DTV: The Revolution in Electronic Imaging

Jerry C. Whitaker

McGraw-Hill
New York San Francisco Washington, D.C. Auckland Bogota
Caracas Lisbon London Mardid Mexico City Milan
Montreal New Delhi San Juan Singapore
Sydney Tokyo Toronto

Chapter

7

DTV Audio Encoding and Decoding

1 Introduction

Monophonic sound is the simplest form of aural communications. A wide range of acceptable listening positions are practical, although it is obvious from most positions that the sound is originating from one source rather than occurring in the presence of the listener. Listeners have accepted this limitation without much thought in the past because it was all that was available. However, monophonic sound creates a poor illusion of the sound field that the program producer might want to create.

Two channel stereo improves the illusion that the sound is originating in the immediate area of the reproducing system. Still, there is a smaller acceptable listening area. It is difficult to keep the sound image centered between the left and right speakers, so that the sound and the action stay together as the listener moves in the room.

The AC-3 surround sound system is said to have 5.1 channels because there is a left, right, center, left surround, and right surround, which make up the 5 channels. A sixth channel is reserved for the lower frequencies and consumes only 120 Hz of the bandwidth; it is referred to as the 0.1 or *low-frequency effects* (LFE) channel. The center channel restores the variety of listening positions possible with monophonic sound.

The AC-3 system is effective in providing either an enveloping (ambient) sound field or allowing precise placement and movement of special effects because of the channel separation afforded by the multiple speakers in the system.

7.1.1 AES Audio

AES audio is a standard defined by the Audio Engineering Society and the European Broadcasting Union. Each AES stream carries two audio channels, which can be either a stereo pair or two independent feeds. The signals are pulse code modulated (PCM) data streams carrying digitized audio. Each sample is quantized to 20 or 24 bits, creating an audio *sample word*. Each word is then formatted to form a *subframe*, which is multiplexed with other subframes to form the AES digital audio stream. The AES stream can then be

Table 7.1 Theoretical S/N as a Function of the Number of Sampling Bits

Number of Sampling Bits	Resolution (number of quantizing steps)	Maximum Theoretical S
18	262,144	110 dB
20	1,048,576	122 dB
24	16,777,216	146 dB

serialized and transmitted over coaxial or twisted-pair cable. The sampling rates support range from 32 to 50 kHz. Common rates and applications include the following:

- 32 kHz—used for radio broadcast links
- 44.1 kHz—used for CD players
- 48 kHz—used for professional recording and production

Although 18-bit sampling was commonly used in the past, 20 bits has become prevalei today.

At 24 bits/sample, the S/N is 146 dB. This level of performance is generally reserved fc high-end applications such as film recording and CD mastering. Table 7.1 lists the theoret ical S/N ratios as a function of sampling bits for audio A/D conversion.

7.1.2 Audio Compression

Efficient broadcast or recording of digital audio signals demands a reduction in the amount of information required to represent the aural signal [1]. The amount of digital information needed to accurately reproduce the original PCM samples may be reduced by applying a digital compression algorithm, resulting in a digitally compressed representation of the original signal. (In this context, the term *compression* applies to the digital information that must be stored or recorded, not to the dynamic range of the audio signal.) The goal of any digital compression algorithm is to produce a digital representation of an audio signal which, when decoded and reproduced, sounds the same as the original signal, while using a minimum amount of digital information (bit rate) for the compressed (or encoded) representation. The AC-3 digital compression algorithm specified in the ATSC DTV system can encode from 1 to 5.1 channels of source audio from a PCM representation into a serial bit stream at data rates ranging from 32 to 640 kbits/s.

A typical application of the bit-reduction algorithm is shown in Figure 7.1. In this example, a 5.1 channel audio program is converted from a PCM representation requiring more than 5 Mbits/s (6 channels × 48 kHz × 18 bits = 5.184 Mbits/s) into a 384 kbits/s serial bit stream by the AC-3 encoder. Radio frequency (RF) transmission equipment converts this bit stream into a modulated waveform that is applied to a satellite transponder. The amount of bandwidth and power thus required by the transmission has been reduced by more than a factor of 13 by the AC-3 digital compression system. The received signal is demodulated

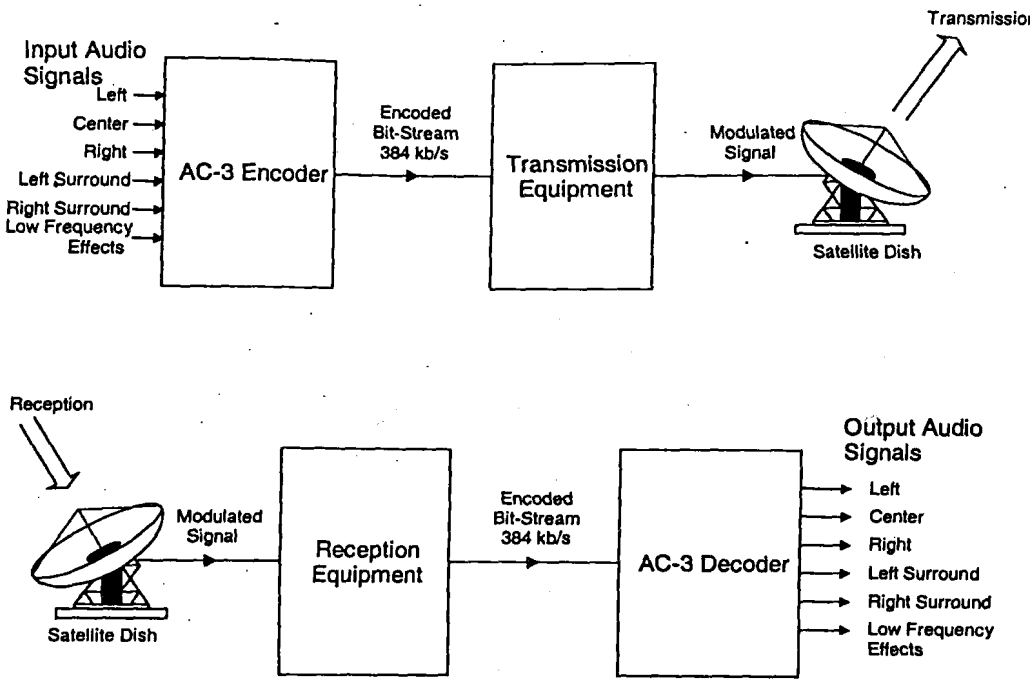

Figure 7.1 Example application of the AC-3 audio subsystem for satellite audio transmission. (*From* [1]. *Used with permission.*)

back into the 384 kbits/s serial bit stream, and decoded by the AC-3 decoder. The result is the original 5.1 channel audio program.

Digital compression of audio is useful wherever there is an economic benefit to be obtained by reducing the amount of digital information required to represent the audio signal. Typical applications include the following:

- Terrestrial audio broadcasting
- Delivery of audio over metallic or optical cables
- Storage of audio on magnetic, optical, semiconductor, or other storage media

7.1.3 Encoding

As discussed briefly in Section 5.7, the AC-3 encoder accepts PCM audio and produces the encoded bit stream for the ATSC DTV standard [1]. The AC-3 algorithm achieves high *coding gain* (the ratio of the input bit rate to the output bit rate) by coarsely quantizing a frequency-domain representation of the audio signal. A block diagram of this process is shown in Figure 7.2. The first step in the encoding chain is to transform the representation

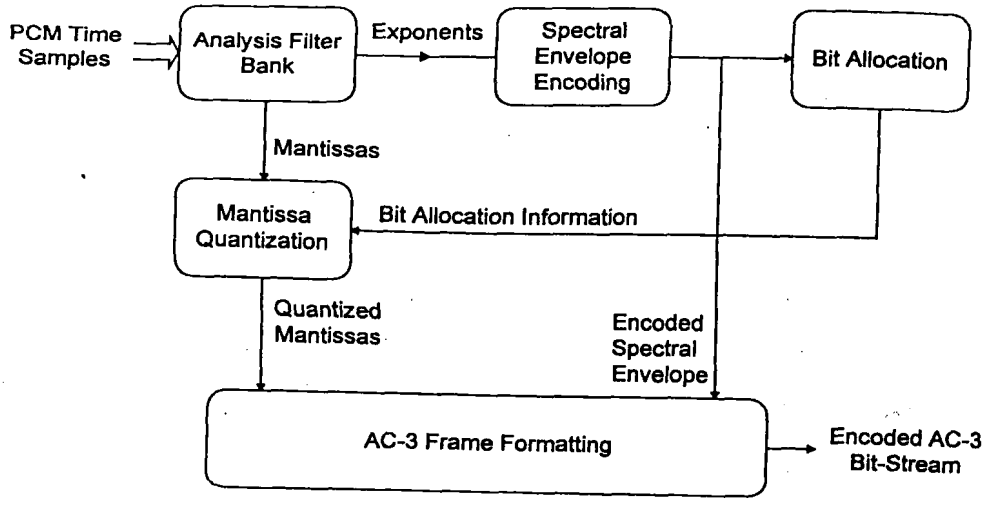

Figure 7.2 Overview of the AC-3 audio-compression system encoder. (*From* [1]. *Used v permission.*)

of audio from a sequence of PCM time samples into a sequence of blocks of frequenc coefficients. This is done in the *analysis filterbank*. Overlapping blocks of 512 time san ples are multiplied by a time window and transformed into the frequency domain. Becaus of the overlapping blocks, each PCM input sample is represented in two sequential tran formed blocks. The frequency-domain representation then may be decimated by a factor c 2, so that each block contains 256 frequency coefficients. The individual frequency coeff cients are represented in binary exponential notation as a *binary exponent* and a *mantissc* The set of exponents is encoded into a coarse representation of the signal spectrum referred to as the *spectral envelope*. This spectral envelope is used by the core bit-alloca tion routine, which determines how many bits should be used to encode each individua mantissa. The spectral envelope and the coarsely quantized mantissas for six audio block (1536 audio samples) are formatted into an AC-3 *frame*. The AC-3 bit stream is a sequenc of AC-3 frames.

The actual AC-3 encoder is more complex than shown in the simplified system of Fig ure 7.2. The following functions also are included:

- A frame header is attached, containing information (bit rate, sample rate, number of encoded channels, and other data) required to synchronize to and decode the encoded bit stream.
- Error-detection codes are inserted to allow the decoder to verify that a received frame of data is error-free.

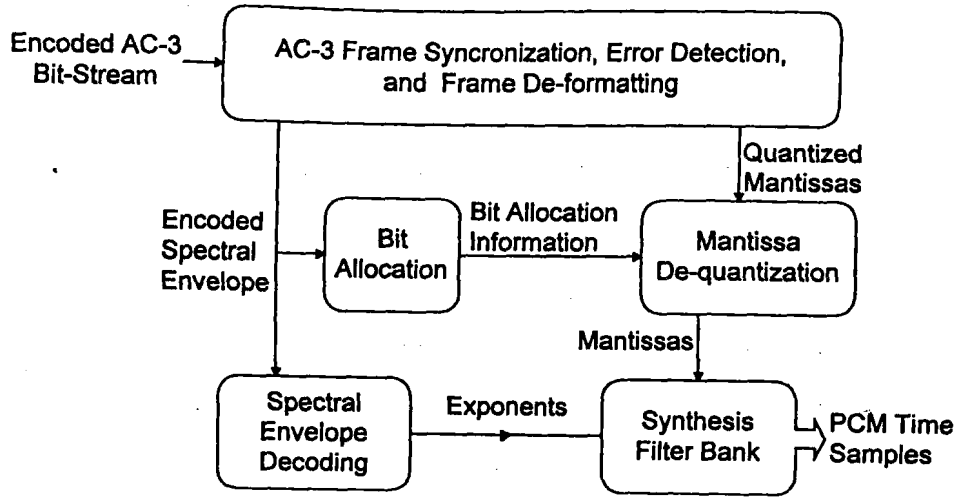

Figure 7.3 Overview of the AC-3 audio-compression system decoder. (*From* [1]. *Used with permission.*)

- The analysis filterbank spectral resolution may be dynamically altered to better match the time/frequency characteristic of each audio block.
- The spectral envelope may be encoded with variable time/frequency resolution.
- A more complex bit-allocation may be performed, and parameters of the core bit-allocation routine may be modified to produce a more optimum bit allocation.
- The channels may be coupled at high frequencies to achieve higher coding gain for operation at lower bit rates.
- In the 2-channel mode, a rematrixing process may be selectively performed to provide additional coding gain, and to allow improved results to be obtained in the event that the 2-channel signal is decoded with a matrix surround decoder.

7.1.4 Decoding

The decoding process is, basically, the inverse of the encoding process [1]. The basic decoder, shown in Figure 7.3, must synchronize to the encoded bit stream, check for errors, and deformat the various types of data (i.e., the encoded spectral envelope and the quantized mantissas). The bit-allocation routine is run, and the results are used to unpack and dequantize the mantissas. The spectral envelope is decoded to produce the exponents. The exponents and mantissas are transformed back into the time domain to produce the decoded PCM time samples. Additional steps in the audio decoding process include the following:

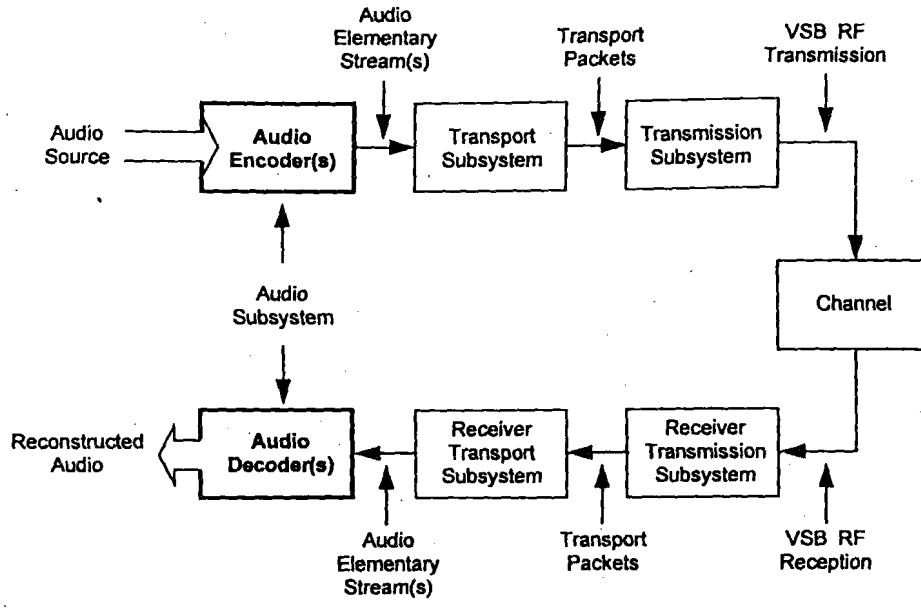

Figure 7.4 The audio subsystem in the DTV standard. (*From* [2]. *Used with permission.*)

- Error concealment or muting may be applied in the event a data error is detected.
- Channels that have had their high-frequency content coupled must be decoupled.
- Dematrixing must be applied (in the 2-channel mode) whenever the channels have been rematrixed.
- The synthesis filterbank resolution must be dynamically altered in the same manner as the encoder analysis filterbank was altered during the encoding process.

7.2 Overview of the AC-3 System

As illustrated in Figure 7.4, the audio subsystem of the ATSC DTV standard comprises the audio-encoding/decoding function and resides between the audio inputs/outputs and the transport subsystem [2]. The audio encoder is responsible for generating the *audio elementary stream*, which is an encoded representation of the baseband audio input signals. (Note that more than one audio encoder may be used in a system.) The flexibility of the transport system allows multiple audio elementary streams to be delivered to the receiver. At the receiver, the transport subsystem is responsible for selecting which audio streams to deliver to the audio subsystem. The audio subsystem is then responsible for decoding the audio elementary stream back into baseband audio.

An audio program source is encoded by a *digital television audio encoder*. The output of the audio encoder is a string of bits that represent the audio source (the audio elementary

stream). The transport subsystem packetizes the audio data into PES (*packetized elementary system*) packets, which are then further packetized into *transport packets*. The transmission subsystem converts the transport packets into a modulated RF signal for transmission to the receiver. At the receiver, the signal is demodulated by the receiver transmission subsystem. The receiver transport subsystem converts the received audio packets back into an audio elementary stream, which is decoded by the digital television audio decoder.

The partitioning shown in Figure 7.4 is conceptual, and practical implementations may differ. For example, the transport processing may be broken into two blocks; the first would perform PES packetization, and the second would perform transport packetization. Or, some of the transport functionality may be included in either the audio coder or the transmission subsystem.

7.2.1 Audio-Encoder Interface

The audio system accepts baseband inputs with up to six channels per audio program bit stream in a channelization scheme consistent with ITU-R Rec. BS-775 [3]. The six audio channels are:

- Left
- Center
- Right
- Left surround
- Right surround
- Low-frequency enhancement (LFE)

Multiple audio elementary bit streams may be conveyed by the transport system.

The bandwidth of the LFE channel is limited to 120 Hz. The bandwidth of the other (main) channels is limited to 20 kHz. Low-frequency response may extend to dc, but it is more typically limited to approximately 3 Hz (–3 dB) by a dc-blocking high-pass filter. Audio-coding efficiency (and thus audio quality) is improved by removing dc offset from audio signals before they are encoded. The input audio signals may be in analog or digital form.

For analog input signals, the input connector and signal level are not specified [2]. Conventional broadcast practice may be followed. One commonly used input connector is the 3-pin XLR female (the incoming audio cable uses the male connector) with pin 1 ground, pin 2 hot or positive, and pin 3 neutral or negative.

Likewise, for digital input signals, the input connector and signal format are not specified. Commonly used formats such as the AES 3-1992 2-channel interface are suggested. When multiple 2-channel inputs are used, the preferred channel assignment is:

- Pair 1: Left, Right
- Pair 2: Center, LFE

- Pair 3: Left surround, Right surround

7.2.1.1 Sampling Parameters

The AC-3 system conveys digital audio sampled at a frequency of 48 kHz, locked to the 27 MHz system clock [2]. If analog signal inputs are employed, the A/D converters should sample at 48 kHz. If digital inputs are employed, the input sampling rate should be 48 kHz, or the audio encoder should contain sampling rate converters that translate the sampling rate to 48 kHz. The sampling rate at the input to the audio encoder must be locked to the video clock for proper operation of the audio subsystem.

In general, input signals should be quantized to at least 16-bit resolution. The audio-compression system can convey audio signals with up to 24-bit resolution.

7.2.2 Output Signal Specification

Conceptually, the output of the audio encoder is an elementary stream that is formed into PES packets within the transport subsystem [2]. It is possible that digital television systems will be implemented wherein the formation of audio PES packets takes place within the audio encoder. In this case, the output of the audio encoder would be PES packets. Physical interfaces for these outputs (elementary streams and/or PES packets) may be defined as voluntary industry standards by SMPTE or other standards organizations.

7.3 Operational Details of the AC-3 Standard

The AC-3 audio-compression system consists of three basic operations, as illustrated in Figure 7.5 [4]. In the first stage, the representation of the audio signal is changed from the time domain to the frequency domain, which is a more efficient domain in which to perform psychoacoustically based audio compression. The resulting frequency-domain coefficients are then encoded. The frequency-domain coefficients may be coarsely quantized because the resulting quantizing noise will be at the same frequency as the audio signal, and relatively low S/N ratios are acceptable because of the phenomenon of psychoacoustic masking. Based on a psychoacoustic model of human hearing, a bit-allocation operation determines the actual SNR acceptable for each individual frequency coefficient. Finally, the frequency coefficients are coarsely quantized to the necessary precision and formatted into the audio elementary stream.

The basic unit of encoded audio is the AC-3 *sync frame*, which represents 1536 audio samples. Each sync frame of audio is a completely independent encoded entity. The elementary bit stream contains the information necessary to allow the audio decoder to perform the identical (to the encoder) bit allocation. This permits the decoder to unpack and dequantize the elementary bit-stream frequency coefficients, resulting in the reconstructed frequency coefficients. The synthesis filterbank is the inverse of the analysis filterbank, and it converts the reconstructed frequency coefficients back into a time-domain signal.

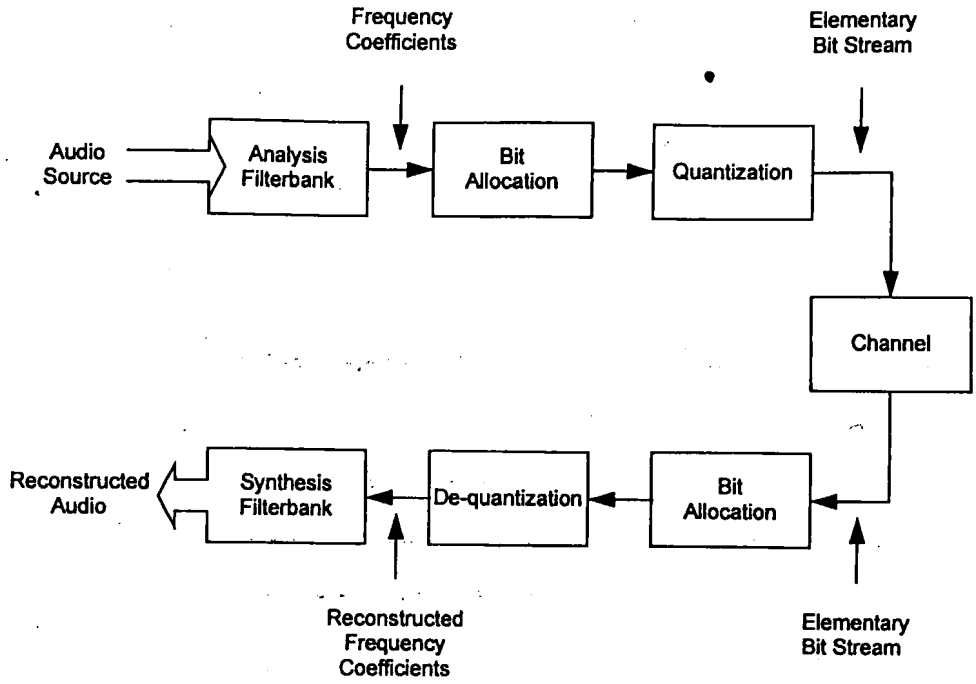

Figure 7.5 Overview of the AC-3 audio-compression system. (*From* [4]. *Used with permission.*)

7.3.1 Transform Filterbank

The process of converting the audio from the time domain to the frequency domain requires that the audio be blocked into overlapping blocks of 512 samples [4]. For every 256 new audio samples, a 512-sample block is formed from the 256 new samples and the 256 previous samples. Each audio sample is represented in two audio blocks, so the number of samples to be processed initially is doubled. The overlapping of blocks is necessary to prevent audible blocking artifacts. New audio blocks are formed every 5.33 ms. A group of six blocks is coded into one AC-3 sync frame.

7.3.1.1 Window Function

Prior to being transformed into the frequency domain, the block of 512 time samples is *windowed* [4]. The windowing operation involves a vector multiplication of the 512-point block with a 512-point window function. The window function has a value of 1.0 in its center, tapering down to almost zero at the ends. The shape of the window function is such that the overlap/add processing at the decoder will result in a reconstruction free of blocking artifacts. The window function shape also determines the shape of each individual filterbank filter.

7.3.1.2 Time-Division Aliasing Cancellation Transform

The analysis filterbank is based on the fast Fourier transform [4]. The particular transfc mation employed is the oddly stacked *time-domain aliasing cancellation* (TDAC) tran form. This particular transformation is advantageous because it allows removal of the 1(percent redundancy that was introduced in the blocking process. The input to the TDA transform is 512 windowed time-domain points, and the output is 256 frequency-doma coefficients.

7.3.1.3 Transient Handling

When extreme time-domain transients exist (an impulse, such as a castanets click), there i a possibility that quantization error—incurred by coarsely quantizing the frequency coeffi cients of the transient—will become audible as a result of *time smearing* [4]. The quantiza tion error within a coded audio block is reproduced throughout the block. It is possible fo the portion of the quantization error that is reproduced prior to the impulse to be audible Time smearing of quantization noise may be reduced by altering the length of the trans form that is performed. Instead of a single 512-point transform, a pair of 256-point trans forms may be performed—one on the first 256 windowed samples, and one on the last 25(windowed samples. A transient detector in the encoder determines when to alter the trans form length. The reduction in transform length prevents quantization error from spreading more than a few milliseconds in time, which is adequate to prevent audibility.

7.3.2 Coded Audio Representation

The frequency coefficients that result from the transformation are converted to a binary floating point notation [4]. The scaling of the transform is such that all values are smaller than 1.0. An example value in binary notation (base 2) with 16-bit precision would be:

0.0000 0000 1010 11002

The number of leading zeros in the coefficient, 8 in this example, becomes the *raw exponent.* The value is left-shifted by the exponent, and the value to the right of the decimal point (1010 1100) becomes the *normalized mantissa* to be coarsely quantized. The exponents and the coarsely quantized mantissas are encoded into the bit stream.

7.3.2.1 Exponent Coding

A certain amount of processing is applied to the raw exponents to reduce the amount of data required to encode them [4]. First, the raw exponents of the six blocks to be included in a single AC-3 sync frame are examined for block-to-block differences. If the differences are small, a single exponent set is generated that is usable by all six blocks, thus reducing the amount of data to be encoded by a factor of 6. If the exponents undergo significant changes within the frame, exponent sets are formed over blocks where the changes are not significant. Because of the frequency response of the individual filters in the analysis filterbank, exponents for adjacent frequencies rarely differ by more than ±2. To take advantage of this fact, exponents are encoded differentially in frequency. The first exponent is

encoded as an absolute, and the difference between the current exponent and the followin exponent then is encoded. This reduces the exponent data rate by a factor of 2. Finally where the spectrum is relatively flat, or an exponent set only covers 1 or 2 blocks, differen tial exponents may be shared across 2 or 4 frequency coefficients, for an additional saving of a factor of 2 or 4.

The final coding efficiency for AC-3 exponents is typically 0.39 bits/exponent (or 0.3 bits/sample, because there is an exponent for each audio sample). Exponents are codec only up to the frequency needed for the perception of full frequency response. Typically the highest audio frequency component in the signal that is audible is at a frequency lower than 20 kHz. In the case that signal components above 15 kHz are inaudible, only the first 75 percent of the exponent values are encoded, reducing the exponent data rate to less than 0.3 bits/sample.

The exponent processing changes the exponent values from their original values. The encoder generates a local representation of the exponents that is identical to the decoded representation that will be used by the decoder. The decoded representation then is used to shift the original frequency coefficients to generate the normalized mantissas that are subsequently quantized.

7.3.2.2 Mantissas

The frequency coefficients produced by the analysis filterbank have a useful precision that is dependent upon the word length of the input PCM audio samples as well as the precision of the transform computation [4]. Typically, this precision is on the order of 16 to18 bits, but may be as high as 24 bits. Each normalized mantissa is quantized to a precision from 0 to16 bits. Because the goal of audio compression is to maximize the audio quality at a given bit rate, an optimum (or near-optimum) allocation of the available bits to the individual mantissas is required.

7.3.3 Bit Allocation

The number of bits allocated to each individual mantissa value is determined by the bit-allocation routine [4]. The identical core routine is run in both the encoder and the decoder, so that each generates an identical bit allocation.

The core bit-allocation algorithm is considered *backward adaptive*, in that some of the encoded audio information within the bit stream (fed back into the encoder) is used to compute the final bit allocation. The primary input to the core allocation routine is the decoded exponent values, which give a general picture of the signal spectrum. From this version of the signal spectrum, a *masking curve* is calculated. The calculation of the masking model is based on a model of the human auditory system. The masking curve indicates, as a function of frequency, the level of quantizing error that may be tolerated. Subtraction (in the log power domain) of the masking curve from the signal spectrum yields the required S/N as a function of frequency. The required S/N values are mapped into a set of *bit-allocation pointers* (BAPs) that indicate which quantizer to apply to each mantissa.

7.3.3.1 Forward Adaptive

The AC-3 encoder may employ a more sophisticated psychoacoustic model than that used by the decoder [4]. The core allocation routine used by both the encoder and the decoder makes use of a number of adjustable parameters. If the encoder employs a more sophisticated psychoacoustic model than that of the core routine, the encoder may adjust these parameters so that the core routine produces a better result. The parameters are subsequently inserted into the bit stream by the encoder and fed forward to the decoder.

In the event that the available bit-allocation parameters do not allow the ideal allocation to be generated, the encoder can insert explicit codes into the bit stream to alter the computed masking curve, hence the final bit allocation. The inserted codes indicate changes to the base allocation and are referred to as *delta bit-allocation codes*.

7.3.4 Rematrixing

When the AC-3 encoder is operating in a 2-channel stereo mode, an additional processing step is inserted to enhance interoperability with Dolby Surround 4-2-4 matrix encoded programs [4]. This extra step is referred to as *rematrixing*.

The signal spectrum is broken into four distinct rematrixing frequency bands. Within each band, the energy of the left, right, sum, and difference signals are determined. If the largest signal energy is in the left and right channels, the band is encoded normally. If the dominant signal energy is in the sum and difference channels, then those channels are encoded instead of the left and right channels. The decision as to whether to encode left and right or sum and difference is made on a band-by-band basis and is signaled to the decoder in the encoded bit stream.

7.3.5 Coupling

In the event that the number of bits required to transparently encode the audio signals exceeds the number of bits that are available, the encoder may invoke *coupling* [4]. Coupling involves combining the high-frequency content of individual channels and sending the individual channel signal envelopes along with the combined coupling channel. The psychoacoustic basis for coupling is that within narrow frequency bands, the human ear detects high-frequency localization based on the signal envelope rather than on the detailed signal waveform.

The frequency above which coupling is invoked, and the channels that participate in the process, are determined by the AC-3 encoder. The encoder also determines the frequency banding structure used by the coupling process. For each coupled channel and each coupling band, the encoder creates a sequence of *coupling coordinates*. The coupling coordinates for a particular channel indicate what fraction of the common coupling channel should be reproduced out of that particular channel output. The coupling coordinates represent the individual signal envelopes for the channels. The encoder determines the frequency with which coupling coordinates are transmitted. If the signal envelope is steady, the coupling coordinates do not need to be sent every block, but can be reused by the

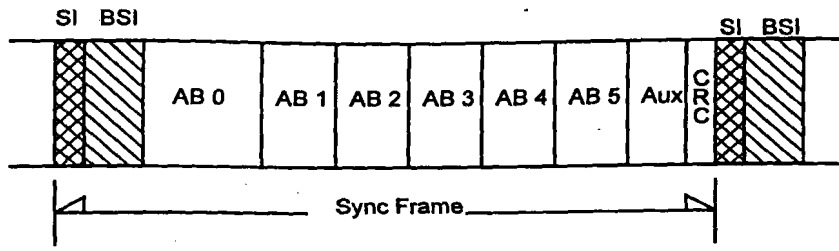

Figure 7.6 The AC-3 synchronization frame. (*From* [4]. *Used with permission.*)

decoder until new coordinates are sent. The encoder determines how often to send new coordinates, and it can send them as often as every block (every 5.3 ms).

7.3.6 Bit-Stream Elements and Syntax

An AC-3 serial-coded audio bit stream is made up of a sequence of *synchronization frames*, as illustrated in Figure 7.6 [4]. Each synchronization frame contains six coded audio blocks, each of which represent 256 new audio samples. A *synchronization information* (SI) header at the beginning of each frame contains information needed to acquire and maintain synchronization. A *bit-stream information* (BSI) header follows each SI, containing parameters describing the coded audio service. The coded audio blocks may be followed by an auxiliary data (Aux) field. At the end of each frame is an error-check field that includes a CRC word for error detection. An additional CRC word, the use of which is optional, is located in the SI header.

A number of bit-stream elements have values that may be transmitted, but whose meaning has been reserved. If a decoder receives a bit stream that contains reserved values, the decoder may or may not be able to decode and produce audio.

7.3.6.1 Splicing and Insertion

The ideal place to splice encoded audio bit streams is at the boundary of a sync frame [4]. If a bit-stream splice is performed at the boundary of the sync frame, the audio decoding will proceed without interruption. If a bit-stream splice is performed randomly, there will be an audio interruption. The frame that is incomplete will not pass the decoder's error-detection test, and this will cause the decoder to mute. The decoder will not find sync in its proper place in the next frame, and it will enter a sync search mode. After the sync code of the new bit stream is found, synchronization will be achieved, and audio reproduction will resume. This type of outage will be on the order of two frames, or about 64 ms. Because of the windowing process of the filterbank, when the audio goes to mute, there will be a gentle fadedown over a period of 2.6 ms. When the audio is recovered, it will fade up over a period of 2.6 ms. Except for the approximately 64 ms of time during which the audio is muted, the effect of a random splice of an AC-3 elementary stream is relatively benign.

7.3.6.2 Error-Detection Codes

Each AC-3 sync frame ends with a 16-bit CRC error-check code [4]. The decoder may use this code to determine whether a frame of audio has been damaged or is incomplete. Additionally, the decoder may make use of error flags provided by the transport system. In the case of detected errors, the decoder may try to perform error concealment, or it may simply mute.

7.3.7 Loudness and Dynamic Range

It is important for the digital television system to provide uniform subjective loudness for all audio programs [4]. Consumers often find it annoying when audio levels fluctuate between broadcast channels (observed when channel hopping) or between program segments on a particular channel (such as commercials being much louder than entertainment programs). One element found in most audio programming is the human voice. Achieving an approximate level match for dialogue (spoken in a normal voice, without shouting or whispering) in all audio programming is a desirable goal. The AC-3 audio system provides syntactical elements that make this goal achievable.

As of this writing, there is no regulatory limit as to how loud dialogue may be in an encoded bit stream. Because the digital audio-coding system can provide more than 100 dB of dynamic range, there is no technical reason for dialogue to be encoded anywhere near 100 percent, as it commonly is in NTSC television. However, there is no assurance that all program channels, or all programs or program segments on a given channel, will have dialogue encoded at the same (or even a similar) level. Without a uniform coding level for dialogue (which would imply a uniform headroom available for all programs), there would be inevitable audio-level fluctuations between program channels or even between program segments. These issues are addressed by the DTV audio standard and are described in Section 7.4.

7.3.7.1 Dynamic Range Compression

It is common practice for high-quality programming to be produced with wide dynamic range audio, suitable for the highest-quality audio reproduction environment [4]. Because they serve audiences with a wide range of receiver capabilities, however, broadcasters typically process audio to reduce its dynamic range. This processed audio is more suitable for most of the audience, which does not have an audio reproduction environment that matches the original audio production studio. In the case of NTSC, all viewers receive the same audio with the same dynamic range; it is impossible for any viewer to enjoy the original wide dynamic range of the audio production.

For DTV, the audio-coding system provides an embedded dynamic range control system that allows a common encoded bit stream to deliver programming with a dynamic range appropriate for each individual listener. A *dynamic range control value* (DynRng) is provided in each audio block (every 5 ms). These values are used by the audio decoder to alter the level of the reproduced sound for each audio block. Level variations of up to ± 24 dB may be indicated.

7.3.8 Encoding the AC-3 Bit Stream

Because the ATSC DTV standard AC-3 audio system is specified by the syntax and decoder processing, the encoder itself is not precisely specified [1]. The only normative requirement on the encoder is that the output elementary bit stream follow the AC-3 syntax. Therefore, encoders of varying levels of sophistication may be produced. More sophisticated encoders may offer superior audio performance, and they may make operation at lower bit rates acceptable. Encoders are expected to improve over time, and all decoders will benefit from encoder improvements. The encoder described in this section, although basic in operation, provides good performance and offers a starting point for future designs. A flow chart diagram of the encoding process is given in Figure 7.7.

7.3.8.1 Input Word Length/Sample Rate

The AC-3 encoder accepts audio in the form of PCM words [1]. The internal dynamic range of AC-3 allows input word lengths of up to 24 bits to be useful.

The input sample rate must be locked to the output bit rate so that each AC-3 sync frame contains 1536 samples of audio. If the input audio is available in a PCM format at a different sample rate than that required, sample rate conversion must be performed to conform the sample rate.

Individual input channels may be high-pass-filtered. Removal of dc components of the input signals can allow more efficient coding because the available data rate then is not used to encode dc. However, there is the risk that signals that do not reach 100 percent PCM level before high-pass filtering will exceed the 100 percent level after filtering, and thus be clipped. A typical encoder would high-pass filter the input signals with a single pole filter at 3 Hz.

The LFE channel normally is low-pass-filtered at 120 Hz. A typical encoder would filter the LFE channel with an 8^{th}-order elliptic filter whose cutoff frequency is 120 Hz.

Transients are detected in the full-bandwidth channels to decide when to switch to short-length audio blocks to improve pre-echo performance. High-pass-filtered versions of the signals are examined for an increase in energy from one subblock time segment to the next. Subblocks are examined at different time scales. If a transient is detected in the second half of an audio block in a channel, that channel switches to a short block.

The transient detector is used to determine when to switch from a *long transform block* (length 512) to a *short transform block* (length 256). It operates on 512 samples for every audio block. This is done in two passes, with each pass processing 256 samples. Transient detection is broken down into four steps:

- High-pass filtering
- Segmentation of the block into submultiples
- Peak amplitude detection within each subblock segment
- Threshold comparison

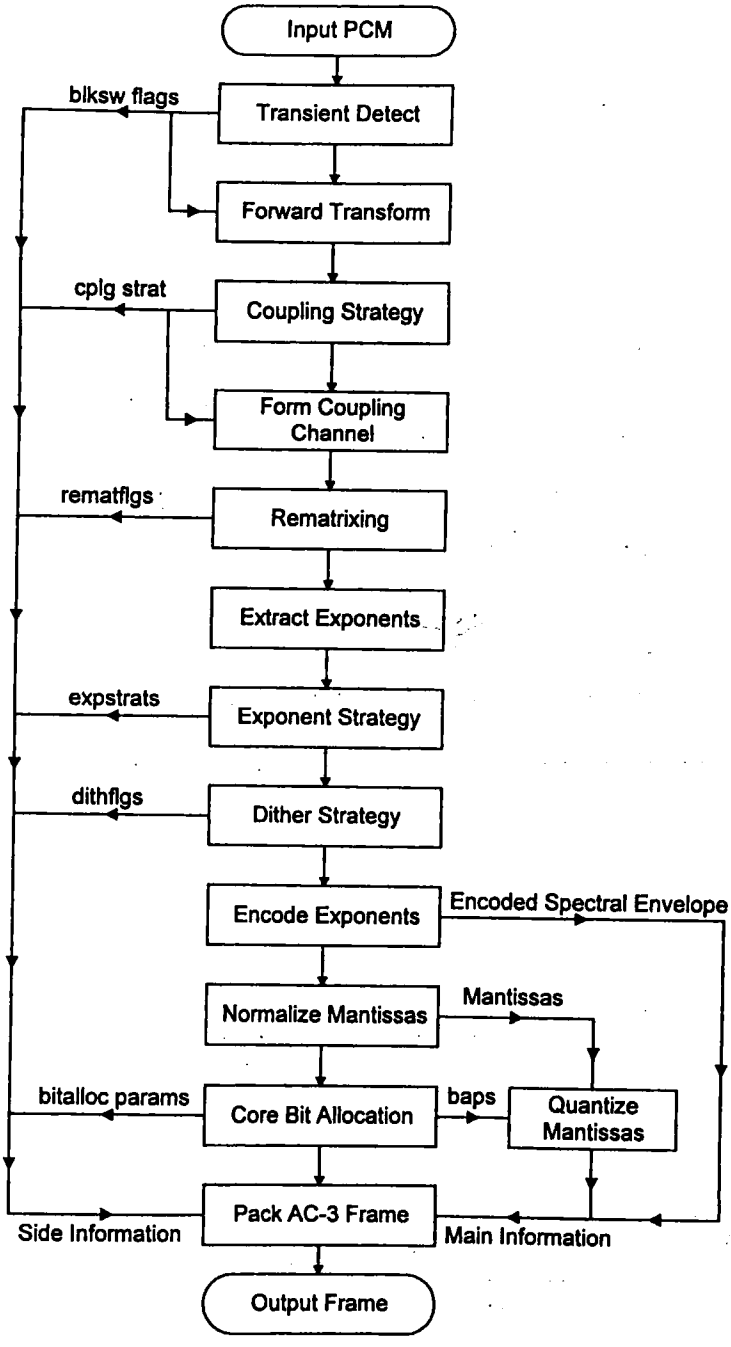

Figure 7.7 Generalized flow diagram of the AC-3 encoding process. (*From* [5]. *Used with permission.*)