AC-3: Flexible Perceptual Coding for Audio Transmission and Storage

**Craig C. Todd, Grant A. Davidson, Mark F. Davis,
Louis D. Fielder, Brian D. Link, Steve Vernon**

**Dolby Laboratories
San Francisco**

0. Abstract

Dolby AC-3 is a flexible audio data compression technology capable of encoding a range of audio channel formats into a low rate bit stream. Channel formats range from monophonic to 5.1 channels, and may include a number of associated audio services. Based on a transform filter bank and psychoacoustics, AC-3 includes the novel features of transmission of a variable frequency resolution spectral envelope and hybrid backward/forward adaptive bit allocation.

1. Introduction

The genesis of the AC-3 technology came from a desire to provide superior multi-channel sound localization for High Definition Television sound. In the United States, the High Definition Television (HDTV) standardization process formally began in 1987 with the formation, by the Federal Communications Commission (FCC), of the Advisory Committee on Advanced Television Service (ACATS). The initial HDTV system proposals called for analog based picture transmission and digital sound transmission. The initially proposed sound systems matrix-encoded a multi-channel audio source into a stereo pair, which was then digitally encoded with Dolby AC-1, a low cost delta-modulation based coding technology. At the receiver, the two audio channels would optionally be decoded into four channels using a matrix decoder. This proposal basically used the 4-2-4 multi-channel matrix system as a 2-1 bit-rate reduction system, since the matrix technology reduced the bit-rate required to convey four channel audio by a factor of 1/2. By 1989, advances in audio coding technology and DSP hardware made it possible to move from the AC-1 audio coding technology to the transform based AC-2 coding technology, simultaneously raising audio quality while reducing bit-rate. The use of the multi-channel matrix technology remained.

By 1990, there were suggestions by some that the limitations of the 4-2-4 matrix technology should perhaps be avoided in HDTV, and that four discrete audio channels should be transmitted. Others felt that the small gain in multi-channel performance was not worth the required doubling of the bit-rate. It was at this point that the concept of AC-3 was born: a fundamentally multi-channel audio coder, operating at approximately the same bit-rate required by two independently coded audio channels, while offering multi-channel audio performance without the limitations of the 4-2-4 channel matrix system. Basically, it was realized that the bit-rate reduction process that was being performed by the 4-2-4 matrix system could be better performed in the coder itself, and not by an addition of the matrix

technology to a two channel coder. If successful, the concept of AC-3 would allow HDTV to gain the benefit of going to discrete audio transmission, without paying the price of a corresponding doubling of the required bit-rate.

While conceived in response to a need for HDTV, the concept of AC-3 was first realized in response to a similar need in the cinema. By 1990, a 4-2-4 matrix based analog sound system had been in place in the 35 mm cinema for some 14 years, and interest was growing to offer the cinema industry new digital sound technology. In 1989, a SMPTE subgroup studied the issue of how many sound channels a new digital film sound system should offer. The conclusion was that 5.1 channels should be provided (left, center, right, left surround, right surround, subwoofer), which is identical to the 70 mm split surround format which has been in use in the cinema since 1979. In order to place digital sound data on film reliably, and yet not interfere with either the picture or analog sound area of the film, the available data rate is limited. It was determined that 320 kb/s of error corrected audio data could be reliably placed and extracted from the film area between the sprocket hole perforations on one side of the 35 mm film. All that was required was to realize the concept of AC-3 as a 5.1 channel audio coder operating at 320 kb/s.

Due to the rapid development schedule required to quickly realize a commercial cinema product, AC-3 was first implemented as a real-time system running on multiple DSP cards. The first cinema industry demonstrations began in May of 1991. By Dec. of 1991, the first AC-3 coded digital film, *Star Trek VI*, played in three theatres. The formal roll out of the Dolby SR•D system (as it is called) was in June of 1992, with release of the film *Batman Returns*.

In mid 1991 the existence of the AC-3 audio coding system was publicly disclosed, and was eagerly embraced by the HDTV audio community in the United States. The ITU BR (formerly CCIR) Task Group 10-1 met in June of 1991 and accepted the basic 5 channel audio format, making it the basis for a recommendation. In February of 1992 the U.S. Advanced Televisions Systems Committee released a document formally recommending composite coded 5.1 channel audio for the U.S. HDTV service. In Oct. of 1992, TG 10-1 accepted the 0.1 low frequency channel and modified the draft recommendation accordingly. In 1993 the AC-3 system underwent subjective testing in the United States in order to evaluate its suitability for inclusion in the HDTV system being proposed by the *Grand Alliance* (a consortium of the remaining U.S. HDTV proponents which has been authorized to collaborate on the U.S. HDTV broadcast system). In Oct. 1993 the Grand Alliance recommended the use of Dolby AC-3. In Nov. 1993, the full ACATS committee formally approved the use of AC-3 by the Grand Alliance HDTV system. Final system tests are expected to be completed by the end of 1994, final FCC approval should occur in 1995, and some initial broadcasts may begin in 1996. While the U.S. HDTV process will take some additional time to complete, AC-3 is expected to begin a widespread roll out in cable television and direct broadcast satellite equipment by the end of 1994.

The requirements which AC-3 had to meet to satisfy the cinema application are straightforward. AC-3 only had to support one bit-rate (320 kb/s) and audio coding mode (5.1 channels). All cinemas in which AC-3 is used have a full complement of loudspeaker channels, so mixdown to fewer than 5.1 channels is not an issue. Audio levels are standardized in the cinema industry, and all playback is done with the full intended dynamic range.

On the other hand, there are a diverse set of requirements for a coder intended for widespread application. Our intent was to make AC-3 into a universally applicable low bit rate coder by satisfying many diverse requirements. This allows AC-3 to deliver an audio signal in a form usable by an entire audience, even though different members of the audience have different needs. While the most critical members of the audience may be anticipated to have complete multi-channel reproduction systems, most of the audience may be listening in mono or stereo. Some of the audience may have matrix based multi-channel reproduction equipment without a discrete input, thus requiring a 2-channel matrix encoded (and generally not mono-compatible) output from the AC-3 decoder. Most of the audience welcomes a restricted dynamic range reproduction, while a few in the audience will wish to experience the full dynamic range of the original signal. The visually and hearing impaired wish to be served. All of these (and other) diverse needs were considered early in the design of the AC-3 coding technology. Techniques to satisfy these needs have been designed in from the beginning, and not added on as an afterthought.

2. Bit Allocation Philosophy

The challenge of bit-rate reduction of audio signals is, of course, to remove bits from a representation of the audio signal in a manner that is inaudible (to humans). Due to the frequency masking properties of human hearing, the best representation of audio to use is a frequency domain representation, obtained by the use of a sub-band or transform filter bank. Audio coding technology has developed to the point that the issue is not really what bits may be removed, but rather which few bits to leave in (or allocate to) the various frequency components of the audio signal. Since the philosophy behind bit allocation is so fundamental to the audio coder design, we begin our exploration of the AC-3 coder with a discussion of the philosophy behind its bit allocation algorithm. This algorithm, along with simple economics, helps to determine the optimum filter bank. There are two broad classes of bit allocation strategy: forward adaptive; and backwards adaptive. Forward adaptive (fig. 1) refers to the method where the encoder calculates the bit allocation and explicitly codes the allocation into the coded bit stream. In theory, this method allows for the most accurate allocation since the encoder has full knowledge of the input signal and, in principle, may be of unlimited complexity. The encoder may precisely calculate an optimum bit allocation within the limits of the psychoacoustic model employed. An attractive feature of forward adaptation is that since the psychoacoustic model is resident only in the encoder, the model may be changed at any time with no impact on an installed base of decoders.

While forward adaptation has some attractive features, there are practical limitations to the performance which can be obtained with this technique. The performance limitations comes from the need to utilize a portion of the available bit-rate to deliver the explicit bit allocation to the decoder. For instance, the ISO MPEG1 layer II audio coder requires a data rate of nearly 4 kb/s/ch to transmit the bit allocation with time resolution of 24 msec and frequency resolution of 750 Hz. During transient conditions it is beneficial to be able to alter the bit allocation with a much finer time resolution but, of course, that would require a significantly greater allocation data rate (nearly 12 kb/s/ch for 8 msec time resolution in the case of the LII coder).

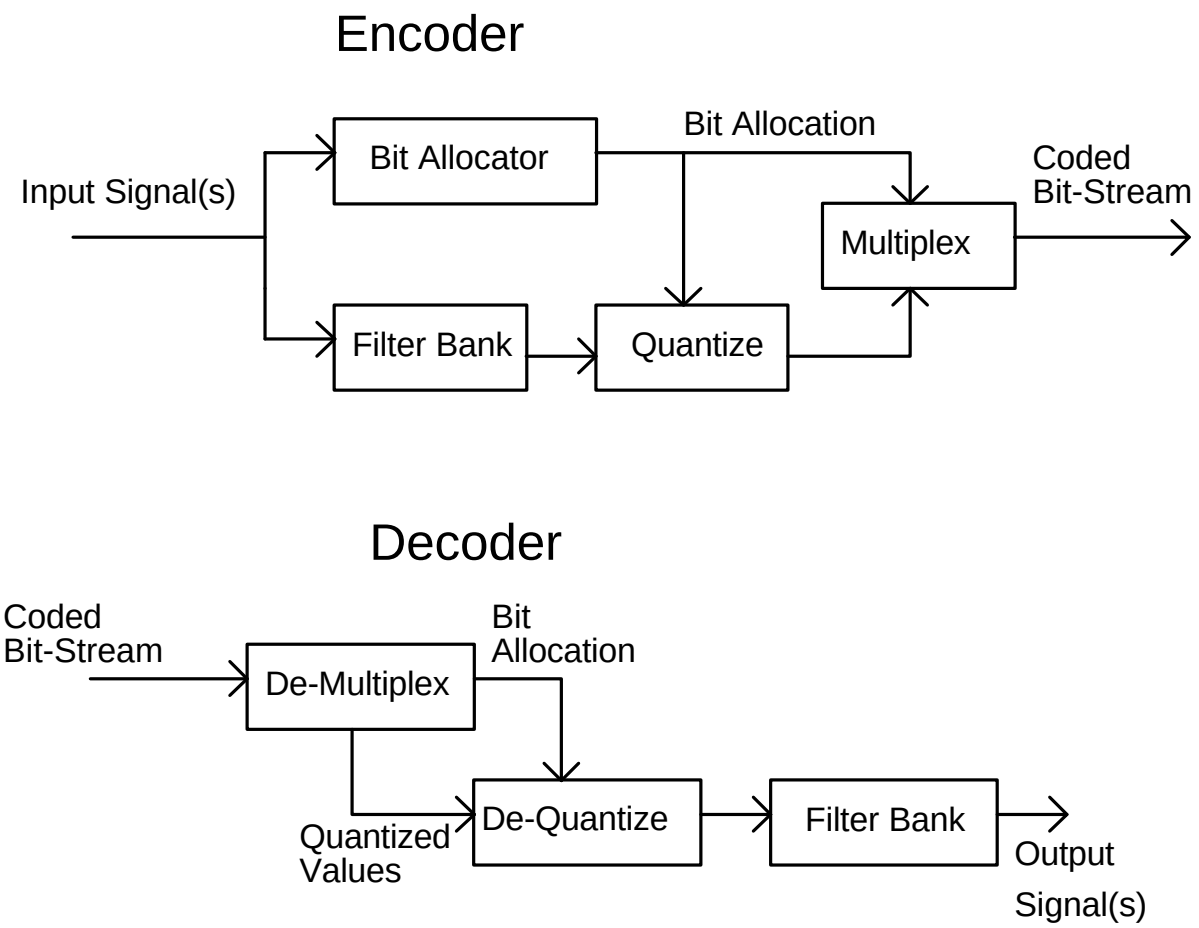

Figure 1: Forward Adaptive Bit Allocation

During steady state conditions it is useful to be able to allocate bits with much greater frequency resolution. For example, in the case of a signal with a spectrum which contains spectral lines in every 750 Hz frequency band, all bands would have to be allocated bits and very limited bit-rate reduction may be obtained before audible degradation occurs. If bits can be allocated on a finer frequency grid, then even spectrally dense signals may have bits removed at frequencies between the spectral lines, and more useful bit-rate reduction may be obtained. Like an improvement in time resolution, an improvement in the frequency resolution of the bit allocation would also require a substantial increase in the allocation data rate (an increase by a factor of 8 in order to improve the frequency resolution by a factor of 8). While attractive in theory, forward adaptive bit allocation clearly does impose significant practical limitations on performance at very low bit-rates.

Backward adaptive bit allocation (fig. 2) refers to the creation of the bit allocation from the coded audio data itself, without explicit information from the encoder. The advantage of this approach is that none of the available data rate is used to deliver the allocation information to the decoder, and thus all of the bits are available to be allocated to coding audio. The allocation may have time or frequency resolution equal to the information used to generate the allocation. Backward adaptive systems are thus more efficient in transmission, and allow the bit allocation to have superior time and frequency resolution. The disadvantages of backward adaptive allocation come from the fact that the bit allocation

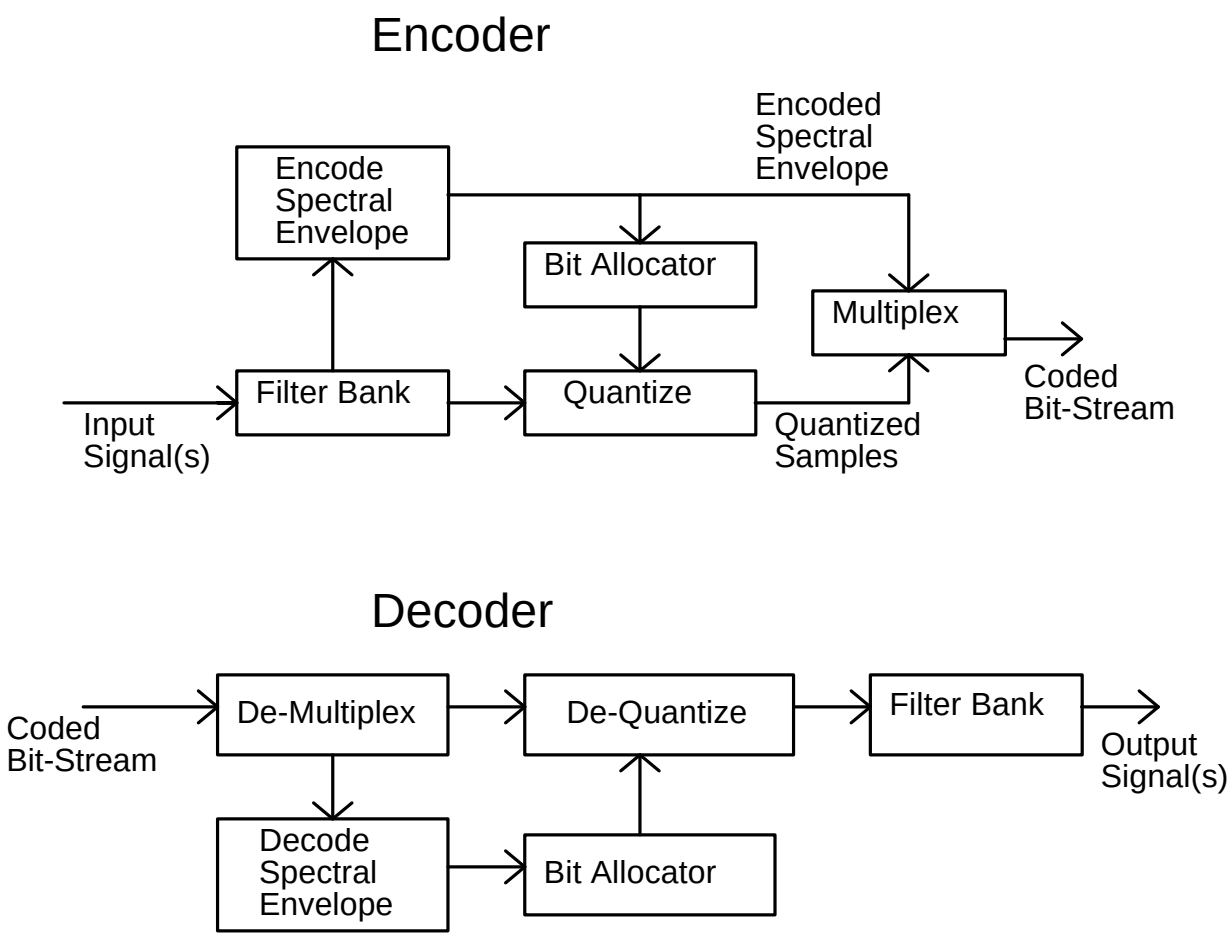

Figure 2: Backward Adaptive Bit Allocation

must be computed in the decoder from information contained in the bit stream. The bit allocation is computed from information with limited accuracy, and may contain small errors. Since the bit allocation is intended to be able to be performed in a low-cost decoder, the computation can not be overly complex, or else decoder cost would not be low. Since the bit allocation algorithm available to the encoder is fixed once decoders are deployed into the field, the psychoacoustic model may not be updated.

The AC-3 coder makes use of hybrid backward/forward adaptive bit allocation, in which most of the disadvantages of backward adaptation are removed. This method involves a core backward adaptive bit allocation routine which runs in both the encoder and the decoder. The core routine is relatively simple, but is based on a psychoacoustic model and, in general, is quite accurate. The input to the core routine is the spectral envelope, which is part of the encoded audio data delivered to the decoder. With AC-3, the spectral envelope has time and frequency resolution equal to that of the analysis/synthesis filter bank.

There are two aspects to the forward adaptation: psychoacoustic model parameter adjustment, and delta bit allocation. The core bit allocation routine employs a psychoacoustic model which has certain assumptions about the masking properties of signals. Certain parameters of the model are explicitly sent within the AC-3 bit stream. Thus, the details of the actual psychoacoustic model implemented in the decoder may be adjusted by the encoder. The encoder can perform a bit allocation based on any

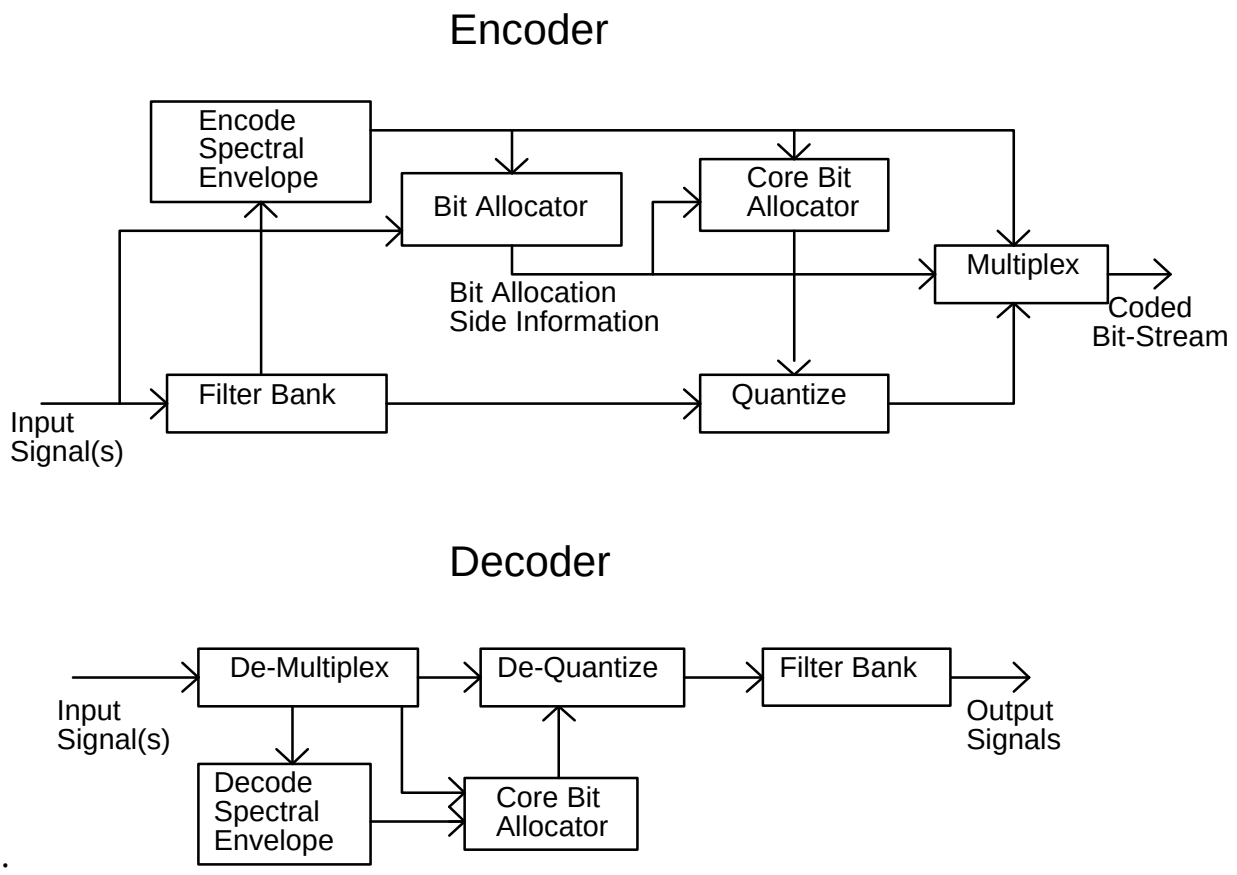

Figure 3: Hybrid Backward/Forward Adaptive Bit Allocation

psychoacoustic model of any complexity, and compare the results to the bit allocation based on the core routine used by the decoder. If a better match can be made to an ideal allocation by altering some of the parameters used by the core routine, the encoder can do so. If there is a condition where it is not possible to approach the ideal allocation by means of parameter variation, the encoder can then explicitly send some allocation information. Since the allocation computed by the core routine should be very close to ideal, only small deltas to the default allocation are ever required. The AC-3 data syntax allows the encoder to explicitly send delta bit allocation information which will cause the bit allocation in small frequency regions to be increased or decreased. Since the final bit allocation used by the encoder and decoder must be identical the final allocationactually used by the encoder is the decoder core routine with whatever parameter variations and delta bit allocation are in effect.

3. Filter Bank

The choice of analysis/synthesis filter bank to employ in an audio coder is a trade-off between frequency resolution, time resolution, and cost, where cost is measured in random access memory bits (RAM) and multiply/accumulate cycles (MACs). Steady state audio signals benefit from finer frequency resolution, whereas transient signals require finer time resolution. It is not possible to simultaneously achieve high time and frequency resolution, so a compromise must be reached. Fine frequency resolution has a very real cost, in that longer blocks of audio must be buffered which requires larger amounts of RAM. Cost of RAM dominates in the single chip decoders which are demanded by the market. Very fine

frequency resolution can allow somewhat higher coding gains to be achieved, but requires significantly more costly RAM.

AC-3 makes use of the oddly stacked time-division aliasing cancellation filter bank described by Princen and Bradley. Overlapping blocks of 512 windowed samples are transformed into 256 frequency domain points. The filter bank is critically sampled, and of low complexity, requiring only 13 MACs for computation. A proprietary 512 point Fielder window is used to achieve the best trade-off between close-in frequency selectivity and far-away rejection. Each transform block is formed from audio representing 10.66 msec (at 48 kHz sample rate), although the transforms are performed every 5.33 msec. The audio block rate is thus 187.5 Hz. During transient conditions where finer time resolution is useful, the block size is halved so that transforms occur every 2.67 msec. The low 13 MAC computation rate is maintained during transient block size switches. The frequency resolution of the filter bank is 93.75 Hz. The minimum time resolution is 2.67 msec. The full resolution of the filter bank is used; the individual filters are not combined into wider bands (or critical bands), except during a portion of the core bit allocation routine. Bit allocation can occur down to the individual transform coefficient level, with neighboring coefficients receiving different allocations.

4. Spectral Envelope

Each individual transform coefficient is coded into an exponent and a mantissa. The exponent allows for a wide dynamic range, while the mantissa is coded with a limited precision which results in quantizing noise. The synthesis filter bank in the decoder constrains the quantizing noise to be at nearly the same frequency as the quantized signal.

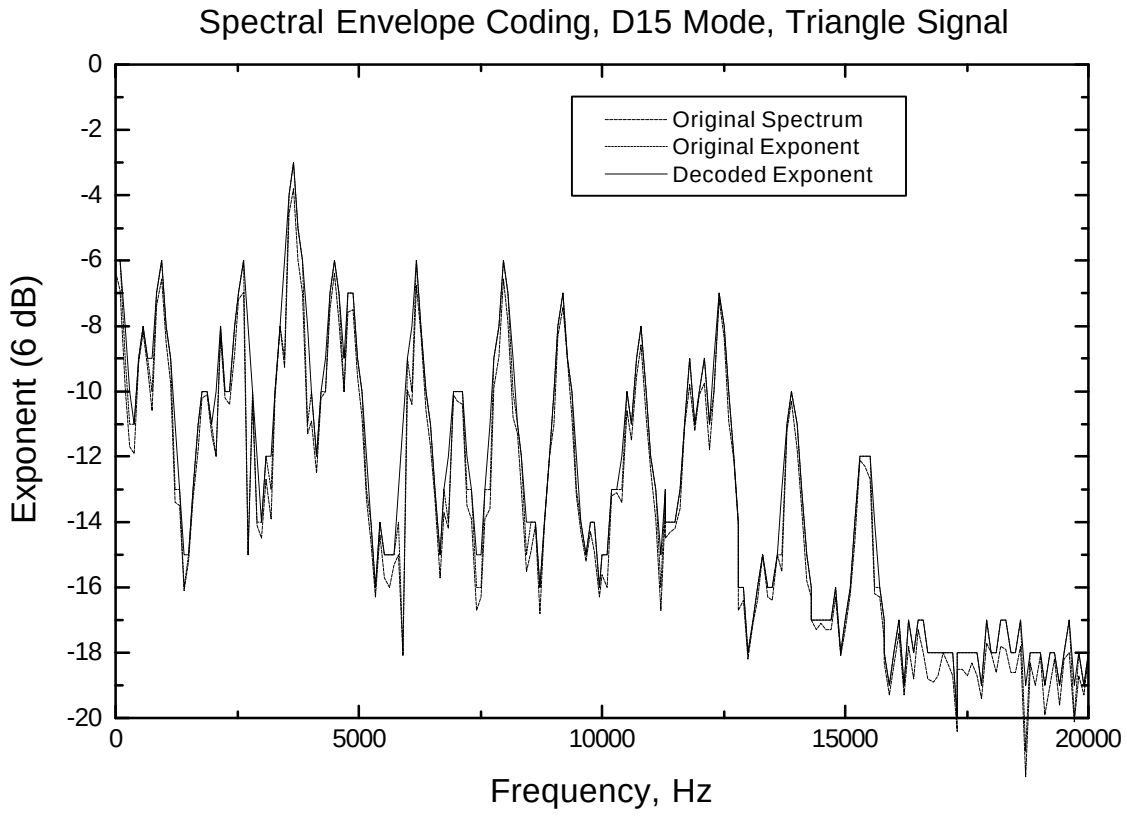

Figure 4: D15 Spectral Envelope (Exponent) Coding

The set of coded exponents forms a representation of the overall signal spectrum, and is referred to as the spectral envelope.

The filters of the transform filter bank are not perfect brick wall filters, but have gentle slopes. In general, the response falls off at approximately 12 dB per adjacent filter, or per 93.75 Hz. If the signal spectrum is analyzed with the filter bank, the level in any two adjacent filters seldom exceeds 12 dB. This fact can be used to advantage in coding the spectral envelope. The AC-3 coder encodes the spectral envelope differentially in frequency. Since deltas of at most ±2 (a delta of 1 represents a 6 dB level change) are required, each exponent can be coded as one of 5 changes from the previous (lower in frequency) exponent: +2, +1, 0, -1, -2. The first exponent (D.C. term) is sent as an absolute, and the rest of the exponents are sent as differentials. Groups of three differentials are coded into a 7 bit word. Each exponent thus requires approximately 2.33 bits to code. This method of transmitting the exponents is referred to as *D15*. When employed, the D15 coding provides a very accurate spectral envelope, as indicated in fig. 4.

Coding all exponents with 2.33 bits each for every individual audio block would be extravagant. However, fine frequency resolution is only required for relatively steady signals, and for these signals the spectral envelope will remain relatively constant over many blocks. The fine resolution D15 spectral envelope is only sent when the spectrum is relatively stable, and in that case the estimate may be sent only occasionally. In the typical case, the spectral envelope is sent once every 6 audio blocks (32 msec), in which case the data rate required is < 0.39 bits / exponent. Since each individual frequency point has an exponent, and there is one frequency sample for each time sample (the TDAC filter bank is

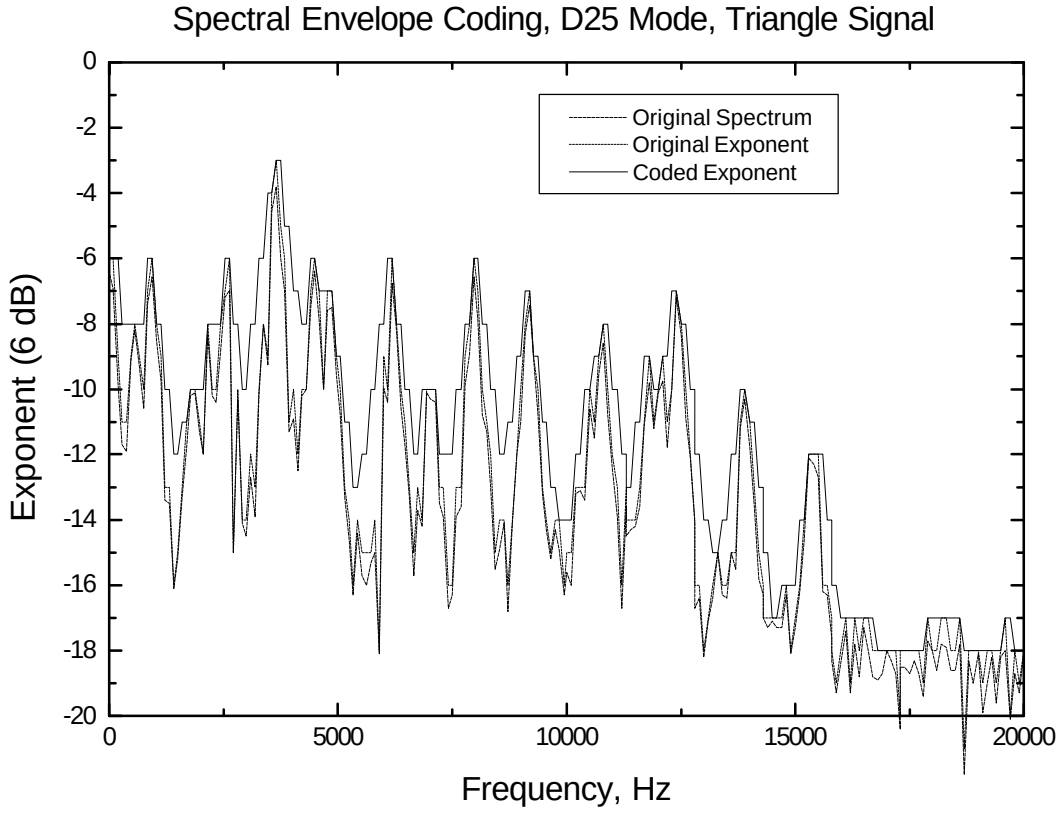

Figure 5: D25 Spectral Envelope (Exponent) Coding

critically sampled), the D15 high resolution spectral envelope requires < 0.39 bits per audio sample.

When the spectrum of the signal is not stable, it is beneficial to send a spectral estimate more often. In order to keep the data overhead for the spectral estimate from becoming excessive, the spectral estimate may be coded with less frequency resolution. Two additional methods are available. The medium frequency resolution method is *D25*, where a delta is transmitted for every other frequency coefficient. This method requires half of the data rate of the D15 method (or 1.l6 bits / exponent), and has a factor of two worse frequency resolution. The D25 method is typically used when the spectrum is relatively stable over 2-3 audio blocks and then undergoes a significant change. Use of the D25 method does not allow the spectral envelope to accurately follow all of the troughs in a very tonal spectrum. The coded spectral envelope is, of course, forced to cover all of the peaks. This is illustrated in fig. 5, which shows the result of applying the D25 method to the same triangle signal shown in fig. 4 (although with this type of signal the D25 method would generally not be used).

The final method is *D45*, where a delta is transmitted for every 4 frequency coefficients. This method requires one fourth of the data rate of the D15 method (or 0.58 bits / exponent), and is typically used during transients for single audio blocks (5.3 msec). The transmitted spectral envelope thus has very fine frequency resolution for steady state (or slowly changing signals), and has fine time resolution for transient signals. Typically, transient signals do not require fine frequency resolution since by their nature transients are

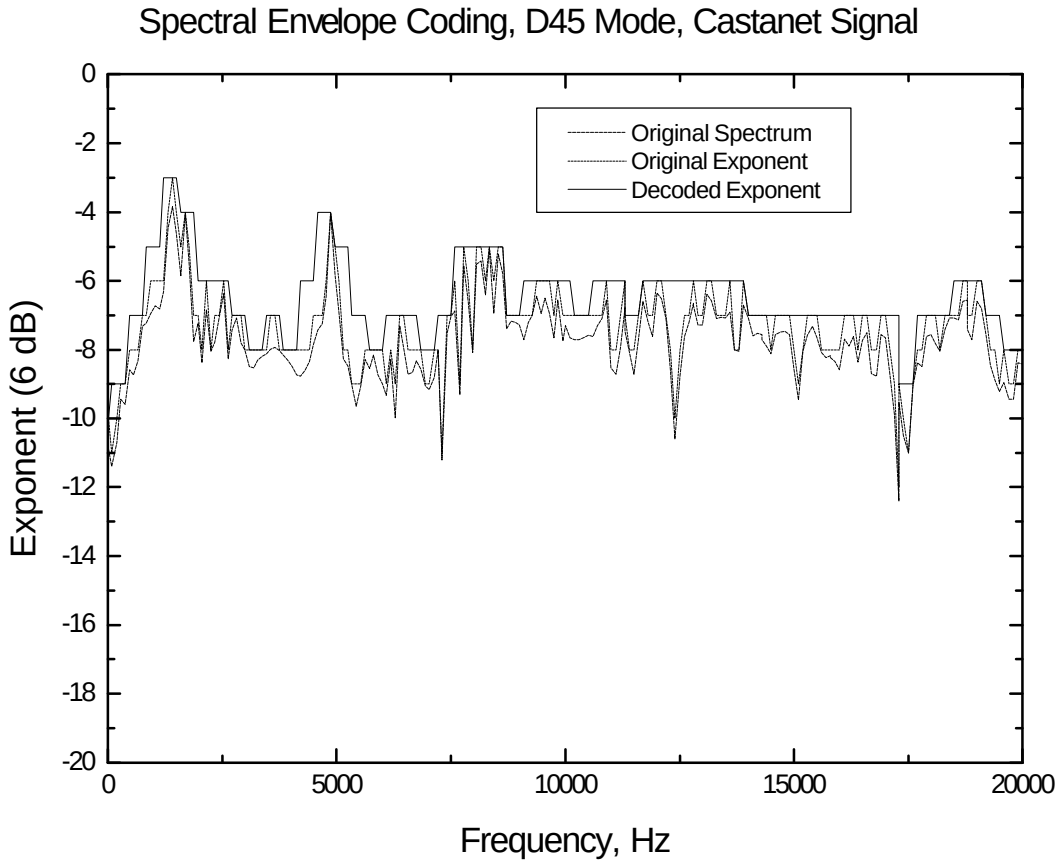

Figure 6: D45 Spectral Envelope (Exponent) Coding

wide band signals. Figure 6 shows the result of coding a castanet transient with the D45 method.

The AC-3 encoder is responsible for selecting which exponent coding method to use for any audio block. Each coded audio block contains a 2-bit field called *exponent strategy*. The four possible strategies are: D15, D25, D45, and REUSE. For most signal conditions, a D15 coded exponent set is sent during the first audio block in a frame, and the following 5 audio blocks reuse the same exponent set. During transient conditions, exponents are sent more often. The encoder routine responsible for choosing the optimum exponent strategy may be improved or updated at any time. Since the exponent strategy is explicitly coded into the bit stream, all decoders will track any change in the encoder.

5. Bit Allocation Details

The AC-3 core bit allocation routine begins with the decoded spectral envelope, or exponents, which are considered to be the power spectral density (psd) of the signal. There may be as many as 252 psd values (the number can vary depending on the number of exponents which have been sent, which in turn depends on the desired audio bandwidth and the audio sampling rate). The major portion of the bit allocation routine is the convolution of a spreading function matching the ear's masking curve, against the power spectral density. In order to reduce the computational load, the original psd array is converted to a smaller banded psd array. At low frequencies, the band size is 1, and at high frequencies the band size is 16. The bands increase in size proportional to the widening of the ear's critical bands. The initial (up to) 252 psd values are combined to form 64 banded psd values.

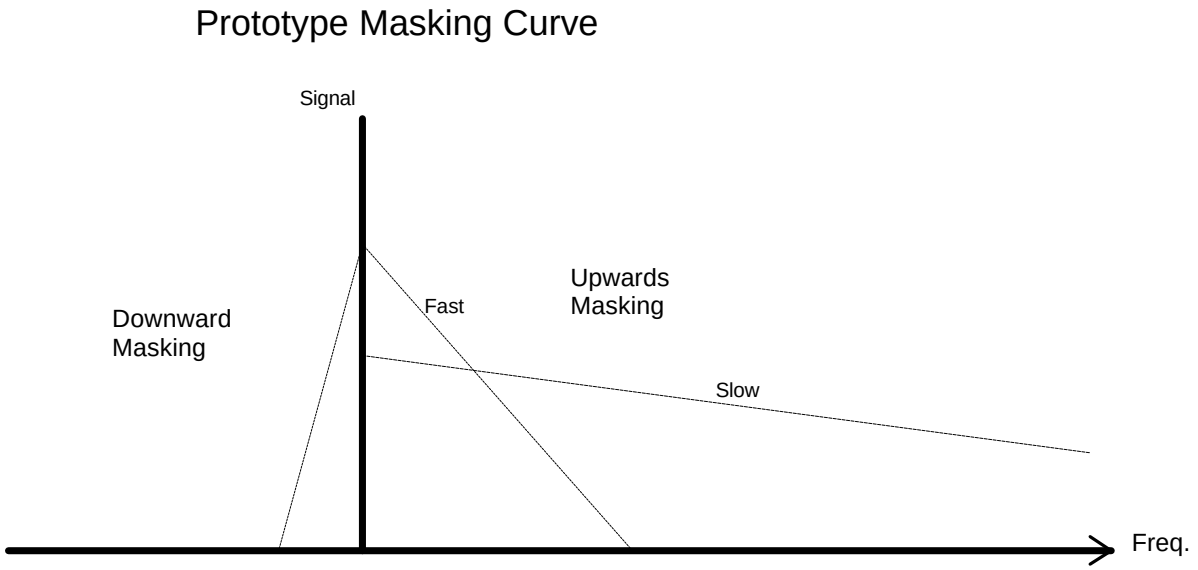

Figure 7: Prototype Masking Curve

A simplified technique has been developed to perform step of convolving the spreading function against the banded psd. The spreading function of the ear is approximated by three curves: a very steeply decaying downwards masking curve; a fast decaying upwards masking curve; and a slowly decaying upwards masking curve which is offset downwards in level. The technique ignores the downwards masking curve for simplicity at the expense of occasional over allocation. A simplified convolution of the two different sloping upwards masking curves against the banded psd is performed. The calculation begins at the lowest

frequency of the banded psd array and moves upwards in frequency. Two running computations are performed, one for each of the fast decaying and slowly decaying masking curves, which we refer to as the fast leak and the slow leak. The calculation is performed in the log domain, where a decay (leak) can be implemented as a simple decrement. As the calculation is moved up in frequency band by band, each new banded psd is examined. If the psd in the new band is significant with respect to the current leak values, the new psd is combined into the leak terms which are increased in value. If the new psd is insignificant, the old leak terms are decremented, or allowed to leak down. The largest of each leak term at each frequency is held. The result of this calculation is an array indicating the predicted masking, band by band.

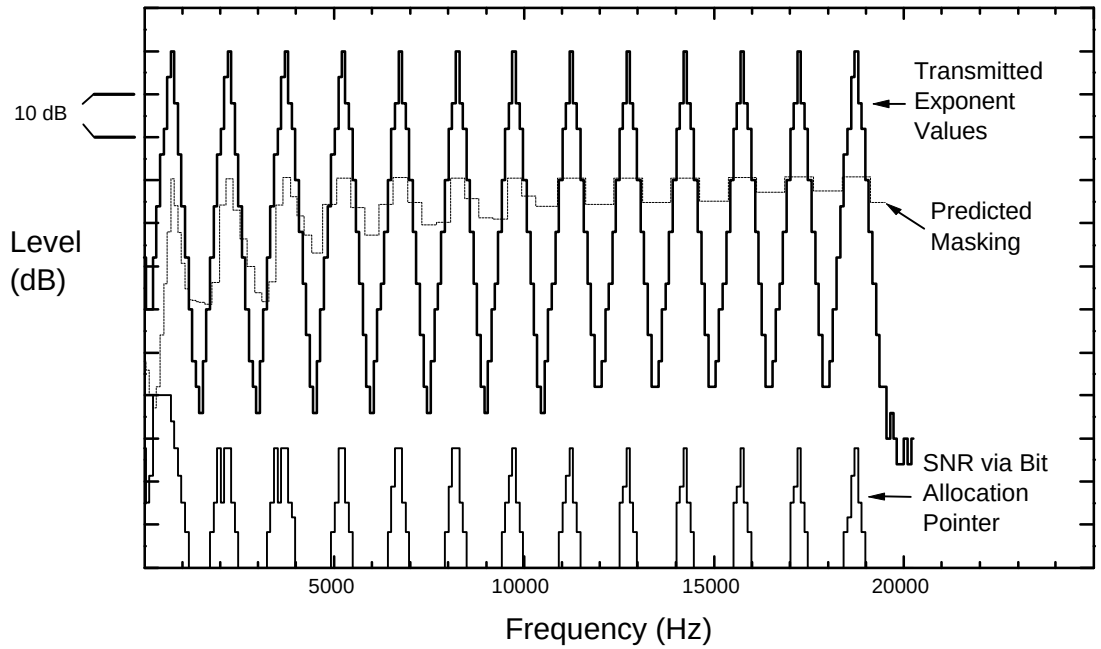

Figure 8: Bit Allocation Computation

This curve is compared against a hearing threshold, and the larger of the two values is held. The final step is to subtract the predicted masking curve from the original (unbanded) psd array to determine the desired SNR for each individual transform coefficient. This is shown in figure 8. The array of SNR values is converted to an array of bit allocation pointers (baps). The bap array values point to the quantizers to be used for each transform coefficient mantissa. At this point the encoder does a bit count to determine if the bit allocation has used up the available number of bits. All available bits are from a common bit pool, available to all of the channels. If more bits are available, the individual mantissa SNR's may be increased, until all of the bits are used up. If too many bits have been allocated, the individual mantissa SNR's may be decreased, and/or coupling (see next section) may be invoked. In the example shown (a difficult multi-tone test signal with tones at all odd multiples of 750 Hz) the AC-3 coder requires approximately 74 kb/s (including all side information) to code the signal to near transparency. This low bit-rate is achieved because the frequency resolution of the filter bank, along with the backward/forward bit allocation, is able to remove bits from in between the tones. As a matter of comparison, this test signal would fully occupy all bands of the LII filter bank, and the LII coder would require more than twice the bit-rate of AC-3 to code this signal to the same quality level.

6. Coupling

Even though the coding techniques employed by AC-3 are very powerful, when the coder is operated at very low bit rates there are signal conditions under which the coder would run out of bits. When this occurs, a technique which we refer to as *coupling* is invoked. Coupling takes advantage of the way in which the ear determines directionality for very high frequency signals, in order to allow a reduction in the amount of data necessary to code a high quality multi-channel audio signal.

At high audio frequencies (above approximately 2 kHz) the ear is physically unable to detect individual cycles of an audio waveform, and instead responds to the envelope of the waveform. Directionality is determined by the inter-aural time delay of the signal envelope, and by the perceived frequency response which is affected by head shadowing and the ear's pinnae. Coupling takes advantage of the fact that the ear is not able to independently detect the direction of two high frequency signals which are very closely spaced in frequency.

When the AC-3 coder becomes starved for bits, channels may be selectively coupled at high frequencies. The frequency at which coupling begins is called the *coupling frequency*. Above the coupling frequency the coupled channels are combined into a *coupling* (or common) channel. Care is taken with the phase of the signals to be combined so as to avoid signal cancellations. The encoder measures the original signal power of the individual input channels in narrow frequency bands, as well as the power in the coupled channel in the same frequency bands. The encoder generates *coupling coordinates* for each individual channel which indicate the ratio of the original signal power within a band to the coupling channel power in the band.

The coupling channel is encoded in the same manner as the individual channels; there is a spectral envelope of coded exponents and a set of quantized mantissas. The channels which are included in coupling are sent discretely up to the coupling frequency. Above the coupling frequency, only the coupling coordinates are sent for the coupled channels. The decoder multiplies the individual channel coupling coordinates by the coupling channel coefficients to regenerate the high frequency coefficients of the coupled channels.

The coupling process is audibly successful because the reproduced sound field is a close power match to the original. Within each narrow frequency band, the reproduced power level out of each loudspeaker matches the original signal power in that loudspeaker. Additionally, the level within each narrow frequency band of each source channel is reproduced at the same overall power in the sound field, even though the power may be shared amongst several of the loudspeaker channels. The coupling coordinates are encoded with an accuracy of < 0.25 dB. This fine resolution not only allows a good power match to be obtained but, more importantly, allows small changes to be made smoothly. (This is in contrast to techniques employed by other coding technologies to combine high frequencies together, where the minimum gain step is 2 dB.)

The AC-3 encoder is responsible for determining the coupling strategy. The encoder controls which of the audio channels are to be included in coupling, and which remain completely independent. The encoder controls at what frequency coupling begins, the coupling band structure (the bandwidths of the coupled bands), and when new coupling coordinates are sent. The coupling strategy routine may be altered or improved at any time and, since the coupling strategy information is explicit in the encoded bit stream, all decoders will follow the changes. For example, early AC-3 encoders often coupled

channels with a coupling frequency as low as 3.5 kHz. As AC-3 encoding techniques have improved, the typical coupling frequency has increased to 10 kHz.

7. Bit Stream Syntax

The fundamental time unit for AC-3 is related to the transform block size. While each transformed block is 512 samples long, the 100% overlap/add of the TDAC transform means blocks are transformed every 256 samples, or every 5.33 msec for a 48 kHz sampling rate. The transform block rate is 187.5 Hz. The AC-3 syntax groups 6 transform blocks into an AC-3 frame. The frame rate is 31.25 Hz, and each frame lasts 32 msec. Each AC-3 frame begins with a 16 bit sync word. Following the sync word are 8 bits of information which indicate sampling rate and frame size. We refer to the first 3 bytes of the AC-3 frame as *sycn info* because the information in these bytes is used to acquire and maintain synchronization with the AC-3 frames.

Following sync info is a set of data we call *bit stream info*, or BSI. The BSI data contains information about the number of channels which are coded, dialogue level, language code, information about associated services, etc. All of the BSI data is essentially static, and is descriptive about the data in the audio blocks which follow.

Following BSI are 6 coded audio blocks. The first block always contains a complete refresh of exponents, coupling coordinates, and all other information which is conditionally transmitted. The following 5 blocks may or may not contain information beyond quantized mantissas. Unused data at the end of the AC-3 frame may be considered *aux data*.

8. Features

A number of features have been designed into AC-3 in order to make it widely applicable. The goal is to have a coded audio signal which is usable by as wide an audience as possible. The potential audience may range from patrons of a commercial cinema or home theatre enthusiasts who wish to enjoy the full sound experience, to the occupant of a quiet hotel room listening to a mono TV set at a low volume who nevertheless wishes to hear all of the program content. Two of the major concerns of the AC-3 coder design were mixdown, and loudness control.

Very early in the development of AC-3, it was decided to encode the channels discretely, without the use of intermediate matrixes (as have been employed in other multi-channel coders which offer *backwards compatibility* with some existing two-channel decoders). It was felt that backwards compatibility with existing 2-channel decoders was of little value, since very few decoders were in the field to be backwards compatible with. We did, however, seriously consider the need to be backwards compatible with a variety of reproduction systems, such as mono, stereo, and matrix surround. Avoiding compatibility matrixing allowed AC-3 development to concentrate on fundamental coding techniques, and not merely on techniques which only mitigate the negative effects of compatibility matrixing.

Without the use of compatibility matrixing, it was first thought that a decoder intended to reproduce a 2-channel output from a 5-channel coded signal would have to be 2.5 times as complicated as a decoder meant for decoding a 2-channel output from a 2-channel coded signal. This was because the 2-channel decoder is required to produce a mixed-down version of the 5-channel signal, and thus must have all 5 channels decoded in order to do so. To solve this dilemma, a technique was developed which requires only partial decoding of

the 5 channels. All channels are decoded into their frequency domain representation. As an alternative to being transformed into the time domain for mixdown, the mixdown is performed directly in the frequency domain. Only the mixed-down signals are then transformed into the time domain. This mixdown technique reduces the 2-channel decoder complexity significantly, since the most complex portion of the algorithm, the inverse filter bank, only has to be performed on 2 channels. It is the fact that the frequency-to-time transformation is a linear process that allows the rearrangement of the inverse filtering and mixdown steps. There is a modest increase in the amount of input buffer memory for this downmix technique compared to the use of compatibility matrixing (where only two channels of input data must be buffered by a two-channel decoder). There is also the need to perform the bit allocation on 5 instead of only 2 channels. This modest increase in complexity is thought to be an excellent tradeoff in providing a single signal which can equally well serve a wide variety of decoders with differing numbers of reproduction channels and different down-mix optimizations.

Since AC-3 delivers all channels to the decoder coded discretely, the decoder has full flexibility to mixdown the channels as appropriate to the listening situation. Differing listening situations require differing mixdown coefficients. For example, a stereo listener may wish the surround signals to be mixed out-of-phase so that they don't localize well in two loudspeaker reproduction. A second listener may wish the same mixdown because this listener has a current home theatre matrix decoder which requires a 2-channel surround matrix encoded signal (which has the surround signal out-of-phase) as an input in order to reproduce surround. A third listener may wish a mono mix-down. This third listener could not simply combine the 2-channel mix enjoyed by the first two listeners, because then the surround signal would be lost. (The currently used Dolby Surround matrix system does suffer from the problem that the surround signal is lost in monophonic reproduction. This fact is known by programme mixers who consciously avoid the placement of essential program content solely in the surround channel. One of the goals for AC-3 was to eliminate this and other limitations of the matrix surround system. It is essential that a new multi-channel coder allow all listeners to hear all of the sound from all of the channels, so there are no constraints on programme production.) The mono listener thus requires a different downmix. Only by allowing the listener to select the desired type of downmix can all audiences be served. In most cases the listener will not be required to make any conscious decision about the desired downmix as products will be sesigned such that the AC-3 decoder will automatically make the choice depending on information such as the number of loudspeakers available.

In the United States (and perhaps elsewhere) there is a problem with the loudness of different broadcast programmes. Many broadcasters highly compress the dynamic range of the audio, and fully modulate the audio channel much of the time. In this case, there is little headroom. Sometimes the entertainment program will have a more natural dynamic range with some headroom, but commercial messages (attempting to sound loud) may not. The result is that there are significant level differences between programme segments on a particular broadcast channel, as well as between broadcast channels. Two of the design goals for AC-3 were: to eliminate the apparent level differences between broadcast channels; and allow broadcasters to compress the dynamic range of the programming for most listeners, while allowing other listeners (who so choose) to enjoy the full original dynamic range of the programme.

Loudness uniformity is achieved by determining the subjective level of normal spoken dialogue, and explicitly coding this level into the data stream as a *dialogue level* control word. The AC-3 decoder may then interact with the system playback level. Differing programmes may have differing dialogue levels which simply mean that they have differing amounts of headroom available for dramatic effect. When the listener adjusts the volume control, the level of reproduced normal dialogue (i.e. not shouts or whispers) will be set to the desired subjective sound pressure level. When a new programme segment begins, with dialogue encoded at a different level (i.e. with a different amount of headroom), the reproduction system may use the coded dialogue level control word to make a corresponding adjustment to the system playback volume. The result is that for all programmes and all channels, the reproduced level of dialogue will be uniform.

The AC-3 coder contains an integral dynamic range control system. During encoding, or at any point thereafter, dynamic range control words may be placed into the AC-3 bit stream. These control words are used by the decoder to alter the level of the decoded audio on a block basis (every 5.3 msec). The control word may indicate that the decoder gain be raised or lowered. The control words are generated by a level compression algorithm which may be resident in the AC-3 encoder, or in a subsequent bit-stream processor. The control words have a resolution of < 0.25 dB per block. The block-to-block gain variations are further smoothed by the gentle overlap add process. Gradual gain changes are free from gain stepping artifacts. For programme audio levels above dialogue level, the dynamic range control words will indicate level reduction. For audio levels below dialogue level, the control words will indicate a level increase. The default for the decoder is to use the control words which will result in the reproduction of the audio programme with a compressed dynamic range. The exact nature of the compression is determined by the algorithm which generates the control words, but in general the compression is such that headroom is reduced (loud sounds are brought down towards dialogue level) and quiet sounds are made more audible (brought up towards dialogue level). It is a decode option to reduce the effect of the control words of either polarity. If the control words which indicate that gain should be increased are not used, then low level sounds will retain their proper dynamics and only loud sounds will be compressed downwards. If only the control words which indicate gain reduction are ignored, the low level sounds will be brought up in level, but loud sounds will reproduce naturally. If all control words are ignored, the original signal dynamic range will be reproduced. It is also possible to partially use the control words for either polarity. Thus the listener may instruct the decoder to partially or fully remove the dynamic range compression which has been applied to either the loud (above dialogue level) or soft (below dialogue level) sounds. (Actually the audio is coded without any level alteration, and level alterations occur at the decoder. However, this is irrelevant to the listener, since the default for the listener is to reproduce the programme with the compression characteristic which has been intended by the programme originator. The listener must take some action to remove the compression.)

The AC-3 syntax has been defined to support the coding of one main audio service with from 1 to 5.1 channels. Additionally, associated services may be embedded into the AC-3 bit stream. Associated service types include: visually impaired (a verbal description of the visual scene), hearing impaired (dialogue with enhanced intelligibility), commentary, dialogue, and second stereo programme. All services may be tagged with a code to indicate language. Single channel services may have a bit-rate as low as 32 kb/s. The overall data stream is defined for bit rates ranging from 32 kb/s up to 640 kb/s. The higher rates are

intended to allow the basic audio quality to exceed audiophile expectations, as well as allow the incorporation of associated services without severely restricting the bit-rate (and thus quality) of the main audio service.

9. Conclusion

AC-3, with its high resolution spectral envelope coding and hybrid forward/backward adaptive bit allocation offers very high coding gain at modest complexity. Bit starvation is avoided during extreme signal demands by invoking the technique of coupling, which avoids any need to restrict the coded audio bandwidth. Fully discrete signal transmission avoids the need to increase coder complexity and compromise coded audio quality in an attempt to avoid matrixing artifacts. Decoder downmixing allows every listener to obtain the optimum downmix for his listening situation. The quality level of AC-3 is not limited to that obtainable with currently available encoders; future encoders are expected to achieve improved results by means of additional complexity and sophistication. All decoders in the field will automatically benefit from future encoding improvements. The first integrated circuit designed for AC-3 (the Zoran ZR38000) became available in 1993, and several more are expected to become available in 1994. AC-3 has been selected for use in the United States HDTV system for reasons of: high audio quality; advanced state of development; and full provision of all necessary features. AC-3 is currently being designed into consumer electronics equipment for cable television, direct broadcast satellite, and pre-recorded media.

This paper was presented at the 96th Convention of the Audio Engineering Society, February 26 – March 1, 1994, as Preprint 3796. Reprinted by permission of the AES.