Audio Engineering Society

Convention Paper 9587

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

Immersive Audio Delivery Using Joint Object Coding

Heiko Purnhagen[1], Toni Hirvonen[1], Lars Villemoes[1], Jonas Samuelsson[1], and Janusz Klejsa[1]

[1]*Dolby Sweden AB, Gävlegatan 12A, 11330 Stockholm, Sweden*

Correspondence should be addressed to Heiko Purnhagen (heiko.purnhagen@dolby.com)

ABSTRACT

Immersive audio experiences (3D audio) are an important element of next-generation audio entertainment systems. This paper presents joint object coding techniques that enable the delivery of object-based immersive audio content (e.g. Dolby Atmos) at low bit rates. This is achieved by conveying a multi-channel downmix of the immersive content using perceptual audio coding algorithms together with parametric side information that enables the reconstruction of the audio objects from the downmix in the decoder. An advanced joint object coding tool is part of the AC-4 system recently standardized by ETSI. Joint object coding is also used in a backwards compatible extension of the Dolby Digital Plus system. Listening test results illustrate the performance of joint object coding in these two applications.

1 Introduction

The object-based representation of immersive audio content (3D audio) is a powerful approach that combines intuitive content creation with optimal reproduction over a large range of playback configurations using suitable rendering systems [1]. Object-based audio is, for example, a key element of the Dolby Atmos [2, 3] ecosystem for creation of immersive cinematic audio content and its playback to theater audiences. An object comprises both the audio waveform itself as well as dynamic object metadata, conveying e.g. its spatial position. Complex audio scenes in cinematic content can sometimes have more than hundred simultaneously active objects. While this amount of information can be handled easily when cinematic content is distributed to theaters, the situation is different when immersive audio content is delivered to consumer entertainment systems, in particular in broadcast or streaming scenarios. In these situations, an efficient representation of the immersive audio content is required to enable transmission at low bit rates.

While immersive audio content can also be represented in other formats, for example using a channel-based representation [4, 5], and while an object-based representation of audio content can also be used to enable other features like personalized experiences [1], this paper focuses on delivering object-based representations of immersive audio content to consumer entertainment devices. Delivering immersive content in this format enables optimal reproduction over playback systems ranging from large immersive loudspeaker configurations for home theater systems, for example a 7.1.4 configuration with a 7.1 setup in the horizontal plane and 4 ceiling speakers, over configurations that use

upward-firing speakers instead of ceiling speakers [6], playback systems in the form of a sound bar, to legacy configurations like 5.1 and 2-channel stereo. Furthermore, consumer devices can also render object-based immersive content for optimal binaural playback over headphones.

This paper is structured as follows. First, the joint object coding paradigm is introduced and various aspects of it are described in detail. Then, two applications implementing this paradigm are presented. Finally, the performance of joint object coding in these applications is illustrated by listening test results, and conclusions are discussed.

2 The Joint Object Coding Paradigm

2.1 Basic Approach

The joint object coding (JOC) paradigm presented in this paper facilitates an efficient representation of object-based immersive audio content suitable for broadcast or streaming applications operating at low bit rates. An encoder implementing this paradigm comprises several steps, as shown in Fig. 1. Input to the encoder is content in an object-based immersive representation, comprising waveforms of the object signals $\mathbf{X}$ and the associated object metadata. First, a multi-channel downmix of the immersive content is generated by a downmix renderer, governed by the object metadata. In addition to the downmix signals $\mathbf{Y}$, the downmix renderer also generates downmix metadata that enables playback of the downmix itself, as will be discussed in Sec. 2.4. Based on the object and downmix signals, the JOC encoder computes JOC parameters that enable an approximate reconstruction of the audio objects from the downmix in the decoder. Finally, perceptual audio coding algorithms are used in the downmix encoder to convey the downmix itself at a low bit rate. The JOC parameters, as well as the object and downmix metadata, are included as side information in the bitstream.

A decoder implementing the joint object coding paradigm is shown in Fig. 2. First, the downmix decoder generates the decoded downmix signals $\mathbf{Y}'$. Next, the JOC decoder uses the JOC parameters extracted from the bitstream to generate an approximate reconstruction $\widehat{\mathbf{X}}$ of the object signals. Finally, the reconstructed object signals together with the associated object metadata extracted from the bitstream are used by a renderer to generate a presentation suitable for the playback configuration available at the decoder side. It should be noted that decoding and rendering can happen in two different devices. The downmix decoding and JOC decoding can, for example, take place in a set-top box (STB) or a TV set, which then can send the reconstructed object signals and the associated metadata to a sound bar or an audio/video receiver (AVR) for rendering and playback.

A perceptually motivated separation of the object and downmix signals into a set of non-uniform frequency bands together with temporal framing enables to compute and apply the JOC parameters for object reconstruction in a time- and frequency-variant manner. The intersection of a frequency band and a temporal frame can be referred to as a time/frequency tile and JOC parameters are computed for each tile. Frequency bands are formed by applying a 64-band complex-valued pseudo-QMF analysis bank [7] to each of the signals, and then grouping the QMF bands into a set of typically 7 to 12 parameter bands according to a perceptual frequency scale. Temporal framing uses overlapping analysis windows with a stride of typically 32 to 43 ms.

In a basic joint object coding system, each audio object is reconstructed as a linear combination of the downmix channels within a time/frequency tile. This is shown in the upper path of Fig. 3, where "T/F" indicates a QMF analysis bank. In each time/frequency tile, the coefficients of this linear combination can be represented by a matrix $\mathbf{C}$ of size $N \times M$ where N is the number of object signals and M is the number of downmix signals. By collecting the input object signals as rows in $\mathbf{X}$ and the downmix signals as rows in $\mathbf{Y}$, the reconstruction is defined by

$$\widehat{\mathbf{X}} = \mathbf{C}\mathbf{Y}. \tag{1}$$

For simplicity, indices indicating the frequency band and temporal frame of a time/frequency tile are omitted in this notation.

In order to ensure a smooth transition between temporally adjacent tiles, the elements of the matrix $\mathbf{C}$ are temporally interpolated with a linear ramp that typically corresponds to the overlap of the analysis windows. The use of a complex-valued QMF bank implies a two-times oversampling of the signals in the QMF domain, which avoids aliasing artifacts if neighboring bands are processed with different JOC parameters. Finally, the reconstructed objects are converted back to

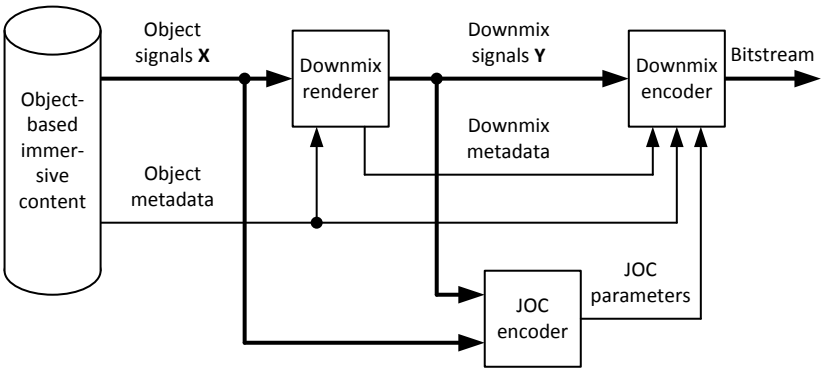

Fig. 1: Block diagram of an encoder implementing the joint object coding paradigm.

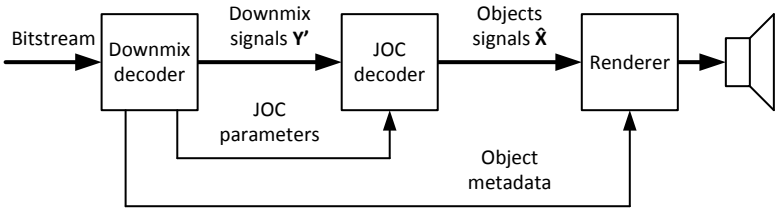

Fig. 2: Block diagram of a decoder implementing the joint object coding paradigm.

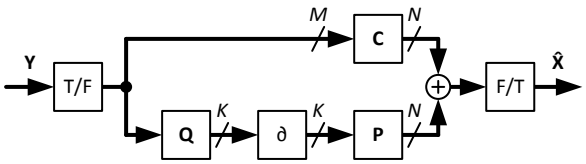

Fig. 3: Block diagram of the upmix in a JOC decoder reconstructing N object signals $\widehat{\mathbf{X}}$ from M downmix signals $\mathbf{Y}$, where the lower path can add contributions from K decorrelators ∂.

the time domain with a QMF synthesis bank, indicated as "F/T" in Fig. 3.

The JOC parameters can be computed in the encoder to achieve a least squares reconstruction of the audio object in each time/frequency tile. Such a reconstruction is optimal in the sense that it minimizes the sum of the squares of the reconstruction error of the signal samples of the audio object in a given time/frequency tile. With the notation $\mathbf{R}_{\mathrm{uv}} = \mathrm{Re}(\mathbf{U}\mathbf{V}^*)$ for sample covariance matrices, a regularized solution is given by

$$\mathbf{C} = \mathbf{R}_{\mathrm{xy}}(\mathbf{R}_{\mathrm{yy}} + \varepsilon\mathbf{I})^{-1}, \qquad (2)$$

where $\varepsilon \geq 0$ is a regularization constant and $\mathbf{I}$ is the identity matrix.

Compared to known techniques such as spatial audio object coding (SAOC) [8, 9], where an integrated object reconstruction and rendering process is used in the decoder, the joint audio object coding paradigm presented here is more flexible and can be combined with different rendering systems for playback of the reconstructed objects. Furthermore, SAOC is based on a simplified model of the original object signal covariance $\mathbf{R}_{\mathrm{xx}}$ and a description of the downmix process. If the model does not fit the data well, a model based least squares solution will perform worse than (2). The efficient transmission of the full matrix $\mathbf{C}$ in JOC leaves all decisions such as the amount of relaxation ε in (2) or even a complete deviation from the waveform matching paradigm to the encoder.

The least squares solution in (2), for example, does not ensure that the reconstructed objects have the original variance, i.e., energy. Hence, it is typically advantageous to compensate for this prediction loss by gaining the coefficients from the least squares solution in the encoder to ensure variance reinstatement within

a time/frequency tile. Another reason to use modifications of or alternatives to the least squares solution in the encoder is to ensure optimal performance for frequency bands where the downmix codec employs parametric coding methods, like high-frequency reconstruction or channel coupling, such that the decoded downmix signals $\mathbf{Y}'$ are no longer waveform approximations of the original downmix signals $\mathbf{Y}$.

2.2 Object Decorrelation

An advanced version of the joint object coding system includes decorrelator units in the decoder. Using additional reconstruction parameters, this enables an improved reconstruction of the covariance of the audio objects for each time/frequency tile, which also improves the perceptual performance of the system.

The number K of decorrelators is the main parameter of this additive enhancement to (1). A pre-decorrelator mix matrix $\mathbf{Q}$ is used to create K decorrelator feeds from the M downmix signals. Each decorrelator feed is sent into a separate decorrelator ∂ and the resulting K outputs are distributed over the N synthesized object signals according to the coefficients of an upmix matrix $\mathbf{P}$ of size $N \times K$, which are conveyed as additional JOC parameters. The lower path of Fig. 3 illustrates these decorrelator contributions to the reconstructed object signals, resulting in

$$\widehat{\mathbf{X}} = \mathbf{CY} + \mathbf{P}\,\partial\,(\mathbf{QY})\,. \tag{3}$$

Dimensionality considerations lead to the thumb rule $K = N - M$, but we have found that fewer decorrelators are sufficient if the coefficients of $\mathbf{P}$ are chosen carefully [10]. The pre-decorrelator mix matrix $\mathbf{Q}$ is derived in the decoder from $\mathbf{C}$ and $\mathbf{P}$, see [10, 11], avoiding the need for conveying further parameters.

Each decorrelator comprises an initial delay followed by an IIR all-pass filter and a "ducker" module that improves performance for transient signals [7]. The IIR all-pass filter coefficients of the K decorrelators are different to ensure that the output signals of the decorrelators are mutually decorrelated even if they are fed with the same input signal.

This approach to object decorrelation is is different from the decorrelation available in the MPEG-H SAOC-3D system [9, 12], where decorrelator contributions are added to the rendered output signals and not to the reconstructed objects themselves. Furthermore, the object decorrelation approach described here typically only requires 1 to 3 decorrelators for a perceptually appropriate reinstatement of the object covariance while the SAOC-3D system would require 11 decorrelators when rendering for playback over a 7.1.4 loudspeaker configuration.

2.3 Content Preprocessing

Complex audio scenes in cinematic content can have many simultaneously active objects, sometimes even more than hundred. While the JOC paradigm could be applied directly to such content, it would require a huge amount of JOC parameters and object metadata to be conveyed as side information. This situation is addressed by an immersive interchange translation process [1] that is illustrated in Fig. 4. In general, the immersive interchange translation process produces a PCM–based object interchange format with metadata that, when rendered to any speaker layout, sounds identical to the rendering of the original object-based mix. At a high level, this type of rendering is achieved by intelligently grouping objects into what we call spatial object groups. The grouping process is primarily based on perceptual loudness and spatial distortion metrics that work to minimize spatial error when compared to the original object-based mix. Spatial object groups are effectively an aggregated set (i.e., mixture) of the original audio objects. Generation of this immersive object interchange format is motivated by the reduced spatial resolution requirements for typical consumer loudspeaker layouts (there are approximately 7 to 20 speakers in a typical immersive home theater layout, as opposed to approximately 30 to 64 speakers in a typical Dolby immersive cinema layout).

While the immersive interchange translation process is perceptually transparent if a sufficiently high number of object groups is used [1], it is even a very helpful tool to optimize the overall rate/distortion performance when used to generate spatial object groups provided as input to the JOC encoder shown in Fig. 1. In particular for low target bit rates, it can be beneficial to reduce the number of spatial object groups sent to the JOC encoder, since this also reduces the bit rate needed for the side information conveying the JOC parameters and the object metadata, which in turn leaves more bit rate available for the encoding of the downmix signals themselves. Typically, 11 to 15 spatial object groups are used as input to the JOC encoder in combination

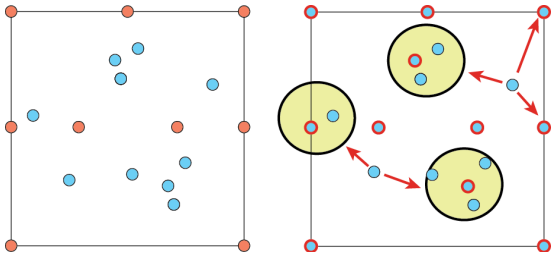

Fig. 4: Illustration of object-based audio content before (left panel) and after (right panel) immersive interchange translation where the orange and blue dots in the left panel represent 19 original objects and where the red circles in the right panel represent 11 spatial object groups.

with a low-frequency effects (LFE) channel directly conveyed as one of the downmix signals.

2.4 Downmix Strategies

The multi-channel downmix signal used as input to the JOC decoder is generated by the downmix renderer as part of the encoding process shown in Fig. 1, and different strategies can be used to generate this downmix. A basic implementation of the JOC paradigm can use a 5.1 channel-based rendition of the audio content as the downmix. This enables building a backwards compatible system, where legacy decoders will only decode and play the 5.1 downmix and ignore the side information carrying the JOC reconstruction parameters and object metadata. A decoder implementing the full JOC decoding process shown in Fig. 2, on the other hand, will use these parameters to reconstruct the objects and use the object metadata conveying the object positions to render the immersive content for optimal reproduction on the playback configuration available at the decoder side, for example an immersive 7.1.4 loudspeaker setup. The JOC paradigm can also be used with other channel-based downmixes, like a 7.1 or 5.1.2 downmix, or even a 2.1 or 2.0 downmix.

A more advanced implementation of the JOC paradigm can use an adaptive downmix instead of channel-based downmix used in the example above. In this case, each downmix channel can be formed as a dynamic spatial group of neighboring objects, thus enabling a better reconstruction of the objects in the decoder. For example, some downmix channels could carry objects in the horizontal plane, while other channels could carry objects in the ceiling, which would not be possible when

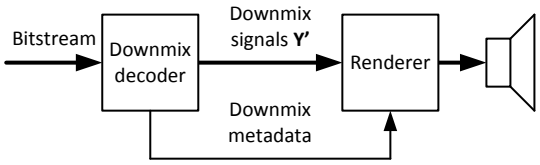

Fig. 5: Block diagram of a low-complexity core decoder for an adaptive JOC downmix.

e.g. a 5.1 rendition is used as downmix. Hence, the downmix renderer generating an adaptive downmix can employ basically the same strategies as the immersive interchange translation process discussed in Sec. 2.3.

The time-varying positions of the dynamic spatial groups forming the adaptive downmix channels can be included as downmix metadata in the bitstream, as shown in Fig. 1. While an adaptive downmix does not provide the direct backwards compatibility of a 5.1 downmix in the example above, the downmix metadata still enables rendition and playback of the adaptive downmix signals in low-complexity decoders that don't implement the full JOC decoding process, albeit at a lower quality than a full JOC decoding would provide. Such a low-complexity core decoder is shown in Fig. 5.

2.5 Parameter Quantization and Coding

The JOC paradigm allows for transmission of all elements of the upmix matrix $\mathbf{C}$ and, if present, of the decorrelation matrix $\mathbf{P}$, for all time/frequency tiles. This provides a powerful and flexible signal model, but potentially also requires a significant bit rate to convey all these parameters as side information in the bitstream. Hence, a flexible quantization and coding scheme is used that allows to operate JOC over a large range of target bit rates.

The values of the matrix elements are quantized uniformly, and two different quantizers are available, providing coarse or fine quantization with a step size of approximately 0.2 or 0.1, respectively. The time- and frequency-resolution of the JOC parameters significantly affects the necessary side information rate and can be adapted to the desired target bit rate operation point. Frequency resolutions from 23 parameter bands down to just a single band are available. Also the temporal resolution of the JOC parameters is flexible, which is achieved by signaling the ramp start time

and ramp duration that controls the temporal interpolation of the JOC upmix matrices from the previous values to the new values decoded from the new set of JOC parameters received in the bit stream. This enables to decouple the update rate of the JOC parameters from the frame rate of the downmix coder. When the downmix coder operates at a high frame rate, the JOC interpolation ramp can span multiple downmix coder frames, and JOC parameters are only embedded in those frames where a new set of parameters is conveyed. On the other hand, for long downmix coder frames, it is also possible to have more than one set of new JOC parameters in a frame.

In addition to the above mechanisms, the coding scheme for the JOC parameters contains additional provisions to improve its efficiency in particular at low target rates. For example, the upmix **C** and decorrelation **P** matrices can be sparsified. If only a few downmix signals are used to reconstruct an object, then these downmix signals can be indicated in the bitstream and there is no need to transmit the remaining columns of zero-valued elements of the upmix matrix. Similarly, an object activity flag is available, which allows to avoid transmission of rows of zero-valued elements of the upmix and decorrelation matrices for objects that are signaled as inactive (and hence only contain silence).

To encode the quantized parameters, time- or frequency-differential coding is employed, and the resulting delta values are then encoded using a set of appropriately trained Huffman code books. The encoder chooses between time- and frequency-differential coding to minimize the resulting bit rate. For frames that need to be decodable independently from previous data (like an I-frame or a random access point), only frequency-differential coding is used, which is signaled in the bitstream by a dedicated flag.

By combining the flexibility of the quantization and coding scheme described here with the possibility to adapt both the number of spatial object groups encoded by the JOC encoder (see Sec. 2.3) and the number of downmix channels, e.g. in an adaptive downmix, (see Sec. 2.4) depending on the desired target bit rate, it is possible to use the joint object coding paradigm over a large range of bit rates. This flexible trade-off between parametric and waveform coding enabled by the JOC paradigm allows to bridge the gap between parametric and discrete approaches for conveying object-based content.

3 Applications

Joint object coding is currently used in two different applications. These two applications differ both in the downmix coder being used and in configuration of the JOC system itself, and will be described in more detail now.

A first version of JOC was designed to provide a backwards compatible extension of the Dolby Digital Plus (DD+) system and enable the delivery of immersive Dolby Atmos content at bit rates like 384 kb/s, which are often used in existing broadcast or streaming applications. To ensure direct backwards compatibility with existing Dolby Digital Plus decoders, a 5.1 (or 7.1 or 5.1.2) downmix is used in the DD+ JOC system. Furthermore, this first version of JOC does not use all of the features described in the previous sections, in order to adapt it to specific aspects of this application. Decoding for DD+ JOC is, for example, already available in AVRs that support Dolby Atmos [6]. It should be noted that these AVRs can also receive immersive Dolby Atmos content over HDMI in other formats such as TrueHD or MAT 2.0 [6].

An advanced joint object coding tool (A-JOC), including adaptive downmix options and decorrelation units in the decoder, is part of the recently standardized AC-4 system [11]. It uses the perceptual audio coding tools available in AC-4 to convey the downmix signals. Furthermore, the AC-4 A-JOC system also supports all other features of the AC-4 system, such as video frame synchronous coding, dialog enhancement, and DRC/loudness processing [4]. A set-top box implementing AC-4 A-JOC decoding can, for example, send the decoded objects in MAT 2.0 format over an HDMI connection to an existing AVRs that supports Dolby Atmos for rendering and playback.

4 Results

To assess the performance of joint audio object coding in both the DD+ JOC system and the more recent and more advanced AC-4 A-JOC system, various listening tests were performed. Fig. 6 shows the results of a MUSHRA [13] listening tests for a set of 13 critical test items of object-based immersive content described in Tab. 1. The content was encoded with the DD+ JOC system at a total bitrate of 384 kb/s, and with the AC-4 A-JOC system at three different total bitrates, namely

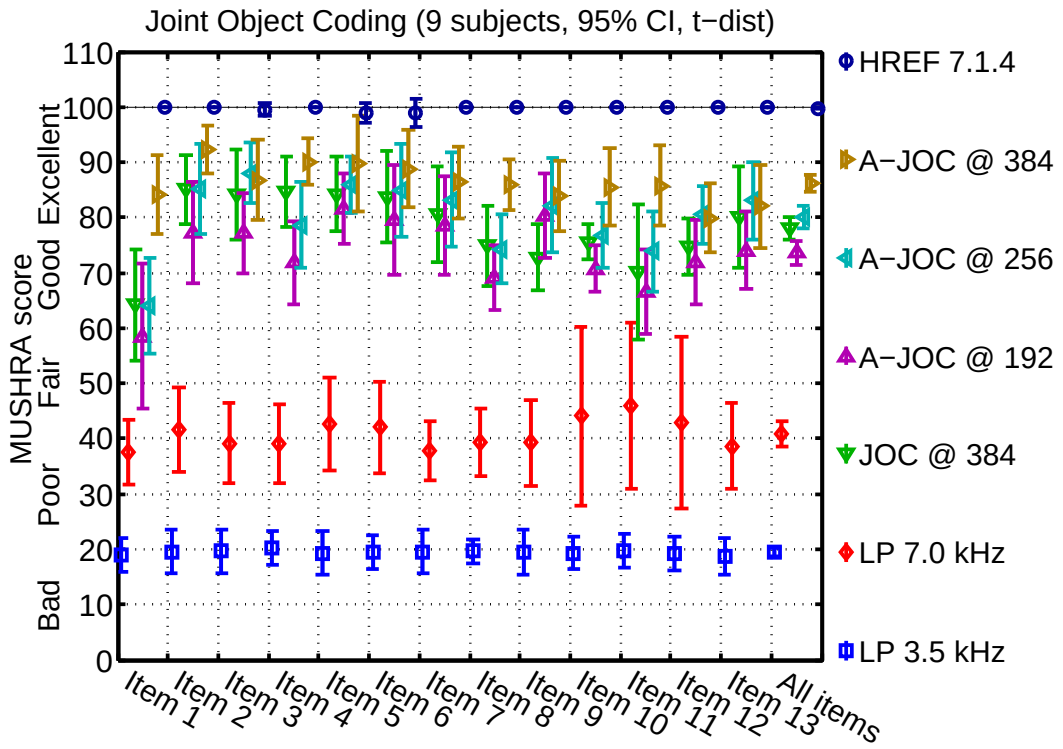

Fig. 6: MUSHRA listening test results for 9 expert listeners after post-screening for DD+ JOC at 384 kb/s and for AC-4 A-JOC at 192 kb/s, 256 kb/s, and 384 kb/s for 13 critical test items of object-based immersive content, where the original and decoded object-based content was rendered for playback on an immersive 7.1.4 loudspeaker configuration.

Item #	Description
1	Live concert with harmonica and applauding audience.
2	Bass-heavy electronic music with suppressed male voice narration.
3	Ambient music and sound of ocean waves rolling over.
4	Fixed and panned clock chimes, mechanical sounds, gears, and bells with strong transients.
5	Panned creature dialog with strong cave reverberation. Subtle running water sounds.
6	Bird in flight with jungle ambience.
7	Forest ambience with numerous wind sound effects.
8	Heavy rainfall with subtle thunderclap.
9	Electronic music with panned percussive elements, cheering crowd, and applause ambience.
10	Strong thunderclap and beginning rainfall.
11	Music with panned percussive elements and strong bass.
12	Rainfall with thunder rumble, wind noise, and music.
13	Electronic music with panned percussive elements and vocals.

Table 1: Description of the 13 critical test items in the MUSHRA listening test.

System @ total bit rate	Sideinfo bit rate
DD+ JOC @ 384 kb/s	avg. 69 kb/s
AC-4 A-JOC @ 192 kb/s	avg. 32 kb/s
AC-4 A-JOC @ 256 kb/s	avg. 41 kb/s
AC-4 A-JOC @ 384 kb/s	avg. 83 kb/s

Table 2: Average bit rate of side information conveying JOC parameters and object metadata for the 13 items for each of the four systems in the MUSHRA listening test.

192 kb/s, 256 kb/s, and 384 kb/s. The reconstructed objects at the output of the JOC decoder were rendered for playback on an immersive 7.1.4 loudspeaker configuration described in ITU-R BS.2051 [14] as Sound System G, except that the left and right "screen" channels were omitted. The open and hidden references as well as the 3.5 kHz and 7.0 kHz lowpass anchors were rendered directly from the original object-based immersive content. The average bit rate of the side information conveying JOC parameters and object metadata for the 13 items in the MUSHRA listening test is given for each of the four systems in this test in Tab. 2.

The rigthmost column of Fig. 6 shows the average score over all items for the 9 expert listeners after post-screening. It can be seen that AC-4 A-JOC at 384 kb/s achieves excellent quality on the MUSHRA scale. Also when studying the per-item results, the quality of this system is always in the excellent range even for the most critical items. For AC-4 A-JOC at 256 kb/s and 192 kb/s, the average quality is on the border between excellent and good, and in the upper half of the good-range of the MUSHRA scale, respectively. The quality achieved by the DD+ JOC system at 384 kb/s is typically in between the quality of the AC-4 A-JOC system at 192 kb/s and 256 kb/s, which is explained by the fact that the AC-4 A-JOC system uses more advanced technologies and has less compatibility constraints than the earlier DD+ JOC system.

5 Conclusions

This paper presented a joint object coding paradigm that enables the delivery of object-based immersive audio content at low bit rates. Two applications implementing the JOC paradigm were described, and listening test results showed that this paradigm is capable of delivering excellent quality when operated at bitrates of 256 kb/s and above. Since the underlying parameterization and in particular the bitstream format used in the AC-4 A-JOC system are very flexible, efficient encoding of object-based immersive audio content is also possible at bit rates below and above the range of operation points assessed in the listening test reported here.

References

[1] Riedmiller, J., Mehta, S., Tsingos, N., and Boon, P., "Immersive and Personalized Audio: A Practical System for Enabling Interchange, Distribution, and Delivery of Next-Generation Audio Experiences," *SMPTE Motion Imaging Journal*, 124(5), pp. 1–23, 2015.

[2] Robinson, C., Tsingos, N., and Mehta, S., "Scalable Format and Tools to Extend the Possibilities of Cinema Audio," *SMPTE Motion Imaging Journal*, 121(85), pp. 63–69, 2012.

[3] Dolby Laboratories, "Dolby Atmos," 2016, available: `http://www.dolby.com/us/en/brands/dolby-atmos.html`.

[4] Kjörling, K. et al., "AC-4 — The Next Generation Audio Codec," in *Audio Engineering Society Convention 140*, 2016.

[5] Villemoes, L., Purnhagen, H., Lehtonen, H.-M., and Klejsa, J., "Parametric Joint Channel Coding of Immersive Audio," Under preparation, 2016.

[6] Dolby Laboratories, "Dolby Atmos for the Home Theater," 2015, available: `http://www.dolby.com/us/en/technologies/dolby-atmos/dolby-atmos-for-the-home-theater.pdf`.

[7] "Digital Audio Compression (AC-4) Standard; Part 1: Channel based coding," ETSI TS 103 190-1 V1.2.1, 2015.

[8] Herre, J., Purnhagen, H., Koppens, J., Hellmuth, O., Engdegård, J., Hilper, J., Villemoes, L., Terentiv, L., Falch, C., Hölzer, A., Valero, M. L., Resch, B., Mundt, H., and Oh, H.-O., "MPEG Spatial Audio Object Coding—The ISO/MPEG Standard for Efficient Coding of Interactive Audio Scenes," *J. Audio Eng. Soc*, 60(9), pp. 655–673, 2012.

[9] Murtaza, A., Herre, J., Paulus, J., Terentiv, L., Fuchs, H., and Disch, S., "ISO/MPEG-H 3D Audio: SAOC 3D Decoding and Rendering," in *Audio Engineering Society Convention 139*, 2015.

[10] Villemoes, L., Hirvonen, T., and Purnhagen, H., "Decorrelation for Audio Object Coding," Under preparation, 2016.

[11] "Digital Audio Compression (AC-4) Standard; Part 2: Immersive and personalized audio," ETSI TS 103 190-2 V1.1.1, 2015.

[12] Herre, J., Hilpert, J., Kuntz, A., and Plogsties, J., "MPEG-H Audio—The New Standard for Universal Spatial/3D Audio Coding," *J. Audio Eng. Soc*, 62(12), pp. 821–830, 2015.

[13] "Method for the subjective assessment of intermediate quality levels of coding systems," Recommendation ITU-R BS.1534-3, 2015.

[14] "Advanced sound system for programme production," Recommendation ITU-R BS.2051-0, 2014.