "Coding Methods for High Resolution Recording Systems", AES 103th Convention 1997, New York
- [4] Julian Dunn, Prism Sound, "Anti-alias and anti-image filtering: The benefits of 96 kHz sampling rate formats for those who cannot hear above 20 kHz", AES 104th Convention 1998, Amsterdam

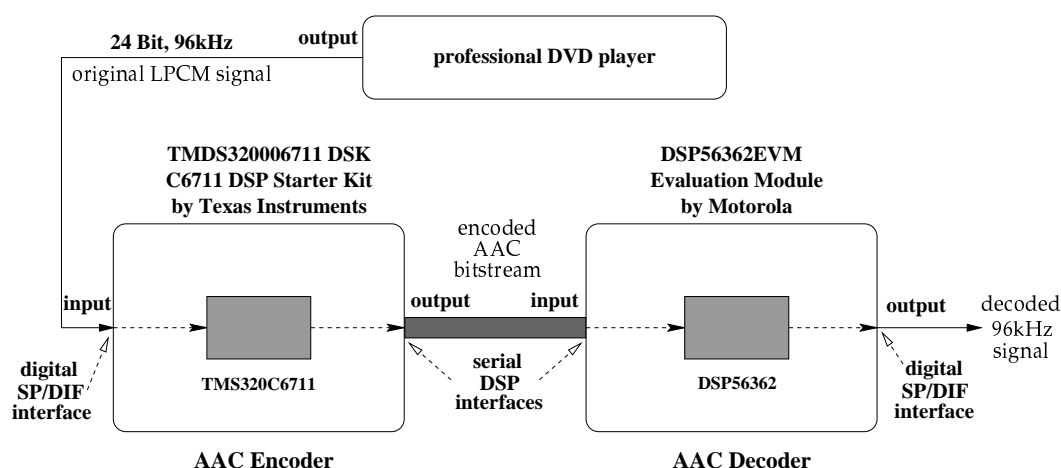


Figure 4: Real time encoding and decoding chain

- [5] Richard Black, "Anti-alias filter: the invisible distortion mechanism in digital audio?", AES 106th Convention 1999, München
- [6] Ashihara Kaoru, Kiryu Shogo, National Institute of Advanced Industrial Science and Technology, "Detection threshold for tones above 22 kHz", AES 110th Convention 2001, Amsterdam
- [7] Acoustic Renaissance for Audio, "A Proposal for the High Quality Audio Application of High-Density CD Carriers", private publication available for download at www.meridian-audio.com/ara
- [8] Jan Verbakel, Leon van de Kerkhof, "Super Audio CD Format", AES 104th Convention 1998, Amsterdam
- [9] M.A. Gerzon, P.G. Craven, J.R. Stuart, M.J. Law, R.J. Wilson, "The MLP Lossless Compression System", AES 17th International Conference on High Quality Audio Coding, Florence 1999
- [10] Marina Bosi, "High Quality Multichannel Audio Coding: Trends and Challenges", AES 106th Convention 1999, München
- [11] Mike Smyth, "An Overview of the Coherent Acoustics Coding System", Whitepaper 1999, DTS Inc, www.dtsonline.com
- [12] Zoran Fejzo, Stephen Smyth, Keith McDowell, Yu-Li You, Paul Smith, "Backward Compatible Enhancement of DTS Multi-Channel Audio Coding That Delivers 96-kHz / 24-Bit Audio Quality", 109th AES Convention, Los Angeles 2000, Preprint 5259
- [13] James A. Moorer, Sonic Solution, "48-Bit Integer Processing Beats 32-Bit Floating Point for Professional Audio Applications", AES 107th Convention 1999, New York
- [14] Datasheet, "The DTS CAE-4 Encoder and CAD-4 Decoder", DTS Inc., www.dtsonline.com