Audio Engineering Society

Convention Paper 9491

Presented at the 140th Convention
2016 June 4–7 Paris, France

AC-4 – The Next Generation Audio Codec

K. Kjörling[1], J. Rödén[1], M.Wolters[2] J. Riedmiller[3], A. Biswas[2], P. Ekstrand[1], A. Gröschel[2], P. Hedelin[1], T. Hirvonen[1], H. Hörich[2], J. Klejsa[1], J. Koppens[1], K. Krauss[2], H-M. Lehtonen[1], K. Linzmeier[2], H. Muesch[3], H. Mundt[2], S.Norcross[3], J. Popp[2], H. Purnhagen[1], J. Samuelsson[1], M. Schug[2], L. Sehlström[1], R. Thesing[2], L. Villemoes[1], and M. Vinton[3]

[1] Dolby Sweden AB, Gävlegatan 12a SE-11330 Stockholm, Sweden

[2] Dolby Germany GmbH, Deutschherrnstrasse 15 – 19, Nuremberg, 90429, Germany

[3] Dolby Laboratories Inc., 1275 Market Street, San Francisco, CA 94103-1410 USA

ABSTRACT

AC-4 is a state-of-the-art audio codec standardized in ETSI (TS 103 190 and TS 103 190-2) and the TS 103 190 is part of the DVB toolbox (TS 101 154). AC-4 is an audio codec designed to address the current and future needs of video and audio entertainment services, including broadcast and Internet streaming. As such, it incorporates a number of features beyond the traditional audio coding algorithms, such as capabilities to support immersive and personalized audio, support for advanced loudness management, video-frame synchronous coding, dialog enhancement, etc. This paper will outline the thinking behind the design of the AC-4 codec, explain the different coding tools used, the systemic features included, and give an overview of performance and applications.

1. INTRODUCTION

With video entertainment entering a new era, where viewers increasingly seek flexibility in what they watch, when they watch it, and how they choose to engage with it, a new generation of audio delivery technology is required to meet the demands of these new consumption patterns and provide flexibility for continued innovation.

One of the primary goals considered during the design of the AC-4 codec – beyond core coding efficiency – was to include system level features and functionality that address long standing ecosystem challenges across day-to-day operations for broadcast, cable, satellite, and over-the-top services. Drawing from several decades of experience, the AC-4 audio coding system includes native support for features that eliminate the need for a number of complex and expensive processes typically found throughout the delivery chain including, complex bitstream synchronization and timing management, use of external loudness and dynamic range controllers, simulcasting for multi-language and/or descriptive services (including dialog enhancement), receiver-side post processing, as well as baseband loudness compliance measurement at turn-around points to name a few examples. As a result, the AC-4 system reduces infrastructure costs and operational complexity, which combined with the scalable capabilities ensure that next-generation services including immersive and personalized audio are within reach for any size organization.

In summary the AC-4 system was designed to improve today's operations and services, and enable tomorrow's experiences.

This paper will outline the high level aspects of the AC-4 audio coding system (Section 2), the bitstream syntax (Section 3), and the audio coding tools employed by the system (Section 4). Finally, performance data is presented (Section 5) and an overview of status of deployment and applications is given (Section 6).

2. THE AC-4 AUDIO CODING SYSTEM

2.1. Overview

The AC-4 codec is a state-of-the-art audio codec for traditional channel based content, immersive channel based content, object based immersive content, and for audio supporting personalization use-cases. As such it supports channel based content in formats like 2.0, 5.1, 5.1.2 to 9.1.4 (where the x.y.z representation means speakers placed in Horizontal.Lfe.Ceiling), and 22.2 [1]. It further supports Object based content both as individual semantic objects as well as Spatial Object Groups [2] as used for Atmos [3] home delivery, and a perceptually motivated sound field format.

The AC-4 system is designed to be efficiently implementable on an as wide variety of devices as possible. Three important aspects of AC-4, listed in the following, enable this.

Core/Full Decode and the **Input/Output Stage:** The syntax and tools are defined in a manner that supports decoder complexity scalability. This enables a design around the principle of an input stage (the decoding of bitstream) and output stage (the decoding/rendering to a specific playback layout). These aspects of the AC-4 coding system ensure that all devices, across multiple device categories, can decode and render the audio cost-effectively. It is important to note that the core decode mode does not discard any audio from the full decode.

Sampling Rate Scalable Decoding: For high sampling rates (i.e. 96 kHz and 192 kHz), the decoder can decode the 48 kHz part of the signal without the complexity burden of the high sampling rate.

The AC-4 system is further designed to handle splices in bitstreams without audible glitches at splice boundaries, both for splices occurring at an expected point in a stream, as well as for splices occurring in a non-predictable manner.

Finally, the AC-4 system offers increased efficiency not only from the traditional bits/channel perspective, but also by allowing for the separation of elements in the delivered audio. As such, use-cases like multiple language delivery etc. can be efficiently supported, by combining an M&E (Music and Effects) with different dialog tracks, as opposed to sending several complete mixes in parallel.

2.2. Video Frame Synchronous Coding

The AC-4 Audio Coding System is the first emission codec that can be configured to perform video frame synchronous operation. The supported video frame synchronous frame rates are: 24 Hz, 30 Hz, 48 Hz, 60 Hz, 120 Hz, and 1000/1001 multiplied by those, as well as 25 Hz, 50 Hz, and 100 Hz.

The video frame synchronous operation is achieved by a combination of sampling rate conversion and matching of the audio frame length to the video frame duration. For higher video frame rates, a mode of operation called Efficient High Frame Rate is available that combines the system benefits of video frame synchronous coding at high frame rates with the efficiency of an audio codec that uses long MDCT transforms. This is done by splitting longer audio frames into two or four system frames. Using a sequence counter a decoder can detect frame drops and frame repetitions and conceal the gaps so that the media duration remains as indicated by the system frames. I-frames can be inserted at any time in an Efficient High Frame Rate stream utilizing Seamless Frame Rate Switching and Signal Aligned Metadata.

The AC-4 Audio Coding System supports seamless switching of frame rates, which are multiples of a common base frame rate. For example a decoder can switch seamlessly from 25 Hz to 50 Hz or 100 Hz. An I-frame is not needed at the switching point.

Such seamless switching functionality is enabled by a concept called Signal Aligned Metadata. AC-4 comprises a variety of coding tools. Some of those tools operate on the signal after a transform into the QMF domain (see section 4), and hence with some delay compared to the spectral data transmitted in the coded frame. The AC-4 audio coding system is designed such that the decoder will delay the parameters for all coding tools and other applications the same amount as the signal is delayed in the decoder. That means that all data, like the spectral data, parametric coding data and metadata that are applied to one frame of the audio signal are transmitted in a single AC-4 frame. Thus, after a switch, all data for decoding is immediately present and no data relevant for finishing the decoding of the previous frame is lost.

2.3. Dialog Enhancement

One important feature of AC-4 is Dialog Enhancement (DE) that enables the consumer/user to adjust the relative level of the dialogue to their preference. With a gradual control, the amount of enhancement can be chosen on the playback side, while the maximum allowed amount can be controlled from the headend.

Dialogue Enhancement is an end-to-end feature and the relevant side-information bitrate scales well with the flexibility of the AC-4 coding core, from very cheap parametric Dialogue Enhancement modes up to modes where dialogue is transmitted in a self-contained manner, part of a so-called Music & Effects plus Dialog (M&E+D) presentation. See **Error! Reference source not found.** for typical rates.

DE mode	Typical bitrate [kb/s]	Long-term bitrate [kb/s]
Parametric	0.75 – 2.5	0.4 – 1.3
Hybrid	8 – 12	4.7 – 6.7
M&E+D	24 – 64	13 – 33

Table 1 DE modes and corresponding typical side information bitrates when dialogue is active, and the long-term average bitrate when dialog is active in only 50% of the frames.

In the hybrid mode an efficient combination of parametric DE and waveform coded dialog allows to bridge the gap between the parametric mode and M&E+D.

2.4. Dynamic Range Control and Loudness

A flexible DRC (Dynamic Range Control) solution is essential to serve the wide range of playback devices and playback environments, from high-end AVR systems via flat-panel TVs in living rooms down to tablets, phones and headphones on-the-go. AC-4 provides means to serve this wide range of use-cases with a highly flexible DRC and loudness management solution. In AC-4, four default and independent DRC decoder operating modes are defined that correspond to certain playback level ranges, as shown in Table 2.

DRC Decoder mode	Output level range [$\mathrm{dB_{FS}}$]
Home Theatre	-31..-27
Flat panel TV	-26..-17
Portable – Speakers	-16..0
Portable – Headphones	-16..0

Table 2 DRC decoder operating modes

DRC metadata as used in legacy codecs [4] has the drawback to increase the required bitrate significantly when the flexibility increases. AC-4 therefore transmits a parametric description of the DRC profiles in the bitstream rather than pre-calculated gains.

Apart from the bitrate advantage, calculating the DRC gains at the receiving end provides a high degree of flexibility, such as rendering to different speaker setups or content personalization, by leveraging multichannel and multiband DRC. In addition, AC-4 DRC supports transmitting DRC gains explicitly (in order to support legacy content with DRC gains metadata), including multi-channel and multi-band gains.

Loudness management in AC-4 includes a novel end-to-end signaling framework along with a real-time adaptive loudness processing mechanism – compliant with loudness regulations worldwide – that provide the service provider with an intelligent and automated system that ensures the highest quality, and that compliant programming is always delivered to listeners. These core technology components sit on top of a control architecture that carry a rich set of loudness and dynamic range descriptors (in addition to traditional loudness-related metadata) that can dynamically and adaptively control inter/intra-facility (downstream) processing in a more cooperative and intelligent manner than what could be achieved with previous coding systems. These new descriptors can be leveraged throughout the entire broadcast distribution, re-distribution (i.e. local affiliates and/or re-transmission through cable/sat/IPTV operators) and consumer delivery chain to reduce operational complexity and costs associated with maintaining regulatory compliance, monitoring, logging while also providing the ability to intelligently eliminate cascaded affiliate/operator loudness processing on a dynamic basis – an industry first.

AC-4 is also designed to ensure that loudness compliance is maintained when several substreams are combined into a single presentation (see section 3) upon decoding, e.g. M&E+D, or Main+Associated presentations.

2.5. Hybrid Delivery

AC-4 is designed to support hybrid delivery where e.g. audio description or an additional language is delivered over a broadband connection, while the rest of the AC-4 stream is delivered as a broadcast stream.

The flexibility of AC-4 syntax allows for easy signaling, delivery and mixing upon playback of audio substreams, which allows for splitting the delivery/transmission between two delivery paths. At the receiver side the timing information needed to combine the stream can either be solved using timing data from the system layer where e.g. DASH would be used for both the broadcast and the broadband delivery.

AC-4 can also insert timing & program ID (grid) metadata into the original AC-4 bitstream at the substream level. This means that both of the delivery streams (the broadcast stream containing the Main Program substream, and the broadband stream containing the Audio Description substream) have identification and timing information which is robust to any repackaging or other system-level manipulation necessary for delivery. Thus, at the receiving device, the timing and program ID (grid) information can be used to validate, and if necessary adjust the alignment of the two substreams delivered from the transport layer.

2.6. Metadata Pipe

Metadata has always been a fundamental component of Dolby’s coding systems over several generations of deployment. The AC-3 system developed in the early 90’s was the first coding system that enabled scalable and consistent playback across the most common devices used at the time. However, as device capabilities, form factors, distribution paths and user applications have expanded over the following decades, the need for a more robust and richer set of metadata for this purpose had become increasingly apparent. The AC-4 system supports an expanded and enriched set of metadata across several new categories including advanced loudness and dynamic range control for several device types and applications, dialogue enhancement, spatial representation, advanced rendering control, program ID/timing and interactivity.

In addition, the metadata generation in AC-4 takes advantage of a secure authentication mechanism that ensures robust and reliable delivery throughout the distribution and re-distribution pathways. This mechanism is capable of validating both the audio essence and the metadata as trusted. Metadata failing authentication – in a receiver or processor in the distribution path – can be utilized to drive more intelligent fallback processing as well as simplifying monitoring for applicable loudness regulations as an example.

To address future metadata needs, metadata extensibility is critical to ensure that a low-friction pathway for distributing/delivering new metadata is always available. The Extensible Metadata Delivery Format (EMDF) [5] syntactical elements in AC-4 provide a structured and extensible container for a collision-free and open pathway for additional information (for example, third-party metadata, third-party application data, and so on.) to be carried in AC-4 bitstreams.

3. AC-4 SYNTAX

An AC-4 audio stream is organized in frames, each with a descriptive section at the frame start (*TOC*, Table of Content), followed by a collection of substreams, most prominently audio substreams. The TOC holds information about *presentations* contained in the stream.

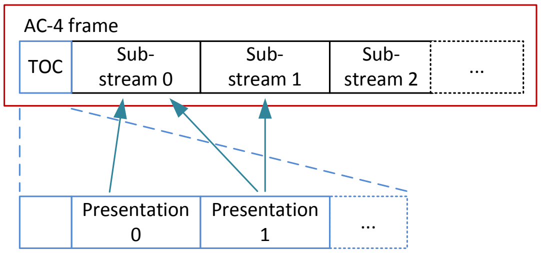

Figure 1 AC-4 bitstream structure

The AC-4 stream may contain either one or multiple presentations. A presentation describes the collection of audio elements that are to be decoded and presented simultaneously. Thus, decoding is always preceded by the selection of a presentation. By example, two presentations in a stream may represent two different language versions of a movie audio track. They might as well represent the same content, but coded at different bitrates, or optimized for specific playback devices (e.g. speakers and headphones). In order to support receiver-based supplementary audio (SA) mixing, two presentations may represent a 'Main' and a 'Main + SA' variant of the same program.

Audio substreams are always included in a presentation by reference. It is noteworthy that presentations in a stream can (but don't have to) share audio substreams, such that e.g. the two presentations describing Main and Main+SA typically share the substream (or the substreams) forming the Main Audio content. Likewise, two different language versions of the same movie audio track may share the Music&Effects substream but include a different substream.

The TOC holds information that is needed for all bitstream processing that happens before decoding, e.g. identifying and selecting presentations and their properties (e.g. language), pruning of streams (i.e. removing presentations) and merging bitstreams. This facilitates also 'late binding', i.e. assembling presentations whose substreams are delivered as separate elementary streams (e.g. over hybrid broadcast/broadband). Information that is not needed for any of these manipulations (and thus, only needed for the actual decode) is residing outside the TOC, i.e. in the substream part of the AC-4 frame.

The mentioned bitstream manipulations are facilitated by a bitstream structure that only requires a re-write of the small TOC while the contained audio substreams can be copied or omitted without parsing or decoding them e.g. in a re-multiplexing engine.

The AC-4 frame structure allows for future additions in a backwards compatible manner, meaning that presentations according to the current specifications can coexist in the audio stream with presentations that rely on future extensions, allowing legacy decoders to safely ignore the latter.

4. AC-4 CODING TOOLS

4.1. AC-4 Decoder Overview

The AC-4 decoder (see Figure 2) is built around wave-form coding in the MDCT domain and parametric coding tools in a complex pseudo-QMF domain. In the MDCT domain two different spectral front-ends are provided, one tailored for arbitrary audio signals (the ASF, Audio Spectral Front-end), and the other tailored for speech signals (the SSF, Speech Spectral Front-end). The former is based very much on a perceptual model of quantization and coding, and the latter operating under a source model of speech. Further in the MDCT domain joint channel coding (SAP, Stereo Advanced Processing) is performed, supporting joint channel coding of up to five channels as opposed to prior art codecs that typically only do joint channel coding of channel pairs.

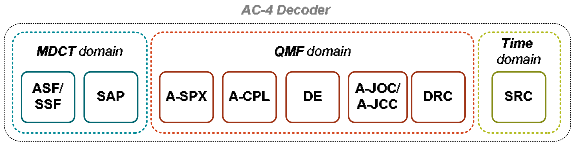

Figure 2 AC-4 decoder overview

In the QMF domain, a companding tool is introduced to control the temporal distribution of the quantization noise introduced in the MDCT domain. Subsequently tools for spatial coding are present. The Advanced Coupling Tool (A-CPL) operating on stereo, 5.1, and 5.1.2 to 9.1.4 content, the Advanced Joint Channel Coding Tool (A-JCC) operating on 5.1.2 to 9.1.4 content, and the Advanced Joint Object Coding Tool (A-JOC) operating on object based content.

Further, also in the QMF domain a high frequency reconstruction algorithm is present (A-SPX, Advanced Spectral Extension), that re-constructs missing high frequencies, and the DE and DRC functionality as elaborated on above also operate in the QMF domain. The sampling rate converter (SRC) in the time-domain is needed for video-frame synchronous coding as discussed above.

More details on the coding tools are given in the subsequent sections.

4.2. Audio Spectral Front-End

The Audio Spectral Front-End (ASF) is a general purpose waveform codec. For an efficient encoding, the incoming time signal is transformed into spectral coefficients with the help of one or several MDCT transforms per audio coding frame. Five different transform lengths corresponding to the frame length or fractions thereof are available. Depending on the signal characteristics, for each frame the transform lengths are chosen that give the optimal compromise between coding gain and avoidance of temporal artifacts.

The spectral coefficients are grouped together in bands, of increasing bandwidth as a function of frequency, in a manner matching the critical bands of the human auditory system. A psychoacoustic model determines quantizer accuracy needed for each of these bands which governs the distribution of available bits over frequency and the consecutive MDCT transforms of a frame. Some frequency bands might end up with all coefficients being quantized to zero. In order to avoid a spectral hole, the coefficients in such a band can be replaced with random values at the decoder side. The noise fill in the decoder is controlled from the encoder by the transmission of the RMS level for the bands that have been quantized to zero. The RMS level is quantized to 3 dB accuracy and is transmitted as Huffman coded deltas relative to the RMS level of the previous band. If the noise fill band is adjacent to a non-zero quantized band then the decoder must compute the RMS level of the previous band.

By utilizing the AC-4 bit reservoir it is possible to spend more bits than available in average for the more difficult to encode frames and retain bits on the easier frames of the signal. The output of the psychoacoustic model gives an indication of the relative difficulty of the audio frames that helps to control the bit reservoir.

4.3. Speech Spectral Front-End

The speech spectral front-end (SSF) is a prediction based coding tool which operates on spectral coefficient vectors of an MDCT transform with stride near 5ms. This choice enables the coding of both transient details and rapidly varying voiced structure properties of speech signals without introducing time smearing and reverberation artifacts associated with the use of longer transforms. Banded envelope power data is updated for each fourth coding unit, (20ms), and geometrically interpolated to furnish a model variance across frequency for each coding unit. A backwards adaptive perceptual weighting is then used in combination with a set of model based scalar quantizers which come in three flavors: plain, dithered, and noise filling. Arithmetic coding is used for the quantized data. Finally, an efficiently tabulated periodic signal model based predictor is inserted in the coding loop. Each bin is predicted by approximately twenty nearby bins in an MDCT domain signal buffer. Since the quantizers now meet a prediction residual, their model variance and rate allocation are adapted by heuristic rules based on predictor gain and envelope data.

The SSF is designed to operate on speech at low bitrates, where it offers an advantage over the ASF. One approach for selecting spectral front end is to employ a speech detector. The MDCT windowing trivially enables seamless switching between ASF and SSF.

4.4. Stereo Advanced Processing

The SAP tool performs joint channel coding in the MDCT domain and exploits signal redundancy in different audio channels for improved bitrate efficiency taking into account binaural masking release effects. Dependent on the channel mode different schemes are available that define the jointly coded channels. Those schemes are perceptually motivated, but are also defined such that low complexity core decoding is enabled for immersive channel modes. Joint stereo coding is achieved by traditionally known Mid/Side (MS) coding and Enhanced Mid/side coding which is designed to better handle panned signal compared to MS coding. Thus redundancy between the mid signal and the side signal is reduced and the main quantization noise is spatially shaped towards a dominant sound source anywhere in the stereo image. Joint MDCT coding is applied in perceptually motivated frequency bands. Dependent on the signal characteristics, certain bands may be jointly coded while others may be coded separately. Effectively the original stereo signal is transformed by means of a parameter controlled time and frequency dependent matrix. The time and frequency resolution depends on the MDCT transforms present in the audio frame. For multichannel 5.1 input the SAP tool works on two selected channel pairs or 3, 4 or 5 audio channels are jointly coded depending on the cross-channel characteristics. Similar to joint stereo coding this results in parameter controlled time and frequency dependent matrix encoding of the multichannel signal which reduces cross-channel redundancies. For example for jointly coding the left, right and center channel of 5.1 content, 1 out of 12 different basic matrix types can be selected that gives the best performance.

4.5. Companding

The companding tool is employed in the QMF-domain to achieve temporal shaping of the core coder (ASF or SSF) quantization noise. Companding in the encoder reduces the dynamic range of the input audio signal before the core encoding process. Modification is done per QMF time slot (in the core coding frequency range, see Section 4.6) by a broadband gain value. These gain values amplify slots of relatively low intensity and attenuate slots of relatively high intensity. Therefore, the output of the core decoder is a signal with reduced dynamic range perturbed by core coder quantization noise of almost uniform level (time envelope) within each frame. Expanding in the decoder restores the core decoder outputs back to the original dynamic range by applying inverse of the encoder gain values per QMF time slot. In this manner, quantization noise is concurrently shaped to approximately follow the temporal envelope of the original signal. Gains calculation using a p -norm of the spectral magnitudes with $p<2$ has been found to be more effective in shaping quantization noise, than basing it on energy (p=2). In AC-4, mean absolute level (p=1) has been chosen.

Typically, companding is activated for transient signals and switched off for stationary signals. Instead of switching off companding abruptly, a constant gain is applied to an audio frame resembling the gains of adjacent active companding frames. Such a gain factor is calculated by averaging mean absolute levels over slots in one frame. For highly correlated multi-channel signals, equal companding gains are applied to all the channels. Encoder control of the desired decoder expanding level is signaled in the bit-stream so that inverse of the encoder gain values are applied to the corresponding QMF time slots.

4.6. Advanced SPX

The Advanced Spectral Extension algorithm (A-SPX) operates in the QMF domain and performs high frequency reconstruction similar to e.g. MPEG SBR [6], or DD+ SPX [7]. As such, the waveform coded low-band signal is used to recreate a high-band signal that is subsequently adjusted, using the A-SPX side-information, to match the properties of the original high band signal. However, as opposed to the earlier versions it allows for very flexible interleaving of wave-form coded elements with the parametrically coded elements. As such, it circumvents one of the fundamental limitations of earlier systems, namely the inability to accurately re-construct important tonal or transient components in high frequencies. This is achieved by letting the waveform core coder run at the same sampling rate as the output signal, which enables the MDCT waveform coder to code spectral lines over the entire frequency range if needed, or with A-SPX only covering a small part of the highest frequency range.

4.7. Advanced Coupling

The Advanced Coupling Tool (A-CPL) is a parametric spatial audio coding tool for stereo, 5.1, and immersive channel based content. It is an evolution of similar technologies in DD+ [7], MPEG-4 Parametric Stereo [8]. A-CPL allows for the reconstruction of a two-channel signal from a mono downmix *M* and associated A-CPL parameters alpha and beta. These time- and frequency-dependent parameters, in combination with a decorrelator, control the creation of a side signal,

$$S = \alpha M + \beta d\left(M\right) \tag{1}$$

where *d(M)* is a decorrelated version of the downmix *M*.

In A-CPL, the transmission of the parameters alpha and beta uses a non-uniform quantization scheme that takes into account the sensitivity of the human auditory system to quantization errors of the spatial parameters, which in turn depends on the position in the (alpha, beta)-plane. For signals mostly panned to left or right, which correspond to the regions around (1,0) and (-1,0) in the (α,β)-plane, substantially finer quantization is desirable than in less critical regions. To avoid the complexity of a full vector quantizer approach, a cascaded scalar non-uniform quantization scheme as illustrated in Figure 3 is used in AC-4.

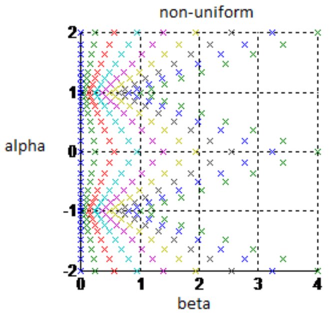

Figure 3 Non-uniform quantization of spatial parameters.

4.8. Advanced Joint Object Coding

The Advanced Joint Object Coding tool (A-JOC) enables an efficient representation of object-based immersive audio content (e.g. Dolby Atmos) for delivery at low bitrates [9]. This is achieved by conveying a multi-channel downmix of the immersive content together with parametric side information that enables the reconstruction of the audio objects from the downmix in the decoder. The downmix itself is encoded using the available tools like ASF and A-SPX.

The parametric side information comprises both JOC parameters and object metadata. The JOC parameters primarily convey the time- and frequency-varying elements of an upmix matrix that reconstructs the audio objects from the downmix signals. Similar to A-CPL and A-JCC, the upmix process is carried out in the QMF domain. The JOC upmix process also includes decorrelators that enable an improved reconstruction of the covariance of the objects, which is controlled by additional JOC parameters. Finally, the reconstructed objects are rendered to the desired playback configuration, where the rendering is governed by the object metadata, conveying e.g. the spatial positions of the objects.

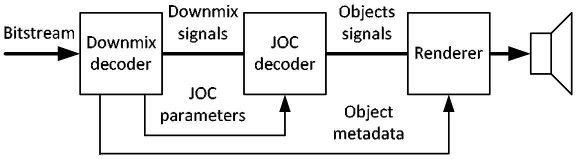

Figure 4 A-JOC decoder

4.9. Advanced Joint Channel Coding

The Advanced Joint Channel Coding (A-JCC) [10] tool is designed for efficient coding of channel based immersive material, such as 7.1.4 or 9.1.4, at low bitrates. The tool facilitates parametric coding using a five-channel downmix (LFE is passed through). A-JCC has two key features: efficient representation of the full upmix matrix that significantly reduces the side information, and a dynamic downmixing process that adapts to the characteristics of the input signal. A-JCC provides an extension to the A-CPL parameterization, enabling each downmix channel to be the sum of either two or three original channels. The upmix matrix consists of dry and wet parameters controlling decorrelator contribution, which are estimated on the encoder side. The resulting parametric side information rate for 7.1.4 content is 8-10 kb/s.

A-JCC has a set of downmix configurations that the encoder selects from for each frame based on the input content properties. Several possibilities for selecting the optimal downmix can be considered, and in order to avoid rapid switching from one downmix to another it can be required that a certain downmix is maintained for a number of consecutive frames before switching. In the decoder side, a smoothing pre-matrix is applied to decorrelator feeds in order to obtain a seamless transition from one downmix to another.

5. PERFORMANCE

In the following results are presented from evaluating channel based performance as well as performance for immersive audio.

Firstly, internal listening test (BS1534 / MUSHRA [11]) results are shown (see Figure 5) comparing the performance of AC-4 for two higher quality operation points (64kb/s stereo and 96kb/s stereo). The test is done on critical stereo content (the MPEG test set) over loudspeakers using expert listeners. As can be seen from the results, an average score in the Excellent range can be achieved already at 64kb/s, while with 96kb/s not only the average but also all individual item scores are in the excellent range.

Secondly a 5.1 evaluation is shown (see Figure 6) using expert listeners and critical 5.1 material (from the EBU multichannel testing [12]), at 80kb/s, 144kb/s, and 208kb/s. As can be seen from the results, the average score at 80kb/s fall in the middle of the "Good" range on the MUSHRA scale, while 144kb/s is on the border to "Excellent", and the 208kb/s is safely in the "Excellent" range with no individual item scoring below 80.

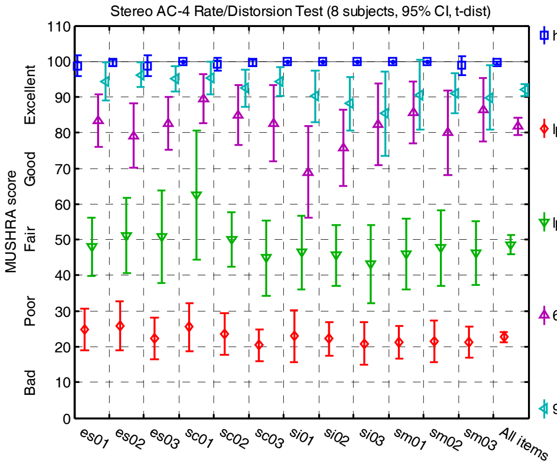

Figure 5 Listening test on stereo content

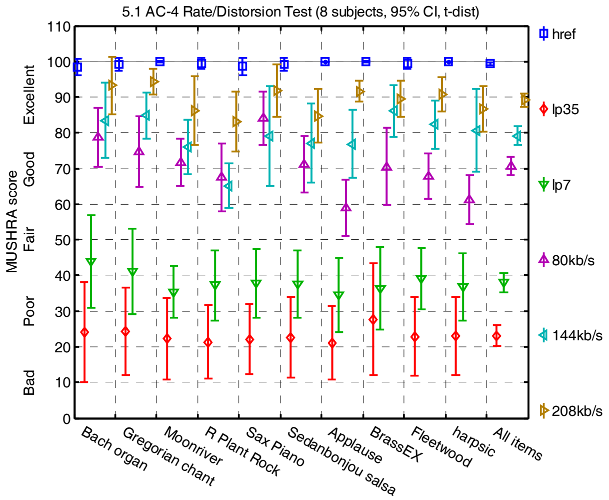

Figure 6 Listening test on 5.1 content

Thirdly, a 7.1.4 evaluation is shown (see Figure 7) using expert listeners and critical 7.1.4 material, at 144kb/s, 256kb/s, and 384kb/s. As can be seen from the results, the average score at 144kbps fall in the middle of the "Good" range on the MUSHRA scale, while 256kb/s and 384kb/s both score in the "Excellent" range.

Finally, results for Atmos immersive content coded with AC-4 are given in [9].

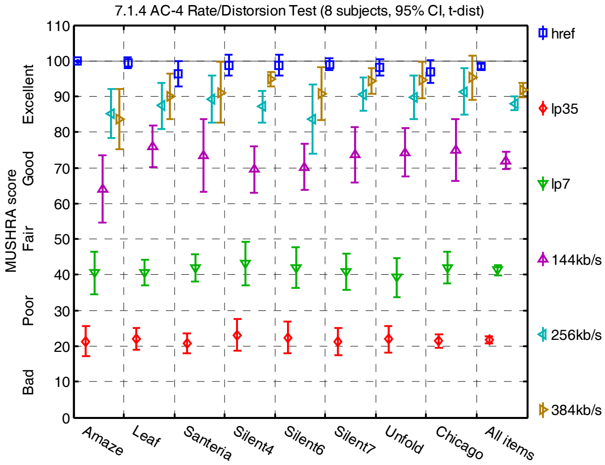

Figure 7 Listening test on 7.1.4 channel based immersive content.

6. APPLICATIONS AND STATE OF DEPLOYMENT

The AC-4 Audio Coding System as outlined in this paper is applicable across all digital audio transmission applications and adds value well beyond the improved compression efficiency even for plain stereo and 5.1 services. It is the next generation emission coding system supporting the content flow from production and distribution all the way to device specific playback configurations, and with its rich feature set it allows scaling services over time as more and more advanced content is created.

Content production for Immersive Audio in scale became reality when Dolby Atmos was launched in 2012, and since then more than 400 theatrical titles and over 50 Blu-ray titles have been produced in the Atmos object based format, covering cinematic content as well as episodic TV content.

In order to enable distribution and interchange of immersive and personalized content work is ongoing in ITU defining metadata (ADM) [13] and the audio data format BW64 (BWAV) [14].

Furthermore with the increasing availability of Dolby Atmos capable AVRs and Soundbars, immersive audio coded with AC-4 can be played back using these devices by connecting with MAT [15] over HDMI.

7. ACKNOWLEDGEMENTS

We acknowledge the work of the Dolby Engineering teams and QA team, which is instrumental in bringing the AC-4 codec from a theoretical prototype to a fully featured and thoroughly tested product. We further acknowledge the huge effort by everyone working on standardization, strategy, marketing and business development, which is essential to enabling the adoption of the AC-4 codec in the marketplace.

8. REFERENCES

[1] Recommendation ITU-R BS.2051-0 - Advanced sound system for programme production

[2] Riedmiller et al., *“Immersive and Personalized Audio: A Practical System for Enabling Interchange, Distribution, and Delivery of Next-Generation Audio Experiences,”* Motion Imaging Journal, SMPTE, 124(5), pp. 1–23, 2015, ISSN 1545-0279, doi:10.5594/j18578.

[3] Dolby Laboratories, *Dolby Atmos*, 2015, available at http://www.dolby.com/us/en/brands/dolby-atmos.html.

[4] Robinson et al., Dynamic Range Control via Metadata, 107th AES convention, September 1999, New York.

[5] ETSI TS 102 366 V1.3.1 (2014-08) Digital Audio Compression (AC-3, Enhanced AC-3) Standard, Annex H.

[6] Dietz et al., *“Spectral Band Replication a novel approach in audio coding”*, Audio Engineering Society Convention Paper 5553 Presented at the 112th Convention 2002 May 10–13 Munich, Germany.

[7] Andersen et al., *“DD+ reference Introduction to Dolby Digital Plus, an Enhancement to the Dolby Digital Coding System”*, Audio Engineering Society Convention Paper 6196 Presented at the 117th Convention 2004 October 28–31 San Francisco, CA, USA-

[8] Purnhagen, H., *“Low complexity parametric stereo coding in MPEG-4”*, Proc. of the 7th Int. Conference on Digital Audio Effects (DAFx’04), October 5-8, 2004, Naples, Italy.

[9] Purnhagen et al., *“Immersive Audio Delivery Using Joint Object Coding”*, Audio Engineering Society Convention Paper Presented at the 140th Convention 2016 June 4–7, Paris, France.

[10] Villemoes el al., *Parametric Joint Channel Coding of Immersive Audio*, under preparation.

[11] *“Method for the subjective assessment of intermediate quality levels of coding systems,”* ITU-Recommendation BS.1534-3, 2015.

[12] EBU - TECH 3324, *EBU Evaluations of Multichannel Audio Codecs*, September 2007, Geneva

[13] ITU-R BS.2076-0 Audio Definition Model (06/2015)

[14] ITU-R BS.2088-0 Long-form file format for the international exchange of audio programme materials with metadata (10/2015)

[15] Dolby Laboratories, *Dolby Atmos for the home theater*, 2015, available at http://www.dolby.com/us/en/technologies/dolby-atmos/dolby-atmos-for-the-home-theater.pdf