Advances in video/audio compression and emerging MPEG-4 AVC and HE-AAC standards

D. Milovanović¹, I. Dimković²
¹Faculty of Electrical engineering, Belgrade
²AheadSoftware GmbH, Germany

I INTRODUCTION

Two international organizations (ISO/IEC and ITU-T) have been heavily involved in the standardization of image, audio and video coding methodologies. ITU-T Video Coding Experts Group (VCEG) develops international standards for advanced moving image coding methods appropriate for conversational and non-conversational audio/video applications. ISO/IEC Moving Picture Experts Group (MPEG) develops international standards for compression and coding, decompression, processing, representation of moving pictures, images, audio and their combinations. It caters essentially to video storage, broadcast video, video streaming (video over internet/DSL/wireless) applications. ITU-T has been working on a video coding standard called H.26L since 1997. In August 1998, the first test model was ready and was demonstrated at MPEG’s open call for technology in July 2001. In late 2001, ISO/IEC MPEG and ITU-T VCEG decided on a joint venture towards enhancing standard video coding performance – specifically in the areas where bandwidth and/or storage capacity are limited. This joint team of both the standard organizations is called Joint Video Team (JVT). The standard thus formed is called H.264/MPEG-4 Part10 and is presently referred to as JVT/H.26L/AVC (Advanced Video Coding). The main goals of JVT are: significant coding efficiency, simple syntax specifications and seamless integration of video coding into all current protocols and multiplex architectures (network friendliness) [1, 2].

In March 2003, MPEG Audio took a major step forward with the finalization of the High Efficiency AAC specification. AAC (Advanced Audio Coding) is the result of intensive research and development, and contains many improvements over older coding methods including MP3. With sampling frequencies ranging from 8 to 96 kHz, and up to 48 channel support, there is a much better stereo coding and filter bank. HE-AAC is the latest improvement of the MPEG-4 AAC standard. Designed for ultra low bit rate coding, as low as 32 Kb/s Stereo (45:1 compression!) HE-AAC is able to achieve superior audio quality, without losing treble sound or collapsing of the stereo image.

II REQUIREMENTS of H.264/AVC

Requirements for H.264/AVC arise from the various video applications that it aims at supporting like video streaming, video conferencing, over fixed and wireless networks and over different transport protocols. The following lists the important requirements and how H.264/AVC meets them [3].

Robust video transmission using ParameterSet

One of the key problems faced by previous standards is their layered nature, which results in less robust video transmission in packet lossy environments. Previous standards contained header information about slice/picture/GOP/sequence that was coded at the start of each slice/picture/GOP/sequence. The loss of packet containing this header information would make the data dependent on this header, as useless. H.264/AVC overcame this shortcoming by making the packets transmitted synchronously in a real-time multimedia environment as self-contained. That is, each packet can be reconstructed without depending on the information from other packets. All information at higher layers is system-dependent, but not content-dependent and is conveyed asynchronously. Parameters that change very frequently are added to the slice layer. All other parameters are collected in a ParameterSet. H.264/AVC standard specifies a method to convey ParameterSets in a special Network Abstraction Layer (NAL) unit type. NAL is described in the next section. Different logical channels or out-of-band control protocols may be used to convey parameter sets from the coder to the decoder. In-band parameter set information and out-of-band control protocol should not be used in combination.

Network friendliness

Previous video coding standards, such as H.261, MPEG-1, MPEG-2 and H.263 were mainly designed for special applications and transport protocols usually in a circuit-switched, bit-stream oriented environment. JVT experts realized the growing importance of packet-based data over fixed and wireless networks right in the beginning and designed the video codec from that perspective. Common test cases for such transmission include fixed Internet conversational services as well as packet-switched conversational services and packet-switched streaming services over 3G mobile networks. These IP-networks usually employ IP on the network layer, UDP at the transport layer and RTP at the application layer. IP and UDP offer an unreliable datagram service while RTP makes the transport of media possible. Sequence numbers are used to restore the out-of-order IP packets. RTP payload does not add to the bit stream but specifies how the media information should be interpreted. The standardization process of both JVT codec and RTP
payload specifications for H.264/AVC is still an ongoing issue but the goal of designing a simple coding scheme should be achieved.

**Support for different bit rates, buffer sizes and start-up delays of the buffer**

In many video applications, the peak bit rate varies according to the network path and also fluctuates in time according to network conditions. In addition, the video bit streams are delivered to a variety of devices with different buffer capabilities. A flexible decoder buffer model, as suggested for H.264/AVC would support this wide variety of video application conditions – bit rates, buffer sizes and start-up delays of the buffer. H.264/AVC Video coding standard requires that the bit stream to be transmitted should be decodable by a Hypothetical Reference Decoder (HRD) without an underflow or overflow of the reference buffer.

**Improved prediction**

Earlier standards used maximum two prediction signals, one from a past frame and the other from a future frame in the bi-directional mode (B-picture). H.264/AVC allows multiple reference frames for prediction. Maximum of five reference frames may be used for prediction. Although this increases the complexity of the encoder, the encoder remains simple and the prediction is significantly improved.

**Improved fractional accuracy**

Fractional pel values add significantly to the accuracy of the reconstructed image. These are imaginary pel positions assumed to be stationed between physical pels. Their values are evaluated using interpolation filters. Previous standards have incorporated half-pel and quarter-pel accuracies. H.264/AVC improves prediction capability by incorporating 1/8-pel accuracies. This would increase the coding efficiency at high bit rates and high video resolution.

**Significant data compression**

Earlier standards implemented quantization steps with constant increments. H.264/AVC includes scalar quantizer with step sizes that increase at the rate of 12.5%. Chrominance values have finer quantizer steps. This provides significant data compression.

**Better coding efficiency**

H.264/AVC uses UVLC (Universal Variable Length Coding), CAVLC (Context-based Variable Length Coding) and CABAC (Context based Adaptive Binary Arithmetic Coding) to provide efficient entropy coding. CABAC provides as good as 2:1 coding gain over MPEG-2.

**Overlay coding technique**

Faded transitions are such that the pictures of two scenes are laid on top of each other in semi-transparent manner, and the transparency of the pictures at the top gradually changes in the transition period. Motion compensation is not a powerful enough method to represent the changes between pictures in the transition during a faded scene. H.264/AVC utilizes overlay coding technique that provides over 50% bit-rate savings in both cross-fades and through-fades compared to earlier techniques.

**Better video quality**

H.264/AVC employs blocks of different sizes and shapes, higher resolution fractional-pel motion estimation and multiple reference frame selection. These provide better motion estimation/compensation. On the transform front, H.264/AVC uses integer based transform, which approximates the DCT used in other standards besides eliminating the mismatch problem in its inverse transform.

**Group capabilities**

H.264/AVC grouped its capabilities into profiles and levels (Baseline, Main and Extended profile). A profile is a subset of the entire bitstream of syntax that is specified by the International Standard. Within each profile there are a number of levels designed for a wide range of applications, bit rates, resolutions, qualities and services. A level has a specified set of constraints imposed on parameters in a bitstream. It is easier to design a decoder if the profile, level and hence the capabilities are known in advance.

### III BASIC ARCHITECTURE OF THE STANDARD

Conceptually, H.264/AVC consists of two layers: Video Coding Layer (VCL) and Network Abstraction Layer (NAL). VCL is the core coding layer, which concentrates on attaining maximum coding efficiency. NAL abstracts the VCL data in terms of the details required by the transport layer and to carry this data over a variety of networks. The VCL layer takes care of the coding of transform coefficients and motion estimation/compensation information. NAL provides the header information about the VCL format, in a manner that is appropriate for conveyance by the transport layers or storage media. A NAL unit (NALU) defines a generic format for use in both packet-based and bit-streaming systems. The format for both the systems is the same except that the NAL unit in a bit-stream system can be preceded by a start code.

### IV SIGNIFICANT FEATURES

**Profiles and Levels**

Taking into account that all the users may not require all the features provided by H.264/AVC, Profiles and Levels have been introduced. It specifies a set of algorithmic features and limits, which shall be supported by all decoders conforming to that profile. The encoders are not required to make use of any particular set of features supported in a profile. For any given profile, Levels generally correspond to processing power and memory capability on a codec. Each level may support a different picture size – QCIF, CIF, ITU-R 601 (SDTV), HDTV, S-HDTV, D-Cinema and data rate varying from a few tens of kilobits per second (kbps) to hundreds of megabits per second (Mbps).

**CABAC arithmetic coding**

The following scheme is used by CABAC (Context-based Adaptive Binary Arithmetic Coding):

1. Context models are created based on the neighboring symbols referred to as context modeling.
2. Non-binary symbols are mapped into a sequence of binary decisions called bins.
3. For each bin, a context variable is defined by an equation of prior transmitted symbols. The possible numerical values of a context variable are called contexts and each context has a probability distribution associated with it.
4. The bins are then encoded with Adaptive Binary Arithmetic Coding. After coding of each bin, the
probability model is updated using the values of encoded bins.

**Fractional pel accuracy**

H.264/AVC supports one-quarter and one-eighth pel accuracy and the fractional sample accuracy is indicated by a parameter called motion resolution in H.264/AVC. If motion resolution has value 0, quarter sample resolution with a 6-tap filter is applied to luma samples in the block. If motion resolution value is 1, one-eighth sample interpolation with 8-tap filter is used. The interpolation is equivalent to upsampling of the frame. Figure 1 shows the interpolation process for motion vector accuracies of 1/8-pel accuracies. The frame has to be upscaled by a factor of 8 for 1/8-pel MV-accuracy. A combination of filters may be used to upscale by the required factor (Fig.1).

![Figure 1. Illustration of interpolation for fractional pel accuracy.](image)

**4x4 Integer transform**

The previous video coding standards relied on Discrete Cosine Transform (DCT) that provided the transformation but produced inverse transform mismatch problems. H.264/AVC uses an integer transform with a similar coding gain as a 4x4 DCT. It is multiplier-free, involves additions, shifts in 16-bit arithmetic, thus minimizing computational complexity, especially for low-end processes. The transformation of input pixels \(X = \{x_{00} \ldots x_{33}\}\) to output coefficients \(Y = \{y_{00} \ldots y_{33}\}\) is defined by

\[
Y = \begin{bmatrix}
1 & 1 & 1 & 1 & x_{00} & x_{01} & x_{02} & x_{03} & 1 & 1 & 1 \\
2 & 1 & -1 & -2 & x_{10} & x_{11} & x_{12} & x_{13} & 1 & -1 & -2 \\
1 & -1 & -1 & 1 & x_{20} & x_{21} & x_{22} & x_{23} & 1 & -1 & -1 \\
1 & -2 & -2 & -1 & x_{30} & x_{31} & x_{32} & x_{33} & 1 & -2 & -1
\end{bmatrix}
\]

This transform matrix is used in all (except 16x16 Intra DC) the 4x4 block transforms.

![Figure 2. Assignment of indices for a 4x4 luma block and corresponding a chroma block.](image)

The 16 luma DC coefficients of 16 (4x4) blocks, are transformed using Walsh-Hadamard transform.

\[
\begin{bmatrix}
1 & 1 & 1 & 1 & x_{00} & x_{01} & x_{02} & x_{03} & 1 & 1 & 1 \\
1 & -1 & -1 & -1 & x_{10} & x_{11} & x_{12} & x_{13} & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & x_{20} & x_{21} & x_{22} & x_{23} & 1 & -1 & 1 \\
1 & -1 & 1 & -1 & x_{30} & x_{31} & x_{32} & x_{33} & 1 & 1 & -1
\end{bmatrix}
\]

// 2

Chroma DC coefficients of four 4x4 blocks of each chroma component are transformed using Walsh-Hadamard transform.

\[
Y = \begin{bmatrix}
1 & 1 & DC_{00} & DC_{01} & 1 & 1 \\
1 & -1 & DC_{10} & DC_{11} & 1 & -1
\end{bmatrix}
\]

Multiplication by two can be performed either through additions or through left shifts, so that no actual multiplication operations are necessary. Thus, the transform is multiplier-free.

For input pixels with 9-bit dynamic range (because they are residuals from 8-bit pixel data), the transform coefficients are guaranteed to fit within 16 bits, even when the second transform for DC coefficients is used. Thus, all transform operations can be computed in 16-bit arithmetic. In fact, the maximum dynamic range of the transform coefficients fills a range of only 15.2 bits; this small headroom can be used to support a variety of different quantization strategies, which are outside the scope of this specification.

The inverse transformation of normalized coefficients \(Y' = \{y'_{00} \ldots y'_{33}\}\) to output pixels \(X'\) is defined by

\[
X' = \begin{bmatrix}
1 & 1 & 1 & 1 & y'_{00} & y'_{01} & y'_{02} & y'_{03} & 1 & 1 & 1 \\
1 & \frac{1}{2} & -1 & -1 & y'_{10} & y'_{11} & y'_{12} & y'_{13} & 1 & -\frac{1}{2} & -1 \\
1 & -\frac{1}{2} & 1 & 1 & y'_{20} & y'_{21} & y'_{22} & y'_{23} & 1 & -1 & 1 \\
1 & -1 & -\frac{1}{2} & -\frac{1}{2} & y'_{30} & y'_{31} & y'_{32} & y'_{33} & 1 & -1 & -1
\end{bmatrix}
\]

Multiplications by \(\frac{1}{2}\) are actually performed via right shifts, so that the inverse transform is also multiplier-free. The small errors introduced by the right shifts are compensated by a larger dynamic range for the data at the input of the inverse transform.

**V IMPLEMENTATIONS**

Greater complexity of H.264/AVC means more processing power is needed for encode/decode, hence a higher VLSI implementation cost. However, recent advances in semiconductors have enabled Standard Definition MPEG-4 AVC decoders to be available now, with High Definition to follow soon (UB Video JVT-C148, Videolocus JVT-D023).

<table>
<thead>
<tr>
<th>H.264/AVC Profiles</th>
<th>Target applications</th>
<th>Decoder complexity increase over MPEG-2</th>
<th>Estimates of efficiency improvements over MPEG-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>low delay applications, video phone, mobile</td>
<td>2.5 X more complex</td>
<td>1.5 x better</td>
</tr>
<tr>
<td>Extended</td>
<td>mobile, streaming</td>
<td>3.5 X more complex</td>
<td>1.75 x better</td>
</tr>
<tr>
<td>Main</td>
<td>interlaced video applications, broadcast, packaged media</td>
<td>4.0 X more complex</td>
<td>2.0 x better</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>H.264/AVC Profiles</th>
<th>Target applications</th>
<th>Decoder complexity increase over MPEG-2</th>
<th>Estimates of efficiency improvements over MPEG-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>low delay applications, video phone, mobile</td>
<td>2.5 X more complex</td>
<td>1.5 x better</td>
</tr>
<tr>
<td>Extended</td>
<td>mobile, streaming</td>
<td>3.5 X more complex</td>
<td>1.75 x better</td>
</tr>
<tr>
<td>Main</td>
<td>interlaced video applications, broadcast, packaged media</td>
<td>4.0 X more complex</td>
<td>2.0 x better</td>
</tr>
</tbody>
</table>

Table 1. Estimates of H.264/AVC decoder complexity.

**VI MPEG-4 HE-AAC**

MPEG-4 HE-AAC is the combination of MPEG AAC and the SBR Bandwidth Extension amendment, which was finalized during the March 2003 MPEG meeting [4]. The amendment is based on SBR (Spectral Band Replication) technology which Coding Technologies and its customers deploy the technology under the aacPlus™ brand name. Late 2001, its value was recognized by MPEG and it is now on track to become a core profile for MPEG-4 audio called High-Efficiency AAC. In order to address the needs of the digital TV and audio industry, MPEG-2 AAC-LC plus SBR is also standardized in an amendment to MPEG-2 AAC (Part 7 of MPEG-2). In addition, MPEG has also recognizes a “low-power” decoder variant for HE-AAC.
Developed jointly by Panasonic, NEC and Coding Technologies, this low-power decoding method requires 40% less processing power and decodes HE-AAC bitstreams with only a slightly reduced audio quality. The availability of both low-power and high-quality decoders for HE-AAC enables the standard to run on the widest possible range of processors in mobile and portable device applications [5].

MPEG-4 HE-AAC is not a replacement for AAC, but rather a superset which extends the reach of high-quality MPEG-4 Audio to much lower bit rates. HE-AAC decoders will decode both plain AAC and the enhanced AAC plus SBR. The result is a backward compatible extension of the standard which nearly doubles the efficiency of MPEG-4 Audio.

SBR is a unique bandwidth extension technique that enables audio codecs to deliver the same listening experience at around half the bit rate. As a result, HE-AAC delivers CD-quality stereo at 48Kbps and 5.1 channel surround sound at 128Kbps. This level of efficiency is ideal for Internet content delivery and fundamentally enables new applications in the markets of mobile and digital broadcasting.

SBR or Spectral Band Replication was developed by Coding Technologies as a generic method to significantly enhance the efficiency of perceptual audio codecs like MPEG AAC. SBR does not replace the core codec, but rather operates in conjunction with it to create a more efficient superset that can cut the required bit rate in half. MPEG-4 Audio uses SBR in conjunction with AAC to create the HE-AAC profile which Coding Technologies has given the name aacPlus.

Present in both the encoding and decoding process, SBR leverages the correlation between the low and high frequencies in an audio signal to describe the high-end of the signal using only a very small amount of data. This SBR data describing the high-frequencies is coupled with the low-frequency compressed data from the AAC codec. Once combined, the complete HE-AAC bitstream contains enough data to recreate the original signal.

For example, to create 48Kbps stereo HE-AAC, the encoder generates two signals: an MPEG AAC signal at about 42Kbps and a SBR signal at about 6Kbps. The SBR signal is then placed into the MPEG AAC auxiliary fields as defined in MPEG-4 and sent out as a complete 48Kbps MPEG-4 HE-AAC bitstream (Fig.3).

![Figure 3. MPEG-4 aacPlus codec structure showing the integration of SBR and AAC.](image)

Since the SBR data is placed within the AAC auxiliary fields, the enhanced signal will be accepted by both an existing AAC and a new HE-AAC decoder. If sent to an AAC decoder, only the low-frequency audio signal will be recognized and decoded. If sent to an HE-AAC decoder, the SBR and the AAC will be decoded to recreate the full frequency signal. This technique makes the new Profile forward compatible with AAC. Also, since the HE-AAC decoder contains a full-fledged AAC decoder, it is able to decode both the Plain AAC and HE-AAC MPEG-4 Audio profiles. This combination makes HE-AAC backward compatible with AAC.

VII CONCLUSIONS

ITU/MPEG JVT H.264/AVC has very good features like multiple reference frames, CABAC, different profiles and is suitable for applications like video streaming and video conferencing. Error resilience is achieved by using parameter sets, which can be either transmitted in-band or out-of-band. It is more flexible compared to the previous standards and this enables improved coding efficiency. However, it should be noted that this is at the expense of added complexity to the coder/decoder. Also, it is not backward compatible to the previous standards. The level of complexity of the decoder is reduced by designing it specifically for a profile and a level. H.264/AVC provides three profiles and levels within them. In all, H.264/AVC seems to have a combination for good video storage and real-time video applications. Various companies are developing software/hardware products based on this emerging standard.

MPEG-4 HE-AAC is the latest improvement of the AAC audio coding standard. HE-AAC is a superset which extends the reach of high-quality MPEG-4 Audio to much lower bit rates. HE-AAC decoders will decode both plain AAC and the enhanced AAC plus SBR. SBR is a unique bandwidth extension technique that enables audio codecs to significantly enhance the efficiency of perceptual audio codecs. The availability of both low-power and high-quality decoders for HE-AAC enables the standard to run on the widest possible range of processors in mobile and portable device applications.

LITERATURE


grupisane u profile ograničenog nivoa složenosti. MPEG-4 HE-AAC je najnovije poboljšanje standardnog perceptualnog audio kodera. HE-AAC zasniva se na SBR tehnici povećanja kvaliteta dekodovanog audio signala na niskim bitskim protocima dok je istovremeno ostvarena smanjena složenost implementacije dekodera.

Advances in video/audio compression and emerging MPEG-4 AVC and HE-AAC standards
D.Milovanović, I.Dimković