

ATSC Video and Audio Coding

GRANT A. DAVIDSON, SENIOR MEMBER, IEEE, MICHAEL A. ISNARDI, SENIOR MEMBER, IEEE, LOUIS D. FIELDER, SENIOR MEMBER, IEEE, MATTHEW S. GOLDMAN, SENIOR MEMBER, IEEE, AND CRAIG C. TODD, MEMBER, IEEE

Invited Paper

In recent decades, digital video and audio coding technologies have helped revolutionize the ways we create, deliver, and consume audiovisual content. This is exemplified by digital television (DTV), which is emerging as a captivating new program and data broadcasting service. This paper provides an overview of the video and audio coding subsystems of the Advanced Television Systems Committee (ATSC) DTV standard. We first review the motivation for data compression in digital broadcasting. The MPEG-2 video and AC-3 audio compression algorithms are described, with emphasis on basic concepts, system features, and coding performance. Next-generation video and audio codecs currently under consideration for advanced services are also presented.

Keywords—Audio coding, bandwidth compression, codecs, consumer electronics, data compression, digital communication, digital TV, HDTV, transform coding, TV broadcasting, video coding, video signal processing.

I. INTRODUCTION

Digital television (DTV) is beginning to emerge in the United States and elsewhere as one of the most significant new consumer applications of our time. The picture and sound improvements offered to viewers, as well as flexibility gained by new digital services, are compelling. Although the roots can be traced farther back, a major impetus for this technology came on 24 December 1996, when the U.S. Federal Communications Commission (FCC) adopted the major elements of the DTV standard proposed by the Advanced Television Systems Committee (ATSC) [1]. This action mandated use of the ATSC standard for DTV terrestrial broadcasts in the United States. Since that time, the ATSC DTV standard has also been adopted by the governments of Canada (1997), South Korea (1997), and Mexico (2004).

Manuscript received June 28, 2005; revised October 10, 2005.
G. A. Davidson, L. D. Fielder, and C. C. Todd are with Dolby Laboratories, San Francisco, CA 94103 USA (e-mail: gad@dolby.com).
M. A. Isnardi is with the Sarnoff Corporation, Princeton, NJ 08543 USA.
M. S. Goldman is with Tandberg TV, Bedford, NH 03110 USA.

Digital Object Identifier 10.1109/JPROC.2005.861715

This paper focuses on two of the key enabling technologies which made an all-digital transmission system commercially viable: video and audio coding. Milestones in the standardization of coding technologies were MPEG-2 video [2] and audio by the International Standards Organization in November 1994, followed by Dolby[®] AC-3 multichannel audio [3] by ATSC in May 1995. Now part of the ATSC DTV standard, MPEG-2 video and AC-3 audio provide an economic benefit by reducing the amount of digital information required to transmit and store DTV programs (data compression). During the ATSC standardization process, video and audio coding were found necessary to meet the FCC's requirement that one high-definition television (HDTV) channel, or multiple standard-definition television (SDTV) channels, fit within the 6-MHz spectrum allocation of a single NTSC (analog) channel. Although a discussion of the relationship between these coding technologies and the other required components of the ATSC DTV standard is outside the scope of this paper, interested readers are referred to [4] for more details. Comprehensive tutorials on the fields of video coding and perceptual audio coding can be found in [5] and [6], respectively. A more in-depth treatise on the principles and applications of digital coding can be found in [7].

Video and audio coding technologies consist of two complementary parts, an encoder and decoder (codec), connected by a data transmission or storage channel. The encoder receives as input an original digital video or audio signal and produces a compressed signal representation (bitstream) for transmission at a lower digital information rate (bitrate). The bitrate is expressed in bits per second (b/s). The decoder receives the bitstream, possibly with errors when channel transmission is imperfect, and reconstructs an approximation to the original digital signal. In the ATSC DTV system, the bitstream syntax and decoder are standardized (fixed). The encoder is only required to produce a bitstream that conforms to the standard; therefore, differences in encoder

¹Dolby is a registered trademark of Dolby laboratories.

design are allowed. This provides a very useful means for encoder refinement over time, as we illustrate later in the paper.

The use of video and audio coding technologies reduces the amount of bandwidth and power required for transmission in DTV systems by a factor of about 80. However, this much data compression exerts a cost, manifested in the form of distortion in the received video and audio signals. Distortion increases with the amount of compression applied, leading to a tradeoff between decoded signal fidelity and transmission efficiency.

Up to a point, video and audio coding distortion can be made imperceptible to human subjects. This is accomplished in part by designing features into video and audio coders which exploit known limitations in human perception. For example, signal components that can be completely removed with no perceptible consequence are *perceptually irrelevant*, and are subject to elimination during data compression. Signal components that have a perceptible consequence when removed or distorted are *perceptually relevant*. From the standpoint of maximizing data compression, the ideal perceptual-based coder removes all irrelevant components while fully preserving relevant ones.

A second means for video and audio compression is to remove *statistically redundant* signal components. Within an audio signal, for example, neighboring waveform samples are more likely to be correlated with each other than uncorrelated. The benefit of removing redundant (correlated) components is that the bitrate required to convey the remaining uncorrelated components is often significantly less than the bitrate of the original signal itself. Unlike irrelevant components, redundant components removed by the encoder can be restored by the decoder.

A convenient framework for removing irrelevant and redundant signal components is space (or time) to frequency domain signal processing. The original input signal is grouped into blocks of adjoining pixels or samples, and then converted into the frequency domain using a block transform such as the discrete cosine transform (DCT). The DCT is dual purpose, decorrelating the signal before coding and providing a starting point for irrelevancy reduction.

In summary, the objective of data compression is to minimize bitrate for a given level of perceived distortion, or alternatively, to minimize perceivable distortion at a specified bitrate. Hence, the single most important descriptor of a codec is the perceptual quality delivered and the way it diminishes with decreasing bitrate. As we describe later in the paper, standard techniques exist for evaluating the perceptual quality of codecs using a pool of human subjects.

The remainder of this paper is organized as follows. Section II cites some specific examples to motivate the use of data compression in DTV. Section III discusses the MPEG-2 video compression standard as employed in the ATSC standard, including specific constraints, broadcast and receiver requirements, and statistical multiplexing features. The ATSC digital audio compression standard (AC-3) is described in Section IV, including an algorithm

overview, ancillary features, and subjective quality evaluation. Sections III and IV conclude with summaries of the next-generation video and audio compression systems currently under consideration for inclusion as extensions to the ATSC standard.

II. MOTIVATION

Studio-quality digital video, in an uncompressed state, requires between 200 Mb/s²–1.2 Gb/s³ transmission rate, and between 180 GB⁴–1100 GB⁵ storage capacity per 2-h movie. Such enormous amounts of transmission and storage capacity make commercial, and especially consumer, applications impractical. Video compression is therefore essential for applications in which channel capacity is limited to about 20–30 Mb/s or in which storage capacity is limited to about 4–5 GB.

Studio-quality digital audio, in an uncompressed state, consisting of five channels, each sampled at 48 kHz and represented with 20 bits per sample, requires 4.8 Mb/s transmission rate, and 4.3 GB storage capacity per 2-h movie. While this is considerably less than the requirement for uncompressed video, audio compression is essential in DTV when combined with compressed video, or when transmitted as an audio-only program in a limited capacity channel.

Fig. 1 presents typical normalized bitrates in bits/pixel (bpp) for both uncompressed and compressed video signals. Actual bitrates are obtained by multiplying by the pixel rate. Typical broadcast studio video signals have a normalized bitrate of 16 bpp.⁶ MPEG-2 digital video compression lowers this to about 0.2 bpp,⁷ yielding a compression ratio of about 80 : 1. Typical broadcast studio multichannel audio programs are conveyed using 20 bits per sample. AC-3 audio compression reduces this to 1.67 bits per sample with a compression ratio of 12 : 1.

It is possible to encode both video and audio into either a constant bitrate (CBR) or a variable bitrate (VBR) stream. In video/audio broadcast applications, multichannel audio is typically coded CBR, whereas video is coded VBR. CBR streams are straightforward to edit and pass through a broadcast production environment. VBR streams complicate the broadcast production environment, but provide improved quality when multiple video signals are multiplexed into a single output stream.

² $704 \times 480 \times 30 \times 2 \times 10 = 202.8$ Mb/s (the 2 represents the factor needed for 4 : 2 : 2 color subsampling, and the 10 is for 10-b systems).

³ $1920 \times 1080 \times 30 \times 2 \times 10 = 1.244$ Gb/s (the 2 represents the factor needed for 4 : 2 : 2 color subsampling, and the 10 is for 10-b systems).

⁴ 202.8 Mb/s \times 60 s/min \times 120 min \times 0.125 B/b = 182.5 GB.

⁵ 1.244 Gb/s \times 60 s/min \times 120 min \times 0.125 B/b = 1119.6 GB.

⁶For uncompressed video, normalized bitrate is defined as $B \times C$, where B is the number of bits per sample (typically 8 or 10) and C is a color format factor. $C = 2$ for 4 : 2 : 2 and 1.5 for 4 : 2 : 0. 4 : 2 : 2 signals have twice the vertical color resolution compared to 4 : 2 : 0 signals.

⁷For compressed video, normalized bitrate is defined as $R / (H \times V \times F \times C)$, where R is the bitrate, H and V are the horizontal and vertical image dimensions, F is the frame rate, and C is the color format factor defined above.

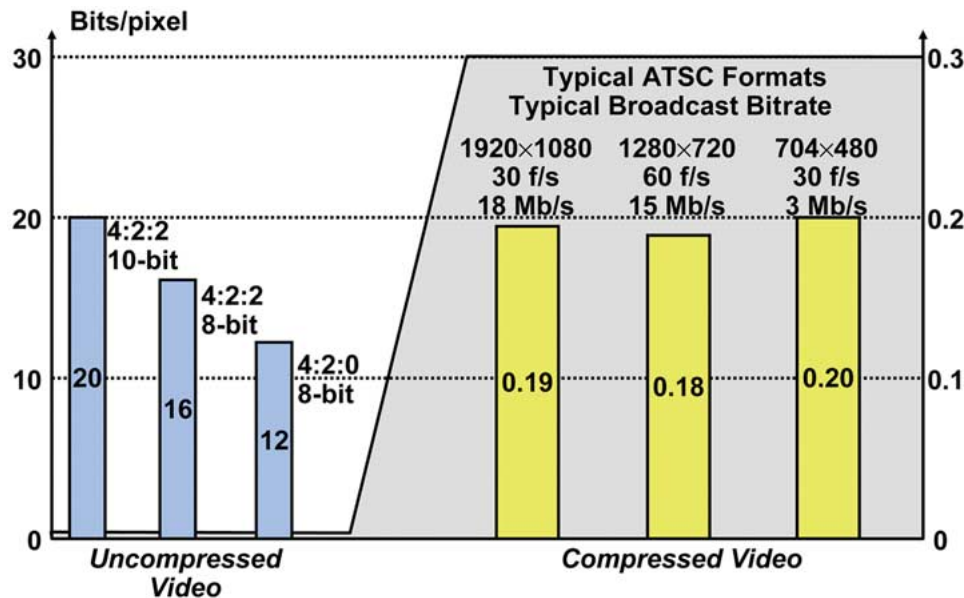


Fig. 1. Normalized bitrates for uncompressed and compressed video formats.

III. VIDEO CODING

This section explains the fundamentals of MPEG-2 video compression and how it achieves relatively large compression factors with very little degradation to subjective video quality. Since video signals are highly redundant in both space and time, good prediction is at the heart of efficient video compression. MPEG-2 employs a set of powerful prediction techniques to achieve high compression. Following this is a discussion on how MPEG-2 video compression is used in the ATSC DTV standard, and how its parameters can be adjusted to meet compression requirements imposed at both ends of the transmission channel. Broadcasters and other multichannel service providers rely on statistical multiplexing for efficient compression of multiple video signals that share the same output bitstream; this topic is briefly reviewed. Finally, this section concludes with a discussion of advanced video compression standards that are poised for growth in next-generation services and products.

A. MPEG-2 Video Compression Overview

MPEG-2 video compression enjoys widespread use as the current standard for both DTV and DVD applications. MPEG-2 is one of a series of international video compression standards developed by the Moving Picture Experts Group (MPEG), formally known as ISO/IEC JTC1/SC29/WG11. An excellent source of MPEG information is available in [8].

MPEG-2's immediate predecessors were MPEG-1, developed primarily for CD-ROM applications, and H.261, developed for videoconferencing. However, MPEG-1 and H.261 could not consistently perform well at normalized bitrates of 0.2 bpp. These earlier standards also lacked support for interlaced video and were not optimized for higher bitrates. Also missing was an efficient way to handle film-based material in which redundant fields were inserted during film-to-video

conversion. MPEG-2 addressed all of these shortcomings, and added scalability features as well. The MPEG-2 specification [2] became an international standard in 1994.

For terrestrial broadcast, the ATSC standard places an absolute limit of 19.4 Mb/s on the MPEG-2 video bitstream. In practice, some data capacity must be reserved for audio and system information, so video is coded at lower rates. For instance, high-definition video is typically encoded in the range 12–18 Mb/s and standard-definition video is typically coded at much lower rates (e.g., 3–6 Mb/s), especially when multiple programs are being sent in a single output bitstream.

The ATSC standard also supports a robust transmission mode called Enhanced 8-VSB that allows a broadcaster to allocate a portion of the 19.4 Mb/s bitrate to enhanced data transmission. Enhanced data is designed to have higher immunity to certain channel impairments than the main service but delivers data at a reduced information rate—either 1/2 or 1/4 the rate of the main service. In addition, during premium programming times, the maximum bitrate of the enhanced data stream is limited to 3 Mb/s.

1) *Coding Algorithm:* The MPEG-2 standard itself specifies the bitstream syntax and decoder operations. It does not specify encoder requirements beyond conformance to the bitstream syntax. In particular, the algorithmic details related to motion estimation, mode decision, and rate control—all of which greatly affect the complexity and performance of the encoder—are left to the encoder designer.

a) *Profiles and Levels:* MPEG-2 introduced the concept of profiles and levels as a way to define compliance points. A profile is defined as a set of specific functional elements (“tools,” in MPEG terminology) which must be supported by the decoder and the bitstream syntax. By imposing bounds on how a bitstream is generated, profiles determine the complexity of the video encoding. Within a given profile, levels determine the range of allowed values for elements of the bitstream syntax.

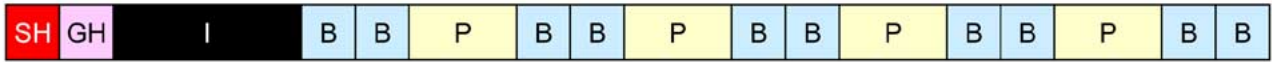


Fig. 2. Typical broadcast GOP structure, shown in coding order for a single GOP. SH: Sequence Header, GH: GOP Header. Lengths of bars show typical relative coded sizes of I , P , and B frames.

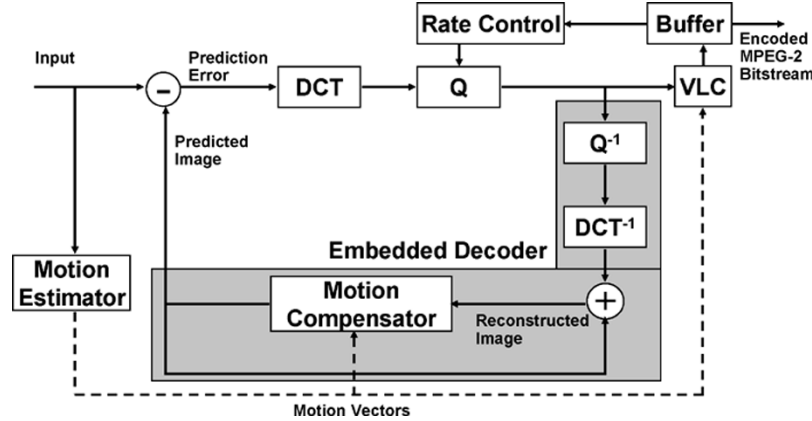


Fig. 3. Generic MPEG-2 video encoder block diagram, showing embedded decoder with grey shading. In an independent decoder, the output is the reconstructed image, not the predicted image.

For example, Main Profile allows all picture coding types, but does not allow studio-quality color formats or any scalable coding modes. Main Level supports a maximum image size of 720×576 , a maximum frame rate of 30 fps and a maximum bitrate of 15 Mb/s. This compliance point is called Main Profile at Main Level (MP@ML) and is used for standard definition DTV and DVD.

b) Frame Encoding: Video compression exploits the large amount of spatial and temporal redundancy inherent in all broadcast video. Within a video frame, neighboring pixels are more likely to be similar in value than different. Image content in neighboring frames are more likely to be simply displaced versions of corresponding content in the current frame than completely dissimilar.

I , or intracoded, frames exploit spatial redundancy and can be completely decoded without reference to any other frames. Spatial redundancy can be removed by transforming the pixels into the frequency domain, where energy is largely localized into the low-frequency coefficients. In addition, the transform allows frequency-weighted quantization to shape the noise so that it is least noticed by the human visual system. The DCT has been used successfully in JPEG, MPEG-1, and H.261, and was adopted again for MPEG-2. However, intraframe compression alone only yields about 1 bpp. In order to get to 0.2 bpp, temporal redundancy must be removed.

P , or predicted, frames use image data from a previously decoded frame to predict the current coded block. For each block of the original P frame, the encoder searches a previous I or P frame⁸ for a good match, thus forming what is called a motion-compensated prediction. The motion-compensated difference signal is DCT transformed, quantized, and variable-length coded. Typically, P frames use 50%–90% fewer bits than I frames.

⁸Since I and P frames are stored as reference pictures in both the encoder and decoder, they are also known as *anchor frames*.

B , or bidirectionally predicted, frames allow motion compensated prediction from a previous and a future decoded anchor frame. In order to ensure that the surrounding anchor frames precede the B frame in the bitstream, the frames must be reordered from input order into coding order at the encoder; the decoder must invert this ordering for display. B frames typically use 50% fewer bits than P frames and therefore improve the overall compression efficiency; however, since they require frame reordering, they cannot be used in low-delay applications such as videoconferencing.

The pattern of I , P , and B pictures is known as a group of pictures (GOP) structure. GOP length refers to the I frame spacing. Longer GOP structures have higher compression efficiency since the I frames are spaced relatively far apart. Typical GOP lengths are about 15 frames (0.5 s) for broadcast applications, as shown in Fig. 2.

c) MPEG-2 Video Coding Tools: Fig. 3 shows a generic encoder that contains the main coding tools. These tools are arranged in a prediction loop, with the predicted image generated by a decoder in the loop feedback path (shaded part of the figure). This embedded decoder contains a subset of the encoding functions and does not include such highly complex functions as motion estimation. In the absence of channel errors, and assuming the inverse transform operations in both encoder and decoder are identical, the output of the embedded decoder and the actual decoder produce identical reconstructed imagery.

MPEG-2 is a motion-compensated block transform-based video compression standard. Each small 16×16 region of the image (called a macroblock) is predicted by a motion compensation unit. For P pictures, a motion estimator finds the best match in the previous stored reference frame; for B pictures, two reference frames are used.

The encoder must make a number of important decisions that affect both the bitrate and quality of the encoding. One important decision is whether to code a macroblock as in-

Table 1
Compression Format Constraints

vertical_size_value	horizontal_size_value	aspect_ratio_information	frame_rate_code	progressive_sequence
1080	1920	1,3	1,2,4,5	1
			4,5	0
720	1280	1,3	1,2,4,5,7,8	1
480	704	2,3	1,2,4,5,7,8	1
			4,5	0
	640	1,2	1,2,4,5,7,8	1
			4,5	0

Legend for MPEG-2 coded values:

aspect_ratio_information: 1 = square samples, 2 = 4:3 display aspect ratio, 3 = 16:9 aspect ratio

frame_rate_code: 1 = 23.976 Hz, 2 = 24 Hz, 4 = 29.97 Hz, 5 = 30 Hz, 7 = 59.94 Hz, 8 = 60 Hz

progressive_sequence: 0 = interlaced scan, 1 = progressive scan

tercoded or intracoded. If no good prediction is found, the encoder will choose intracoded, as this mode will use fewer bits. There are many other mode decisions, such as whether a macroblock should use field or frame motion compensation and whether it should use field or frame spatial transform.

Each macroblock, whether it is intercoded or intracoded, undergoes a block DCT that produces an array of spatial frequency coefficients. For most video, the DCT has a number of desirable properties, including the ability to compact energy into the lowest spatial frequency coefficients, and allowing use of quantization that is well matched to the spatial response of the human visual system.

The DCT coefficients are quantized, which means that the number of allowable values is reduced. This is the only part of the encoding process in which information is irreversibly discarded, and it is the mechanism by which the instantaneous bitrate is controlled. The 8×8 array of DCT coefficients are quantized by a combination of a quantization matrix and a quantization scale factor. The intercoded and intracoded quantization matrices can be customized by the encoder on a picture basis, to better match, for instance, the properties of the video content. Customized quantization matrices are sent in the bitstream; if no matrices are sent, the encoder and decoder use default quantization matrices defined by the standard. The quantization scale factor offers macroblock-level control of quantization and bitrate. The rate controller adjusts the quantizer scale factor to optimize local quality and to meet bitrate targets.

Variable-length codes are applied to the quantized transform coefficients, motion vector differentials and other data structures; shorter codes are assigned to more probable values, resulting in an overall reduction in bitrate for video that contains redundant signal components.

A rate buffer collects the bursty sequence of variable-length codes and releases them in a more controlled fashion into the channel. The MPEG-2 standard allows both CBR and VBR modes of operation. However, the encoder must adhere to a video buffer verifier (VBV) model when constructing the bitstream. This model is an idealized decoder buffer model and essentially limits the variability in coded picture sizes. VBV compliance is especially important for broadcast applications in which bitstreams are “pushed” into the decoder; it ensures that the bitstream will not cause the decoder’s input buffer to overflow.

The video bitstream produced by the MPEG-2 encoding process is a sequence of codes representing the video structure (headers) and motion-compensated transform coefficients.

2) *MPEG-2 Bitstream Syntax*: The MPEG-2 video bitstream has a hierarchically layered structure.

- The *sequence layer* starts with a sequence header that contains important parameters that apply to all coded frames in the sequence, such as picture size and picture rate. A sequence consists of one or more coded pictures. Sequence headers allow proper initialization of decoders at random access points.
- The *GOP layer* is optional. If a GOP header is present, an *I* picture must be the first coded picture to follow. Because of this restriction, GOP headers are often used as random-access points in the bitstream. Typically, sequence and GOP headers precede every *I* picture in the bitstream so that random access (i.e., channel change) points occur frequently. The combination of sequence header/GOP header/*I* picture allow a decoder to completely reinitialize and start decoding.
- The *picture layer* contains buffer and picture type (i.e., *I*, *P*, and *B*) information. A picture contains one or more slices.
- The *slice layer* contains resynchronization information. If errors corrupt the bitstream, the decoder can wait for the next slice header and resume decoding.
- The *macroblock layer* is the basic coding structure and unit of motion compensation. For broadcast video, a macroblock consists of four 8×8 luminance blocks and the two corresponding 8×8 chrominance blocks.
- The *block layer* is the basic transform unit. It contains the 8×8 DCT coefficients of a luminance or chrominance block.

B. Relation Between ATSC and MPEG-2 Video Coding

ATSC bitstreams must conform to the MPEG-2 Main Profile at High Level, but with additional constraints placed on certain syntax elements in the sequence and picture layers. In particular, the picture size, frame rate and aspect ratio parameters must conform to Table A3 in [1], reproduced here in Table 1. These will be discussed in more detail in the next section.

1) *Video Formats*: Much discussion among industry groups went into the selection of the compressed video formats allowed by the ATSC standard. Two high-definition (HD) picture sizes are supported. Both are characterized by 16:9 display aspect ratio and square pixels. The 1920 × 1080 HD format allows both progressive and interlaced scan, but its frame rate is limited to 30 Hz. The 1280 × 720 HD format is progressive only, but can support frame rates as high as 60 Hz. Two standard-definition (SD) picture sizes are supported. The 704 × 480 format is directly derived from ITU-R 601 sampling. It supports both progressive and interlaced scan and both 4:3 and 16:9 display aspect ratio; however, it is the only ATSC format that does not support square pixels. The VGA-friendly 640 × 480 format is 4:3 and square pixels, and supports both progressive and interlaced scan.

Note that relative picture dimensions are the ratio of small numbers (e.g., 1080 : 720 = 3 : 2), making image resizing a fairly straightforward procedure, even for consumer electronic devices. This relationship is used to advantage in set top decoders that resize any decoded format into a single output format for display.

Another noteworthy feature of Table 1 is the support of “sister” frame rates, e.g., those that are related by the factor 1.001. This allows a migration path from the NTSC-related rates of 23.976/29.97/59.94 fps to the 24/30/60 fps rates supported by newer video products.

The ATSC video formats are a strict subset of those allowed by U.S. digital cable [9]. This attribute permits compliant pass-through of ATSC programming over digital cable. Cable also allows a 1440 × 1080 HD size and 720 × 480, 544 × 480, 528 × 480, and 352 × 480 SD sizes.

2) *Bitstream Specifications Beyond MPEG-2*: The ATSC specification also defines extensions to the MPEG-2 syntax to implement system-specific features. Two of these extensions are now discussed.

a) *Captioning Data*: ATSC allows a fixed 9600 b/s of closed caption data to be sent in the picture layer user data. This data capacity is ten times greater than that allowed by the analog NTSC Line 21 captioning standard [10]. In addition to simple pass-through of NTSC captioning data, the greater channel capacity allows additional features, such as multiple languages and easy-reader text, to be supported. The DTV closed captioning standard [11] defines the full functionality.

b) *Active Format Description and Bar Data*: The motion picture and television industries have used a variety of picture aspect ratios over the years. Active format description (AFD) and bar data [1], [12] provide information about how the useful, or “active,” picture area is contained within the coded video raster. This information can be used by receivers to provide the best possible representation of the active area on the intended display, which itself may have an aspect ratio of either 4:3 (normal) or 16:9 (widescreen).

AFD, when present, is carried in the picture user data of the MPEG-2 video elementary stream. Within this descriptor, a 4-b active_format field indicates the aspect ratio and geometrical relationship of the active area to the full coded frame.

For instance, active_format can indicate that a 16:9 active area is vertically centered within the coded frame.

AFD can only indicate a limited number of combinations of active aspect ratio and geometrical relationships and may not be precise. For precise signaling of the start and end of active area, bar data is used. Bar Data can indicate the line and pixel numbers of the bar borders and can track picture-to-picture variations in bar geometry.

3) *Colorimetry*: Colorimetry refers to the accurate reproduction of scene colors on the display. The MPEG-2 video specification allows important system parameters—color primaries, transfer characteristics and RGB-to-YCrCb matrix coefficients—to be signaled explicitly in the sequence layer of the bitstream. Although a wide range of standardized colorimetric system parameters are allowed by the MPEG-2 video specification, it is noted that two of them—ITU-R BT.709 for HDTV and SMPTE 170M for SDTV—are in common use.

It is important for the receiver to know the colorimetric parameters of the encoded video signal and to apply them to the display process. For instance, if the wrong inverse matrix is applied by the receiver in the production of RGB components from transmitted YCrCb components, then displayed colors will look incorrect.

C. Broadcast Requirements Relating to Video Compression

Broadcasters prefer high video quality at the lowest bitrates. Furthermore, they prefer reasonably fast channel change. As we see in this section, these goals cannot be achieved simultaneously, leading to a compromise solution.

1) *Video Quality*: The only quality requirement placed on ATSC transmission is that the quality of the received video must be at least as good as a standard-definition analog video signal. This is a subjective requirement, and unfortunately, there are no universally accepted video subjective quality assessment methods in use. Broadcasters use their best judgment on what is subjectively acceptable; if it is too low, they will know this by the number of complaints from their viewing audience.

A complicating factor is that the subjective video impairments of analog television, which include thermal noise (snow), FM impulse noise (dots) and multipath (ghosts), are absent in DTV.⁹ Instead, they are replaced by video compression artifacts, which include blockiness, mosquito noise, and ringing.¹⁰

2) *Channel Change*: Channel change is virtually instantaneous for analog television. However, in DTV systems, channel change is largely dependent on video encoding parameters such as GOP length. During a channel change, the new video bitstream is entered at a random point. The video decoder cannot initialize and start decoding until an I-frame

⁹If these analog impairments are already present at the input to the DTV encoder, then they will be present in the decoded image, unless special means are used to remove them prior to encoding.

¹⁰Mosquito noise and ringing are generally located along or near the edges of objects in the image.

appears.¹¹ Longer GOP lengths have I-frames spaced relatively far apart (e.g., every 1–2 s); they are more coding efficient, but impart longer channel change times, on average.

Since broadcasters would like viewers to switch to their DTV program with low delay but would also like to compress with high efficiency, they generally compromise using GOP lengths of about 0.5 s.

D. Receiver Requirements Relating to Video Compression

ATSC-compliant DTV receivers create a usable picture from any allowable transmitted video format.¹² There is no requirement that the coded video format be preserved on output. This allows receiver manufacturers great flexibility in their designs. For instance, one can design an ATSC DTV receiver that decodes all formats in Table 1 and converts them to a single standard-definition format for display.

ATSC receivers may also implement forms of error concealment. When uncorrectable bit errors occur, portions of coded video pictures are discarded by the transport subsystem of a DTV receiver. This lost information produces gaps in the video bitstream. When the decoder detects, or is informed of, an error, it generally stops decoding and resumes at the next slice header. The missing information can be replaced by picture data in the reference frame stores; more sophisticated decoders employ motion-compensated error concealment, using either MPEG-2's concealment motion vectors (if present) or proprietary motion interpolation schemes. Additional error concealment strategies are discussed in [12].

In fringe areas, the DTV signal may be lost and reacquired repeatedly by the receiver. In general, reacquisition times will be similar to channel change times. Upon reacquisition, the video and audio must be presented in synchronism.

E. Statistical Multiplexing

The content of video material varies greatly from talk shows to high motion sporting events. Due to the inherent differences in picture complexity, video compression by its nature requires a variable bitrate. With single channel applications, the video compression is rate controlled to fit into a fixed bandwidth transport. If multiple channels of digital video are carried within a single fixed bandwidth transport (such as a standard 19.4 Mb/s television allocation), having each digital channel rate controlled to a fixed bitrate is inefficient and will result in lower overall video quality or fewer digital channels being conveyed in the multichannel transport.

Statistical multiplexing is a mechanism to increase the number of possible digital video channels within a fixed transport by leveraging the inherent variations in the instantaneous bitrates required to achieve a particular constant picture quality over time. Fig. 4 shows an example of the

¹¹A sequence header, which contains initialization information, and a GOP header generally precede an I-frame in broadcast DTV bitstreams. Because of this association, "GOP length" generally refers to I-frame distance.

¹²The current FCC transmission rules do not specify video formats, so ATSC receiver manufacturers use Table 1 as a guideline.

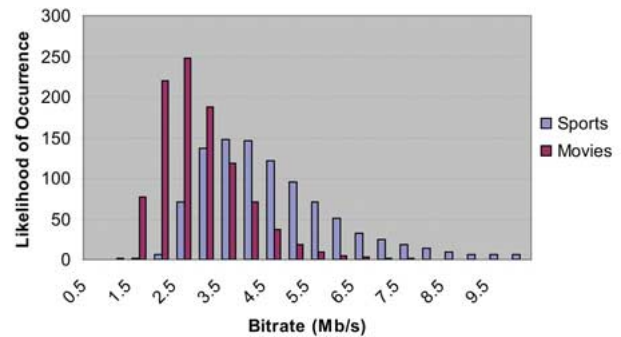


Fig. 4. An example of long-term statistics of two types of television channels.

probability of occurrence of a set of different bitrates for two types of SDTV programming: a sporting event and a motion picture. A typical fixed rate transmission may send such signals at, for example, a bitrate of 6 Mb/s. Most of the time there would be spare bitrate capacity; however, even at this rate there are some peaks where to attain the ideal quality would require more data than the available capacity.

If several video channels share, i.e., are multiplexed into, the same overall fixed rate transport, it is possible to take advantage of the fact that if one channel is demanding a high bitrate to code a picture well, it is likely that other channels have capacity to spare. The more channels that are statistically multiplexed, the higher the bandwidth savings.

Closed-loop statistical multiplexing is the most efficient mechanism for combining multiple services into a single transport. Using a variety of parameters, each encoder computes the video quality and complexity and forward this information to a bitrate arbiter (for example, a statistical multiplexer). The arbiter, in turn, informs each encoder how much bitrate is available to it over a specific time period based on an evaluation of each encoder's parameters and service priorities defined. Fig. 5 shows an example of four video channels and one opportunistic data channel being statistically multiplexed into an MPEG-2 transport stream (see [13] for more information about the MPEG-2 transport stream). The four video streams plus the data stream are limited to a predefined maximum bitrate. Opportunistic data, which can be used by broadcasters to deliver customized services to their audience, can be transmitted whenever the four combined video streams have a bitrate under this maximum limit. The limit is set by subtracting a fixed allocation for audio and system information from the total bitrate capacity of the transmission channel.

F. Next-Generation Video Coding

More than a decade has passed since MPEG-2 video became an International Standard, and the state of technology has leapt ahead. What was impossible to fathom back in the early 1990s for cost-effective receiver technology is now a reality. For the first time since MPEG-2 video, two new technologies are available that significantly decrease the amount of bandwidth needed for digital video: MPEG-4 AVC [14] and SMPTE VC-1 [15].

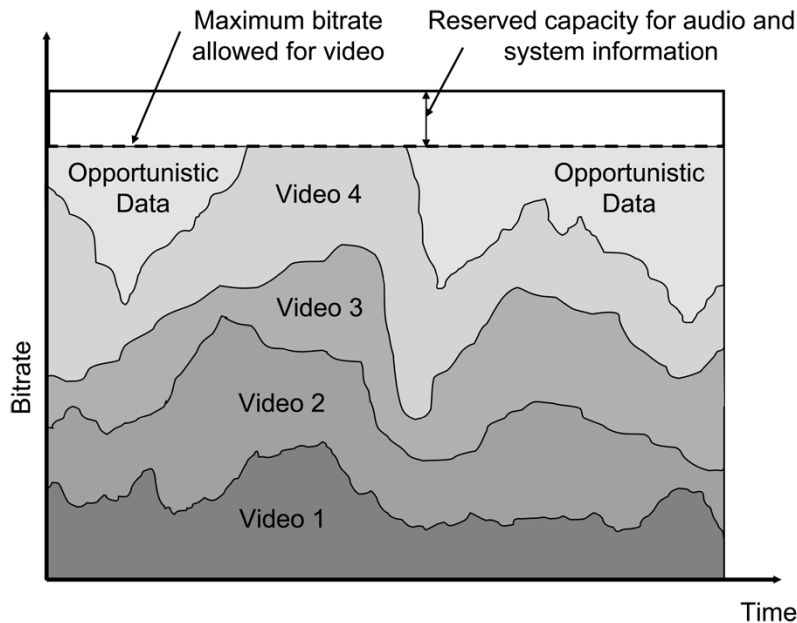


Fig. 5. An example of statistical multiplexing within an MPEG-2 transport stream.

Table 2
Terminology Primer for Various Compression Technologies

<ul style="list-style-type: none"> • MPEG-2 • H.262 	The ubiquitous video codec standard used in digital television today – Terrestrial Broadcasting, Cable, DBS, DVD-Video
<ul style="list-style-type: none"> • MPEG-4 Part 2 • MPEG-4 SP/ASP 	A follow-on video codec standard – not widely used for DTV
<ul style="list-style-type: none"> • MPEG-4 AVC • H.264 • MPEG-4 Part 10 • AVC • “JVT” 	Advanced Video Coding (AVC) – A next generation video codec standard jointly developed by ISO/IEC MPEG and ITU-T VCEG
<ul style="list-style-type: none"> • SMPTE VC-1 • SMPTE 421M • Windows Media™ Video 9 (WMV9) 	A next generation video codec developed by Microsoft – now a proposed standard in the SMPTE

The industry is using various terms to describe the same technology. Table 2 explains this.

Similarly to MPEG-2 video, both next-generation codecs are organized into profiles and levels to define specific interoperability points. As Table 3 shows, only certain profiles are applicable for broadcast-quality video. Both the terminology itself and the profile usage have caused some industry confusion as potential users attempt to compare video quality of what they believe are encodings made by the same technology but in fact are not. Examples include digital camcorders and World Wide Web video streaming applications to PCs.

Early results of both MPEG-4 AVC and SMPTE VC-1 are realizing 40%–50% compression efficiency gains over MPEG-2 video. Just like with MPEG-2 video, continual refinements of real-time implementations will occur over the next few years, albeit at a projected timescale more “compressed” than with MPEG-2 video. In 1994, the state of the art for real-time full ITU-R SD resolutions was 8–8.5 Mb/s, as shown in Fig. 6. With refinements in algorithmic imple-

Table 3
Profiles and Levels versus Application for Video Codecs

Application	MPEG-2 Video (H.262)	MPEG-4 AVC (H.264)	SMPTE VC-1 (Windows Media Video 9)
Mobile Devices Video Conferencing Internet Streaming	–	Baseline	Simple
Broadcast Quality (e.g., TV, Cinema, IPTV)	Main Profile SD: Main Level HD: High Level	Main Profile SD: Level 3 HD: Level 4	Main (progressive displays only)
		High Profile SD: Level 3 HD: Level 4	Advanced Profile SD: Level 1 HD: Levels 2&3

mentations, advanced preprocessing, technology advances, and statistical multiplexing, this has been reduced to under 3 Mb/s for the same picture quality. Most experts believe that the MPEG-2 video improvement curve is near its asymptotic theoretical minimum. For next-generation compression technologies, SD bitrates for similar picture quality start at under 2.5 Mb/s today and may drop below 1.25 Mb/s within the next few years.

With HD content, bitrate reduction is even more dramatic as an amount of consumed bandwidth per service. Only a few years ago, HD content required nearly 19 Mb/s. While today’s MPEG-2 HD content is being compressed at rates between 12–18 Mb/s, next-generation compression at similar picture quality is starting at approximately 8–10 Mb/s, and will likely drop below 6 Mb/s within the next few years (see Fig. 7).

As with MPEG-2 video, obtainable bitrates for a particular overall picture quality vary greatly with content, with real-time encoded high motion sports being one of the most difficult classes.

The coding gains come from the ability to perform more parallel processing and select better matches (i.e., better re-

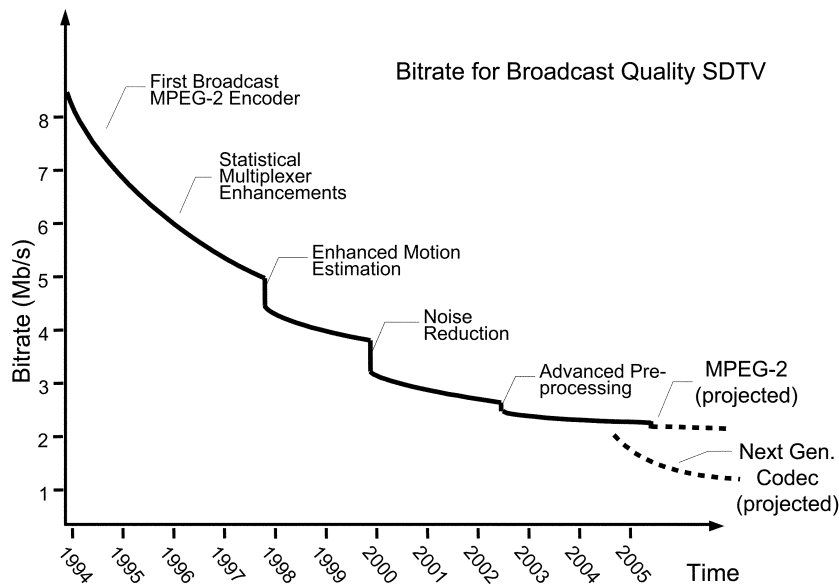


Fig. 6. Trends in broadcast quality SDTV bitrates for MPEG-2 and next-generation codecs.

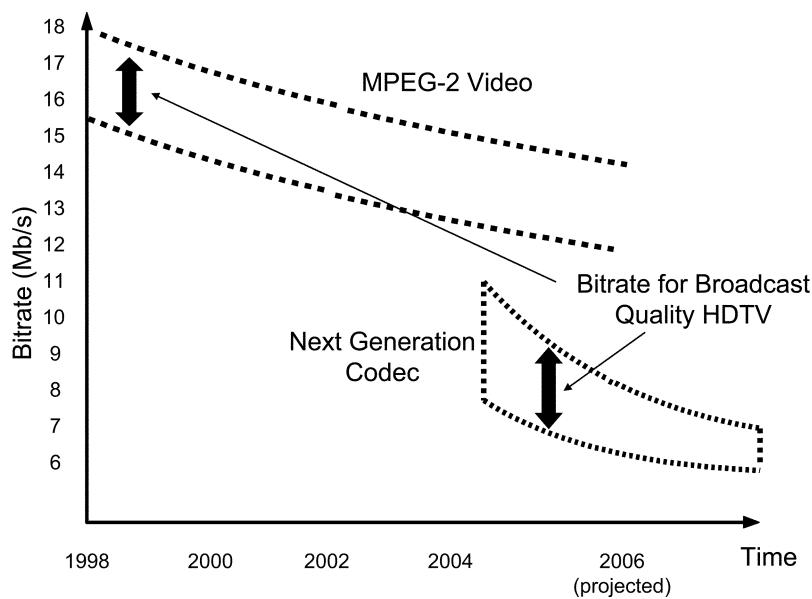


Fig. 7. Trends in broadcast quality HDTV bitrates for MPEG-2 and next-generation codecs.

sults on the rate-distortion curve) in real time and improved entropy coding, resulting in fewer bits used in the stream processing stage. In the next-generation codecs, expanded prediction modes, an in-loop deblocking filter, and a more efficient bitstream syntax have also led to significant improvements. Table 4 contains a summary of the algorithmic tool differences among MPEG-2 video, MPEG-4 AVC, and SMPTE VC-1.

IV. AUDIO CODING

Many of the basic techniques commonly employed in video coding, such as time to frequency transforms and perception-based quantization, have direct counterparts in audio coding. However, there are differences in the relative significance of these techniques, due primarily to differences

in signal properties and the way the signals are perceived by humans. For example, to exploit the inherent high redundancy in a group of neighboring video frames, MPEG-2 video incorporates comprehensive forward and backward frame prediction techniques. Although temporal prediction is sometimes used in audio coding, generally higher gains are realized using adaptive length overlapped transforms (compared to fixed-size nonoverlapped DCTs in MPEG-2 video) and perception-based irrelevancy reduction schemes. These methods exploit the known limitations of human hearing, such as frequency and temporal noise masking, to conceal quantization noise.

In this section, we present the forces which led to the development of AC-3, summarize the primary algorithmic and audio system features, and discuss subjective audio quality

Table 4

Comparison of Algorithmic Elements Used in MPEG-2 Video, MPEG-4 AVC, and SMPTE VC-1

Algorithmic Element	MPEG-2 Video (H.262)	MPEG-4 AVC (H.264)	SMPTE VC-1 (Windows Media™ Video 9)
Intra Prediction	– None: MB encoded – DC predictors	– 4×4 Spatial – 16×16 Spatial – I_PCM	– Frequency Domain Coefficient
Picture Coding Type	– Frame – Field – Picture AFF	– Frame – Field – Picture AFF – MB AFF	– Frame – Field – Picture AFF – MB AFF
Motion Compensation Block Size	– 16×16 – 16×8, 8×16	– 16×16 – 16×8, 8×16 – 8×8 – 8×4, 4×8 – 4×4	– 16×16 – 8×8
Motion Vector Precision	– Full Pel – Half Pel	– Full Pel – Half Pel – Quarter Pel	– Full Pel – Half Pel – Quarter Pel
P Frame Feature	– Single Reference	– Single Reference – Multiple Reference	– Single Reference – Intensity Compensation
B Frame Feature	– One Reference Each Way	– One Reference Each Way – Multiple Reference – Direct & Spatial Direct Modes – Weighted Prediction	– One Reference Each Way
In-Loop Filters	– None	– De-Blocking	– De-Blocking – Overlap Transform
Entropy Coding	– VLC	– CAVLC – CABAC	– Adaptive VLC
Transform	– 8×8 DCT	– 4×4 Integer “DCT” – 8×8 Integer “DCT”	– 4×4 Integer “DCT” – 8×4, 4×8 Integer “DCT” – 8×8 Integer “DCT”
Other	– Quantization Scaling Matrices	– Quantization Scaling Matrices	– Range Reduction – In-Stream Post Processing Control

evaluation methods and results. We conclude with a description of advanced coding technologies and features recently added to AC-3 and standardized by ATSC as Enhanced AC-3 [3].

A. AC-3 Audio Compression

From 1976 to 1992, the prevalent means for conveying movie soundtracks on 35-mm film was an analog matrix surround system in which the four original audio channels (left, center, right, surround) are mixed to two, stored on the optical print, and then converted back to four during playback (4-2-4 matrix technology). When movement began in the early 1990s to store digital surround sound on film, initial suggestions were to simply wrap 4-2-4 matrix technology around two discrete digital audio channels.

AC-3 was initially developed as an alternative digital audio storage means for film, designed specifically to avoid the performance limitations of matrix technology. AC-3 was the first codec to jointly code more than two audio channels into one composite bitstream. This approach achieves greater coding efficiency than multiple monophonic or stereo codecs, as well as the matrix approach, by exploiting the fact that the five main audio channels are delivered and

presented to listeners simultaneously. As embodied in the ATSC standard, AC-3 was also the first codec to employ metadata¹³ in the bitstream, allowing listeners to adjust playback for different room conditions, e.g., dynamic range control and channel downmixing. In addition to its use in the ATSC DTV standard, AC-3 is the audio compression format required on DVDs.

The channel configurations supported by AC-3 meet the recommendations for multichannel sound reproduction contained in ITU-R BS.775-1 [16]. AC-3 is capable of encoding a range of discrete, 20-kHz-bandwidth audio program formats, including one to three front channels and zero to two rear channels. An optional low-frequency effects (LFE or subwoofer) channel is also supported. The most common audio program formats are stereo (two front channels) and 5.1 channel surround sound (three front channels, two surround channels, plus the low-frequency effects channel, denoted “.1”). Input channels are coded into a bitstream ranging from 32 to 640 kb/s.

The remainder of this section provides overviews of the most prominent data compression features in AC-3 encoders

¹³“Metadata” refers to data about the data.

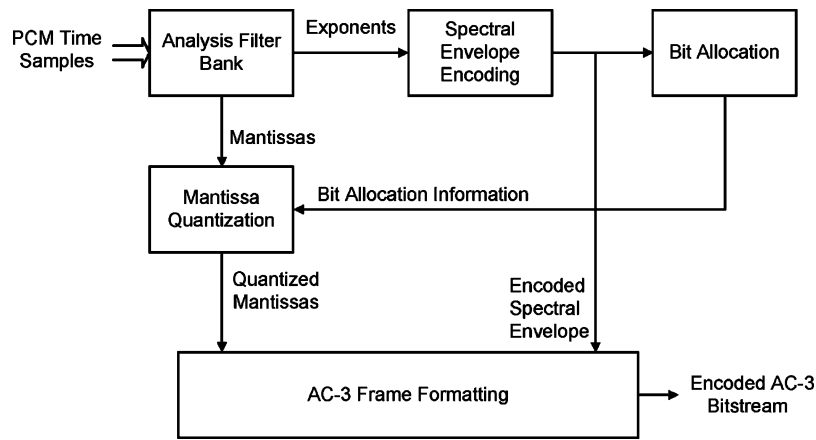


Fig. 8. Block diagram of an AC-3 encoder.

and decoders (Section IV-A1), the bitstream syntax (Section IV-A2), audio dynamic range control (Section IV-A3), audio subsystem services (Section IV-A4), and subjective audio quality evaluation (Section IV-A5). For a more detailed discussion of AC-3 features, the reader is referred to [17] or, for a complete specification, the ATSC AC-3 standard itself [3].

1) *Coding Algorithm:* In general respects, the architecture of AC-3 encoders and decoders is similar to generic perceptual audio codecs. Encoder processing starts by partitioning incoming channel streams of digital audio samples into contiguous frames. The AC-3 frame length is fixed at 1536 samples per channel (32 ms in duration with a sample rate of 48 kHz). A time-to-frequency analysis is performed on the waveform contained in each frame so that further processing and coding is performed in the frequency (auditory) domain. Of note, a psychoacoustic analysis is performed to estimate, for each of a multiplicity of nonuniform-width frequency bands, the power level at which coding distortion becomes just perceptible. The locus of these levels across the bands is called a *masking curve*. As in a generic perceptual audio codec, accuracy of the masking curve estimate has a noticeable effect on subjective quality of the decoded audio, as it separates relevant from irrelevant signal components. The remaining tasks for the encoder are to determine an appropriate bit allocation (quantization accuracy) for each frequency coefficient, and to format the coded data into a bitstream for transmission or storage. The bit allocation varies from frame to frame depending on signal characteristics, the masking curve, as well as the desired encoding bitrate. The AC-3 standard supports encoding in both constant and variable bitrate modes.

Like a generic perceptual audio decoder, an AC-3 decoder performs frame synchronization, detects errors, and then deformats the incoming bitstream. After some intermediate processing to reconstruct the quantized frequency-domain data, a frequency-to-time synthesis completes the process. The decoder generates 1536 digital audio samples per output channel, per frame.

We now turn to the specific architectural features which characterize and distinguish AC-3. A block diagram of the

encoder is shown in Fig. 8. The first step in the encoder is to convert all the audio samples in one frame into a sequence of six frequency coefficient blocks per input channel. The analysis filter bank is based on the oddly stacked time domain aliasing cancellation (OTDAC) transform [18], but modified as described below. The ensemble of frequency coefficients including all input channels in one transform time interval is called an *audio block* (AB). The input sample block for each transform is of length 512 and is overlapped by 50% with the preceding block. During decoding, each inverse transform produces 512 new audio samples, the first half of which are windowed, overlapped, and summed with samples from the last half of the previous block. This technique has the desirable property of crossfade reconstruction, which reduces audible distortion at block boundaries while maintaining critical sampling.

The AC-3 specific modification to OTDAC adds the capability to adaptively switch transform block length when signal conditions warrant (e.g., during intervals of rapid amplitude changes in the time waveform). A transform with adaptive time/frequency resolution can be implemented by changing the time offset of the transform basis functions during short blocks [17]. The time offset is selected to preserve critical sampling and perfect reconstruction at all times.

The next stage of processing (not shown in Fig. 8) is joint channel coding (spatial coding). Channel coupling is a method for reducing the bitrate of multichannel programs by mixing two or more correlated channel spectra in the encoder [19]. Frequency coefficients for the single combined (coupled) channel are transmitted in place of the individual channel spectra, together with a relatively small amount of side information. Rematrixing is a channel combining technique in which sum and difference signals of highly correlated stereo channels are coded in place of the original channels themselves. That is, rather than code and format left and right (L and R) in a two channel codec, the encoder processes $L' = (L + R)/2$ and $R' = (L - R)/2$.

Following joint coding, the individual frequency coefficients are converted into floating point representation as a binary exponent with one or more associated mantissas. The set of exponents is encoded into a representation of the signal

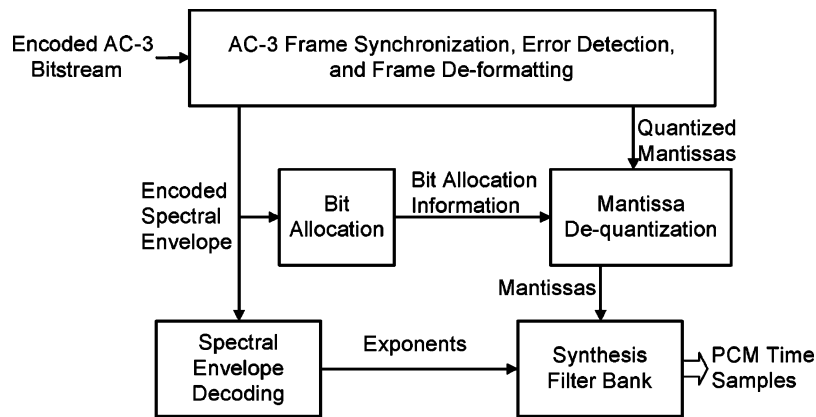


Fig. 9. Block diagram of an AC-3 decoder.

spectrum level across frequency, commonly referred to as the spectral envelope. The means for coding the spectral envelope in AC-3 provides for variable resolution in both time and frequency, allowing the encoder to adapt to the very wide variety of spectra present in motion picture soundtracks, instrumental music, music with vocals, and pure speech signals. In the frequency domain, one, two, or four mantissas can be shared by one floating-point exponent. In the time dimension, a spectral envelope can be sent for any individual AB, or shared between any two or more consecutive ABs within the same frame.

For short-term stationary audio signals, the spectral envelope remains substantially invariant within a frame. In this case, the AC-3 encoder transmits exponents once in AB 0, and then typically reuses (shares) them for the remaining blocks 1–5.

For short-term nonstationary signals, the signal spectrum can change significantly from block-to-block. In this case, the AC-3 encoder transmits exponents in AB 0 and in one or more other ABs as well. Exponent retransmission allows the coded spectral envelope to better match dynamics of the original signal spectrum. Sending multiple exponent sets in one frame results in an audio quality improvement if the benefit of a more accurate spectral envelope exceeds the cost of exponent retransmission.

The process of identifying perceptually irrelevant signal components, and determining the accuracy with which spectral components should be conveyed, is performed by the bit allocation step of Fig. 8. Bit allocation consists of distributing a pool of B bits, integral in number, to the mantissas in all six blocks for every channel in the frame, to minimize a perception-based distortion metric. The output is a bit assignment array which defines the quantization word length (resolution) of every mantissa in the frame. The bit assignment is performed subject to the constraint that the total number of allocated bits is less than or equal to B . B depends on the desired total bitrate, the number of side information bits in the frame, and other parameters.

The perception-based metric in AC-3 is based in part on a masking curve, as computed from a parametric model [17]. The masking curve is used to determine a desired quantization noise distribution across both time and frequency. Bits

are assigned to mantissas in a manner which causes the shape of the actual quantization noise to approximate that of the desired distribution. If the resulting quantization noise is entirely below the masking curve for all blocks in a frame, it is deemed inaudible.

Mantissas are limited in resolution using a set of scalar quantizers. To gain coding efficiency, certain quantized mantissa values are grouped together and encoded into a common codeword (composite coding). For example, in the case of the three-level quantizer, three quantized mantissas are grouped together and represented by a single 5-b codeword in the bitstream.

AC-3 spectral envelope (exponent) transmission employs differential coding, in which the exponents for a channel are differentially coded across frequency. The differential exponents are combined into groups using a composite coding scheme.

The AC-3 decoding process is a mirror-image reversal of the encoding process, except ABs are processed individually instead of as a group of six. The decoder, shown in Fig. 9, must synchronize to the encoded bitstream, check for errors, and deformat the various types of data such as the encoded spectral envelope and the quantized mantissas. The spectral envelope is decoded to reproduce the exponents. A simplified bit allocation routine is run, and the resulting bit assignment is used to unpack and dequantize the mantissas. The frequency coefficients are inverse transformed into the time domain to produce decoded digital audio samples.

There are several ways in which AC-3 decoders may determine that one or more errors are contained within a frame of data. The decoder may be informed of that fact by the transport system which has delivered the data. Data integrity may be checked using the two 16-b cyclic redundancy check words (CRCs) in each frame. Also, some simple consistency checks on the received data can indicate that errors are present. The decoder strategy when errors are detected is user definable. Possible responses include muting, block repeats, frame repeats, or more elaborate schemes based on waveform interpolation to “fill in” missing PCM samples. The amount of error checking performed, and the behavior in the presence of errors are not specified in the AC-3 ATSC standard, but are left to the application and implementation.

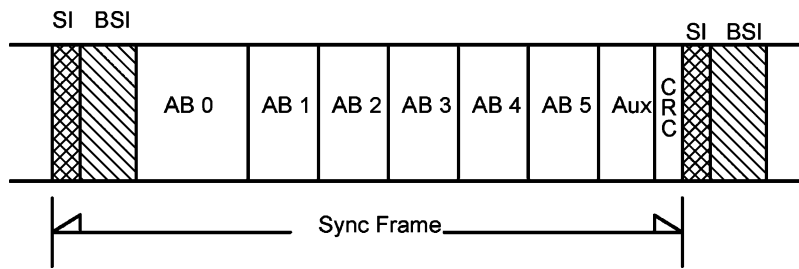


Fig. 10. AC-3 synchronization frame.

2) *Bitstream Syntax*: An AC-3 serial coded audio bitstream is composed of a contiguous sequence of synchronization frames. A synchronization frame is defined as the minimum-length bitstream unit which can be decoded independently of any other bitstream information. Each synchronization frame represents a time interval corresponding to 1536 samples of digital audio. All of the synchronization codes, preamble, coded audio, error correction, and auxiliary information associated with this time interval are completely contained within the boundaries of one audio frame.

Fig. 10 presents the various bitstream elements within each synchronization frame. The five different components are as follows: synchronization information (SI), bitstream information (BSI), AB, auxiliary data field (AUX), and CRC. The SI and CRC fields are of fixed length, while the length of the other four depends upon programming parameters such as the number of encoded audio channels, the audio coding mode, and the number of optionally conveyed listener features. The length of the AUX field is adjusted by the encoder such that the CRC element falls on the last 16-bit word of the frame.

The number of bits in a synchronization frame (frame length) is a function of sampling rate and total bitrate. In a conventional encoding scenario, these two parameters are fixed, resulting in synchronization frames of constant length. AC-3 also supports variable-rate audio applications.

Within one synchronization frame, the AC-3 encoder can change the relative size of the six ABs depending on audio signal bit demand. This feature is particularly useful when the signal is nonstationary over the 1536-sample frame. ABs containing signals that require a high bit demand can be weighted more heavily during bit allocation. This feature provides for local bitrate variation while maintaining an overall fixed bitrate.

3) *Loudness and Dynamic Range Control*: Prior to AC-3, consumer audio delivery systems simply conveyed one or two channels of audio into the home. Early in the development of AC-3, it was recognized that valuable new features would become available to listeners through the use of metadata. The first source of high-quality multichannel audio that would be delivered through AC-3 was motion picture soundtracks. The home is a very different acoustic environment than the cinema. The cinema is generally quiet (at least in the ideal), the soundtrack is reproduced at a calibrated sound pressure level (SPL), and the audience has no control over reproduction. The home is a variable environ-

ment, sometimes quiet and sometimes not. Furthermore, SPL levels in the home are under the listener's direct control, and are typically much lower (20–30 dB) than in the cinema. Historically, motion picture soundtracks have been subjected to dynamic range preprocessing prior to delivery to the home. This was done to compress the wide dynamic range soundtrack into a more limited range, rendering it more suitable for delivery over the restricted dynamic range analog channels that typically served domestic listening environments. This approach made it impossible for consumers (for example, the ones with a high-end home theater) to enjoy the soundtrack as originally created and intended.

The approach taken with AC-3 was to allow the original soundtrack to be delivered without any processing at all, with metadata providing the decoder important information to control dynamic range and dialog level. Consider that some soundtracks (e.g., motion pictures) will have high dynamic range with dialog levels well below full scale, and other programs with small dynamic range may have dialog level closer to full level (e.g., commercials). When different types of programs are concatenated in a DTV broadcast service, the listener could be subjected to dramatically changing dialog levels. From the perspective of the listening audience, a less annoying overall presentation is achieved with uniform dialog levels across all program types, as well as across program channels. One approach to uniformity of dialog level would be to standardize the dialog level within the digital coding range. This approach would have required broad industry acceptance which would have been very difficult, if not impossible, to achieve. Instead, AC-3 allows the program provider latitude to use any appropriate level for dialog, but requires delivered bitstreams to indicate the level of normal spoken dialog in an element of metadata (designated *dialnorm*). The value of *dialnorm* is intended to be static over the length of a program. Every AC-3 decoder uses the value of *dialnorm* to adjust the gain applied to the reproduced audio, bringing differing dialog level of all programs into uniformity at -31 dB; this is a few decibels lower than dialog level in typical movie soundtracks. The ATSC DTV standard [1] and FCC regulations require that *dialnorm* be correctly set by broadcasters. As methods become available for broadcasters to properly set the value of *dialnorm*, unintended SPL fluctuations during program switching will be eliminated.

Once level uniformity is established, dynamic range must be controlled, as certain programs will contain audio pas-

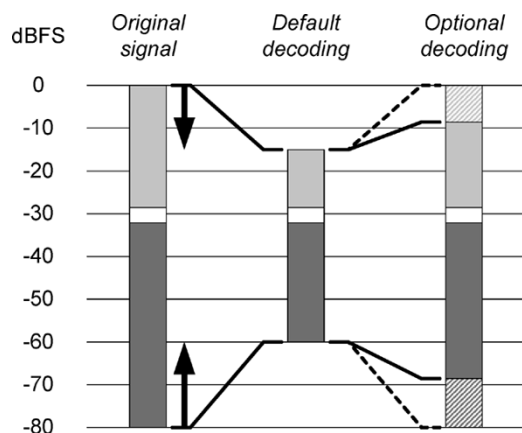


Fig. 11. Action of dynamic range control, dialog level at -31 dB.

sages that are much louder than dialog. AC-3 restricts wide dynamic range soundtracks with another element of metadata designated *dynrng*. By default, decoders use the value of *dynrng* to dynamically adjust the loudness of the reproduced audio. The values of *dynrng* can be generated prior to the AC-3 encoder, or within the AC-3 encoder by a dynamic range control algorithm. Values of *dynrng* are generated to bring loud sounds down in level, and quieter sounds up in level with respect to dialog level, as depicted in Fig. 11. Sounds (including most dialog) that are near dialog level are relatively unaffected. While many listeners will prefer the default decoder behavior with limited dynamic range, other listeners can scale the *dynrng* control signal so as to reproduce the audio with more, or even all, of the original program dynamic range.

4) *Audio Subsystem Services*: The ATSC audio subsystem offers several service types to meet the needs of a diverse listening audience. Multiple audio services are provided by multiple AC-3 streams. Each AC-3 stream conveyed by the transport system contains the encoded representation of one audio service. Specific fields are available in the AC-3 bitstream, and in the AC-3 descriptor included in the MPEG-2 program specific information, to identify the type of service provided by each AC-3 bitstream.

The audio services are generally classified as main services and associated services. Each associated service may be associated with one or more main services, or may stand alone. There are two types of main services and three primary types of associated services. Main service types are complete main (CM) and music and effects (ME). The CM service is the normal mode of operation where all elements of an audio program are present (dialog, music, and effects). The audio program may contain from one to 5.1 channels. The ME service indicates that music and effects are present without dialog. Associated services may be either single-channel programs that may be decoded and mixed simultaneously with a main audio service to form a complete program (requiring a “dual stream” decoder in the receiver to be properly reproduced), or they may be complete program mixes that do not need to be combined with a main service (requiring only a typical single stream AC-3

decoder). The primary associated service types are visually impaired (VI), hearing impaired (HI), and dialog (D).

An efficient means for offering an audio program in multiple languages is to send one 5.1 channel ME service together with a separate single-channel D service for each language. Based on the listener’s preference, the transport demultiplexer will select the appropriate D service (language) to deliver to a dual stream AC-3 decoder, for simultaneous decoding and mixing into the center channel of the ME service. In a similarly efficient way, a 5.1 channel service for the visually impaired can be provided with a 5.1 channel CM service for the main audience, and additionally providing a single-channel VI associated service containing a narrative description of picture content. Alternately, the VI service could be provided as a self-contained complete program mix (but tagged as a VI service) in any number of channels (up to 5.1). With this approach, a receiver with a single stream decoder can properly reproduce the VI service. To date, all receivers made only include single stream decoding.

5) *Subjective Quality Evaluation*: The task of evaluating subjective audio quality, while intensive and time consuming, is critical so that broadcasters can make informed decisions about their audio services. AC-3 was designed to meet the strict quality requirement for broadcast applications established in ITU-R BS.1115 [20]. This requirement, called “broadcast quality,” implies that impairments on all audio sequences for a particular codec are either “*imperceptible*” or “*perceptible but not annoying*,” as evaluated using a test methodology defined in Recommendation ITU-R BS.1116 [21]. This methodology is the most sensitive test method available, involving only the most critical (difficult to code) audio sequences. BS.1116 involves a double-blind triple-stimulus with hidden reference approach, requiring the listener to score both the codec under test and a hidden reference (source).

In a typical broadcast environment, an audio program will encounter repeated audio compression and decompression stages as it passes through production and distribution channels. These codecs provide a means for efficient multichannel audio storage on video tapes, video workstations, and contribution and distribution (backhaul) links over satellite or fiber circuits. An emission codec (AC-3) is used for final delivery to the audience. The ITU-R considers that a broadcast system could employ as many as eight generations of coding, followed by the emission codec. It is important that excellent audio quality be maintained through the entire cascade. The most common codec currently in use for the contribution and/or distribution of multichannel audio in DTV broadcasting is Dolby E.

The subjective quality of a cascaded Dolby E contribution-distribution codec, placed in tandem with AC-3 and other emission codecs, was evaluated in a BS.1116 subjective test [22]. Fig. 12 presents the performance of AC-3 when operated at 192 kb/s for two-channel stereo, both separately and when placed in tandem with a cascade with eight generations of Dolby E. The vertical scale shows the *diff grade*, which is the difference in score between the codec under test and

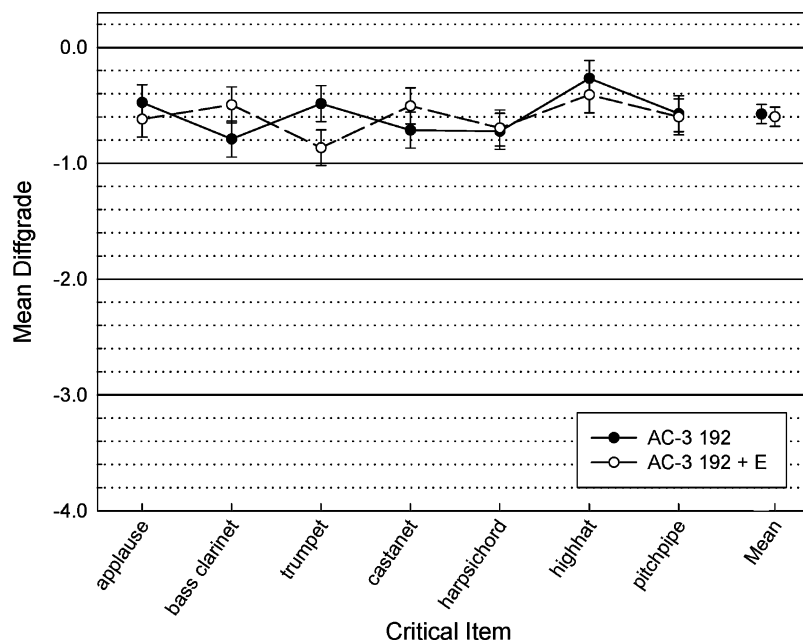


Fig. 12. Subjective test results for eight generations of Dolby E + Dolby Digital (AC-3) at 192 kb/s stereo.

the hidden reference. A score of 0 implies the impairment is considered imperceptible, and a score between 0 and -1.0 is considered perceptible but not annoying.

The results in Fig. 12 show that AC-3 satisfies the requirements for ITU-R broadcast quality for stereo signals at a bitrate of 192 kb/s, both separately and when placed in tandem with cascaded Dolby E. Similar performance is obtained with 5.1 channel signals at bitrates on the order of 384–448 kb/s.

B. Enhanced AC-3 (ATSC Standard)

An advanced version of the existing AC-3 coding system has been developed to better meet the needs of emerging broadcast and multimedia applications. This system is called enhanced AC-3 or E-AC-3. It provides for a wider range of supported bitrates, channel formats, and reproduction circumstances. Increased bitrate efficiency is obtained through the use of new algorithmic elements, while preserving a high degree of compatibility with existing AC-3 [3], [23].

1) *Expanded Bitrate Flexibility*: This new coding system is based on the existing AC-3 standard by preserving the present metadata carriage, underlying filter bank, and basic framing structure. The operating range has been increased by allowing bitrates spanning 32 kb/s–6.144 Mb/s. In addition, the bitrate control has a finer resolution, as little as 0.333 b/s at a sample rate of 32 kHz and a six-block transform frame size. The bitrate control is provided by a frame size parameter which sets the size of each substream in a frame to be 2–4096 B in size, incremented in 2-B intervals.

2) *Channel and Program Extensions*: The flexibility of the channel format has been expanded to allow for a significantly larger number of channels than 5.1. The increased channel capacity is obtained by associating the main audio

program bitstream with up to eight additional dependent substreams, all of which are multiplexed into one E-AC-3 bitstream. This allows the main audio program to convey the existing 5.1 channel format of AC-3, while the additional channel capacity comes from the dependent bitstreams.

Multiple program support is also available through the ability to carry seven more independent audio streams, each with a possible seven additional-channel, dependent substreams.

3) *Coding Algorithm Enhancements*: Coding efficiency has been increased to allow the use of lower bitrates. This is accomplished using an improved filter bank, improved quantization, enhanced channel coupling, spectral extension, and a technique called transient prenoise processing.

The adaptive hybrid transform is the combination of an improved filter bank and more efficient quantization methods. The filter bank is improved with the addition of a Type II DCT in cascade with the AC-3 OTDAC transform. This provides improved performance for stationary audio signals by converting the set of six 256-coefficient transform blocks into one 1536-coefficient hybrid transform block with increased frequency resolution.

This increased frequency resolution is combined with six-dimensional vector quantization (VQ) [24] and gain adaptive quantization (GAQ) [25] to improve coding efficiency for challenging audio signals such as pitch pipe and harpsichord. VQ is used to efficiently code frequency bands requiring lower resolution, while GAQ is used when higher quantization accuracy is warranted.

Improved coding efficiency is also obtained by combining channel coupling [19] with a new phase modulation technique. This new technique is called enhanced channel coupling. This method expands on the AC-3 method of employing a high-frequency mono composite channel which reconstitutes the high-frequency portion of each

channel on decoding. The addition of phase information and encoder-controlled processing of spectral amplitude information sent in the bitstream improves the fidelity of this process so that the mono composite channel can be extended to lower frequencies than was previously possible.

The manipulation of phase requires that a modified discrete sine transform (MDST) also be generated in the decoder, for example by performing an inverse MDCT followed by forward MDST. An angle scale value is applied to each MDCT/MDST coefficient pair, which is derived from two parameters: a bitstream subband angle value and a decorrelating angle value. This decorrelating angle value is derived by the decoder, based on a decorrelation bitstream scale factor and an associated random number sequence.

Another powerful tool added is called spectral extension. This method builds on the channel coupling concept by replacing the upper frequency transform coefficients with lower frequency spectral segments translated up in frequency. The spectral characteristics of the translated segments are matched to the original through spectral modulation of the transform coefficients mixed with a controlled amount of random noise. Noise blending compensates for the fact that spectral characteristics of typical music signals are more noise-like at higher frequencies.

An additional technique to improve audio quality at low bitrates is transient prenoise processing [26]. This is a decoder postprocessor that reduces the time spreading of quantization noise in transform blocks that contain transient signals. Quantization noise which occurs prior to the transient onset is called prenoise. In this technique, waveform segments corrupted by prenoise are replaced with a synthetic waveform which better approximates the original signal. Parameters computed in the E-AC-3 encoder are transmitted as side information to assist decoder postprocessing. Postprocessing utilizes time scaling and auditory scene analysis techniques.

V. CONCLUSION

In this paper, we have presented an overview of the video and audio coding subsystems of the ATSC DTV standard. This standard employs MPEG-2 video coding, with certain constraints and extensions, and completely specifies AC-3 audio coding. We provided the need for data compression in DTV broadcasting, and focused on the basic concepts employed. Both video and audio codecs strive to achieve the same goal; namely, to minimize bitrate for a given level of perceived distortion. The number of bits required to represent the original signal is reduced, to a large extent, using decorrelation techniques. MPEG-2 video relies on both transform coding and motion-compensated prediction to decorrelate the signal. AC-3 relies on adaptive overlapped transform coding for decorrelation, in combination with a noise masking model for irrelevancy reduction.

Through encoder refinements, MPEG-2 video and, to a lesser but still significant extent, AC-3 audio have increased in coding efficiency over the past decade. Nevertheless, coding technologies continue to evolve fundamentally. Two next-generation video codecs are currently under consid-

eration by ATSC for advanced services. These codecs are expected to offer approximately 50% bitrate savings over MPEG-2 video. Also, an enhanced AC-3 codec offering further bitrate savings was recently standardized by the ATSC. Like MPEG-2 video and AC-3, these advanced codecs are expected to realize performance gains over time. This will enable television programs of higher quality and/or quantity to be provided in the future.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the editors whose comments improved early drafts of this paper.

REFERENCES

- [1] *ATSC Standard: Digital television standard (A/53), revision D, including Amendment no. 1*, ATSC Document A/53C, Advanced Television Systems Committee, Washington, D.C., Jul. 27, 2005.
- [2] *Information technology—Generic coding of moving pictures and associated audio information: Video*, ISO/IEC 13 818-2, 2000.
- [3] *Digital audio compression standard (AC-3, E-AC-3), revision B*, ATSC Document A/52B, Advanced Television Systems Committee, Washington, D.C., Jun. 14, 2005.
- [4] J. C. Whitaker, *DTV: The Revolution in Electronic Imaging*. New York: McGraw-Hill, 1998.
- [5] T. Sikora, "Trends and perspectives in image and video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 6–17, Jan. 2005.
- [6] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–513, Apr. 2000.
- [7] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1984.
- [8] MPEG pointers & resources. [Online]. Available: <http://www.mpeg.org>
- [9] *Digital Video Systems Characteristics Standard for Cable Television*, ANSI/SCTE 432004.
- [10] *Line 21 data services*, CEA-608-B, Consumer Electronics Association.
- [11] *Digital television (DTV) closed captioning*, CEA-708-B, Consumer Electronics Association.
- [12] *Recommended practice: Guide to the use of the ATSC digital television standard*, ATSC Document A/54A, Advanced Television Systems Committee, Washington, D.C., Dec. 4, 2003.
- [13] B. J. Lechner, R. Chernock, M. Eyer, A. Goldberg, and M. Goldman, "The ATSC transport layer, including Program and System Information (PSIP)," *Proc. IEEE*, vol. 94, no. 1, pp. 77–101, Jan. 2006.
- [14] *Information technology—Coding of audio-visual objects—Part 10: Advanced video coding*, ISO/IEC 14 496-10 | ITU-T Rec. H.264, Sep. 28, 2004.
- [15] *Compressed video bitstream format and decoding process (proposed standard)*, SMPTE 421M VC-1.
- [16] *Multichannel stereophonic sound system with and without accompanying picture*, Recommendation ITU-R BS.775-1, International Telecommunications Union, Geneva, Switzerland, 1994.
- [17] G. A. Davidson, "Digital audio coding: Dolby AC-3," in *The Digital Signal Processing Handbook*, V. K. Madisetti and D. B. Williams, Eds. Boca Raton, FL: CRC, 1998, pp. 41-1–41-21.
- [18] J. Princen, A. Johnson, and A. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing 1987*, pp. 2161–2164.
- [19] C. C. Todd, G. A. Davidson, M. F. Davis, L. D. Fielder, B. D. Link, and S. Vernon, "AC-3: perceptual coding for audio transmission and storage," presented at the 96th Conv. Audio Engineering Soc., 1994, Preprint 3796.
- [20] *Low bit-rate audio coding*, Recommendation ITU-R BS.1115, International Telecommunications Union, Geneva, Switzerland, 1994.
- [21] *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, Recommendation ITU-R BS.1116, International Telecommunications Union, Geneva, Switzerland, 1994.

- [22] D. Grant, G. Davidson, and L. Fielder, "Subjective evaluation of an audio distribution coding system," presented at the 111th Conv. Audio Engineering Soc., 2001, Preprint 5443.
- [23] L. D. Fielder, R. L. Andersen, B. G. Crockett, G. A. Davidson, M. F. Davis, S. C. Turner, M. S. Vinton, and P. A. Williams, "Introduction to Dolby digital plus, an enhancement to the Dolby digital coding system," presented at the 117th Conv. Audio Engineering Soc., 2004, Preprint 6196.
- [24] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer, 1992.
- [25] M. M. Truman, G. A. Davidson, A. Ubale, and L. D. Fielder, "Efficient bit allocation, quantization, and coding in an audio distribution system," presented at the 107th Conv. Audio Engineering Soc., 1999, Preprint 5068, .
- [26] B. Crockett, "Improved transient pre-noise performance of low bit rate audio coders using time scaling synthesis," presented at the 117th Conv. Audio Engineering Soc., 2004, Preprint 6184.



Grant A. Davidson (Senior Member, IEEE) received the B.S. degree in physics from the California Polytechnic State University, San Luis Obispo, in 1980, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Santa Barbara, in 1984 and 1987, respectively.

From 1983 to 1987, he was a Member of the Communications Research Laboratory at the University of California, Santa Barbara, where he studied low bitrate speech compression

techniques using vector quantization and linear predictive coding. He has also studied special-purpose VLSI processor architectures for real-time DSP applications. Since 1987, he has been a Member of the Research Division of Dolby Laboratories, San Francisco, CA. He is a codeveloper of the Dolby AC-3, MPEG-2/4 Advanced Audio Coding (AAC), Dolby E, and Dolby E-AC-3 audio compression algorithms. He is currently a Principal Staff Engineer at Dolby. He has received 14 U.S. patents. His research activities include digital audio and speech processing, with an emphasis on psychoacoustics-based low bitrate audio coding algorithms.

Dr. Davidson is a member of the Audio Engineering Society (AES). He has received two Technical Emmy Awards. He has served on the IEEE Signal Processing Technical Committee on Audio and Electroacoustics since 1999.



Michael A. Isnardi (Senior Member, IEEE) received the B.S.E.E., M.S.E.E., and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1982, 1983, and 1986, respectively.

From 1986 to 1991, he was a Member of Technical Staff and Technology Leader at the David Sarnoff Research Center (now Sarnoff Corporation), Princeton, NJ, where he developed the Advanced Compatible Television System, the first advanced TV system tested by the FCC. From

1991 to 2005, he was a Technical Manager at Sarnoff, where he managed development of multimedia processing, video compression, and watermarking technology for entertainment and government applications. Early work included development of Sarnoff's Compliance Bitstreams product, which received a Technical Emmy Award. He is now a Distinguished Member of Technical Staff at Sarnoff. He has been awarded 31 U.S. patents.

Dr. Isnardi is an active member of the Advanced Television Systems Committee (ATSC) and Society of Motion Picture and Television Engineers (SMPTE). He received the David Sarnoff Award, Sarnoff Corporation's highest award for technical achievement in June 2000. In the same month, he also received an IEEE Third Millennium Medal from the Consumer Electronics Society. He was the conference chair for the 2002 International Conference on Consumer Electronics. He has presented many tutorials on MPEG and digital TV to the broadcast and consumer electronics industries.



Louis D. Fielder (Senior Member, IEEE) received the B.S. degree in electrical engineering from the California Institute of Technology, Pasadena, in 1974 and an M.S. degree in acoustics from the University of California, Los Angeles, in 1976.

From 1976 to 1978, he worked on electronic component design for custom sound-reinforcement systems at Paul Veneklasen and Associates. From 1978 to 1984, he was involved in digital-audio and magnetic recording research at the Ampex Corporation. At that time he became interested in applying psychoacoustics to the design and analysis of digital-audio conversion systems. Since 1984, he has worked at Dolby Laboratories, San Francisco, CA, on the application of psychoacoustics to the development of audio systems and on the development of a number of bitrate reduction audio codecs for music distribution, transmission, and storage applications. Additionally, he has investigated perceptually derived limits for the performance for digital-audio conversion, low-frequency loudspeaker systems, and loudspeaker-room equalization. Recently, he has worked on Enhanced AC-3, the extension to the AC-3 audio coding system.

Mr. Fielder is a fellow of the Audio Engineering Society (AES), a member of the Society of Motion Picture Television Engineers (SMPTE), and a member of the Acoustical Society of America. He was on the AES Board of Governors during 1990–1992, President during 1994–1995, and is currently the Treasurer-Elect.



Matthew S. Goldman (Senior Member, IEEE) received the B.S. (high honors) and M.S. degrees in electrical engineering from Worcester Polytechnic Institute, Worcester, MA, in 1983 and 1988, respectively.

He has been actively involved in the development of digital television systems since 1992. Until 1996, he was a consulting engineer at Digital Equipment Corporation, where he was the systems architect for a first generation digital video server system. From 1996 to 2000, he was

director of engineering, advanced systems development, at DiviCom Inc, where he specified MPEG-2 based systems solutions. He is currently Vice President of Technology, Compression Systems, for Tandberg Television Inc., Bedford, NH. He was a prominent participant and a project editor of the MPEG-2 standards, and has been influential in other industry organizations, including the Society of Cable Telecommunications Engineers, the Advanced Television Systems Committee, and the Society of Motion Picture and Television Engineers (SMPTE). He holds six patents related to digital video transport.

Mr. Goldman is a member of the SMPTE and a member of the Academy of Digital Television Pioneers.



Craig C. Todd (Member, IEEE) received the B.S. degree in physics from the California Institute of Technology, Pasadena, in 1970.

He is currently Vice President of Technology Strategy for Dolby Laboratories. He has been with Dolby Laboratories since 1977 and has been heavily involved with surround sound and audio coding. He developed and established the Dolby surround sound matrix technology, first in the cinema, and later in the home. He was one of the authors of ITU Recommendation BS.775, which

established the 5.1 channel audio format. He is one of the primary designers of AC-3 and Dolby E coding technologies and was heavily involved in the ACATs/ATSC effort to set the DTV standard for the United States. He is currently heavily involved in the setting of technology standards for D-Cinema.

Mr. Todd is a Fellow of the Audio Engineering Society (AES) and the Society of Motion Picture Television Engineers (SMPTE) and recipient of the SMPTE Samuel Warner Medal.