IMMERSIVE AUDIO CODING FOR VIRTUAL REALITY USING A METADATA-ASSISTED EXTENSION OF THE 3GPP EVS CODEC

D. McGrath[*], *S. Bruhn*[+], *H. Purnhagen*[+], *M. Eckert*[*], *J. Torres*[*], *S. Brown*[*], *D. Darcy*[†]

[*]Dolby Australia Pty. Ltd., Sydney, Australia, [+]Dolby Sweden AB, Stockholm, Sweden,
[†]Dolby Laboratories, Inc., San Francisco, USA

ABSTRACT

Virtual Reality (VR) audio scenes may be composed of a very large number of audio elements, including dynamic audio objects, fixed audio channels and scene-based audio elements such as Higher Order Ambisonics (HOA). Potentially, the subjective listening experience may be replicated using a compact spatial format with a set number of dynamic objects and scene-based elements, retaining only the perceptual essence of the audio scene. The compact format would further enable a reduction in the complexity of subsequent compression and rendering. This paper investigates these hypotheses by exploring the use of a compact format that consists of up to four dynamic objects and nine HOA channels, with the Enhanced Voice Services (EVS) codec being applied to a 4-channel down-mix of the compact format.

***Index Terms*—** Audio Coding, Virtual Reality, Spatial Audio, Immersive Audio, Ambisonics

1. INTRODUCTION

Recent efforts within the 3GPP standards organization are directed towards the enabling of virtual reality (VR) media services over 3GPP networks. A study finalized in 2017 [1] identified various use cases and categorized them broadly into a) mobile endpoint consumption of managed and third-party VR content, and b) user-generated VR content. The study concluded that there were gaps in the existing set of 3GPP specifications to adequately represent the audio component of VR media. Consequently, 3GPP launched the "VRStream" work item [2] with the objective of specifying an efficient audio coding solution for the managed and third-party VR content use case category. Candidate solutions were submitted in July 2018, under a common understanding that in order to deliver adequate audio quality for the service, a solution would have to meet the excellent quality range in a series of MUSHRA tests as specified in [3].

Motivated by indications that many mobile endpoints will support the 3GPP enhanced voice services (EVS) mono codec [4, 5] in the near future, as well as by results demonstrating the suitability of the EVS codec for the coding of First Order Ambisonics (FOA) audio [1], the authors have chosen to examine the feasibility of meeting the VRStream audio objectives with an EVS-based solution. More specifically, the feasibility of an approach is investigated comprising encoding of a complex VR scene into a set of four audio channels, represented by individually encoded EVS bitstreams, along with metadata to guide the decoder to recover a perceptually relevant approximation of the original VR audio scene. This paper describes the details of this approach, called the Metadata-Assisted EVS Codec (MAEC) and provides listening test results that characterize the codec against the quality objectives.

2. MAEC OVERVIEW

An overview of the MAEC codec is provided in Figure 1. The codec contains the following key components:

1. The content ingestion engine (CIE) accepts inputs containing dynamic objects [6], Higher Order Ambisonics (HOA) [7] scenes, and channel-based content. The maximum number of objects, channels, and scenes is not constrained, but practical limits associated with memory generally mean that ingestion has a practical limit of 64 channels, 7^{th} order HOA, and 64 objects. The ingestion engine converts the input to a compact N^{th} order HOA scene plus M objects. The order of the HOA scene and the number of objects provided to the encoder are determined by a target bit rate for metadata stream. Metadata bit rates can range from 4 to 128kbps.
2. The Spatial Audio Reconstruction (SPAR) encoder performs operations such as the down-mix to a four channel FOA (B-Format) bed for encoding and metadata modeling for the residuals to HOA, object metadata, and metadata associated with energy compaction of the four FOA component signals WXYZ by means of prediction.
3. The SPAR decoder takes the four-channel energy compacted WX'Y'Z' stream and recreates the N^{th} order Ambisonics format and M objects that were specified by the CIE. This is a process inspired by Joint Object Coding [8] but here extended to work with a compact audio representation consisting of an HOA scene plus objects.
4. The N^{th} order HOA scene and M objects are presented to an audio renderer that may, for instance, produce output for headphones or a loudspeaker array.

3. CONTENT INGESTION ENGINE

The CIE takes a complex VR scene as input as described in Section 2, and produces an output consisting of M dynamic objects and an N^{th} order HOA stream. The total number of audio channels in the compact version of the VR scene will be $N_{total} = M + (N+1)^2$.

The dynamic objects have their (time-varying) location defined according to [2], wherein the metadata for each dynamic object is defined in terms of azimuth, elevation, radial distance and size. For all objects in the tests described in this paper, the object radius was 1.0 (a dimensionless quantity) and the object size was 0 (objects were treated as point sources).

In order to work well with the EVS encoder and decoder, the M dynamic objects had their location metadata quantized to sample-times that were a multiple of 1920 (two EVS frames at 48kHz audio sampling rate, or 40ms).

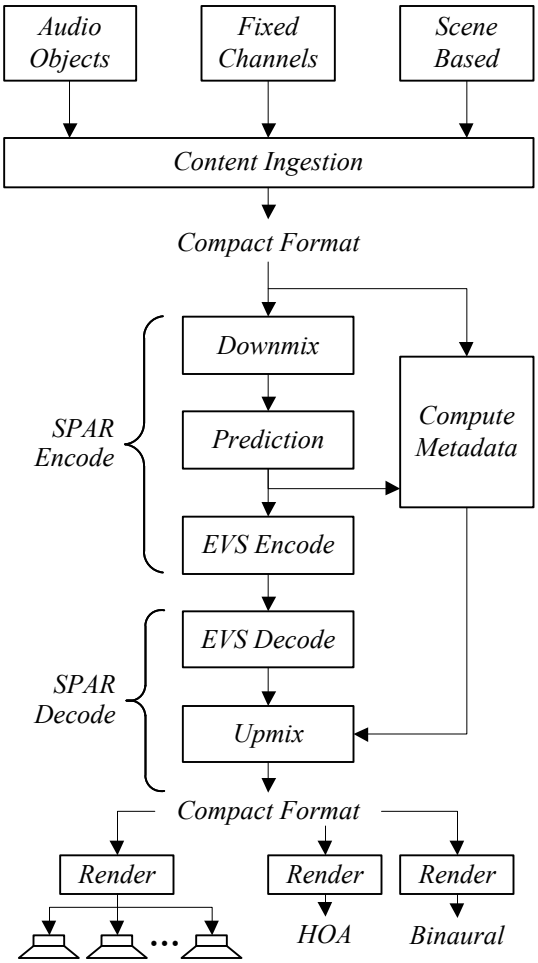

Fig. 1. High level block diagram of the MAEC System.

The output of the CIE is a multi-channel compact spatial signal, $\mathbf{x}(t)$. For the subsequent descriptions we will refer to the example where $M = 2$ and $N = 2$, such that the compact spatial signal is composed of 2 objects and a 9-channel HOA scene:

$$\mathbf{x}(t) = [obj_1(t), obj_2(t), hoa_1(t), \cdots, hoa_9(t)]^T. \quad (1)$$

In addition, the metadata for each dynamic object is defined as a discrete set of unit-vectors:

$$\Theta_o(k) = [x_o(k), y_o(k), z_o(k)]^T, \quad (2)$$

where $x_o(k)^2 + y_o(k)^2 + z_o(k)^2 = 1$ and the metadata $\Theta_o(k)$ is indicative of the location for object $o \in \{1, 2, \ldots, M\}$ over the time interval $kT \leq t < (k+1)T$, where T is the metadata block interval, 40ms.

4. THE SPAR ENCODER AND DECODER

The SPAR encoder and decoder both operate on the audio signal in overlapping blocks, where the block stride is $T = 40$ms. The window function $w(t)$ is shown in Figure 2. The leading and trailing edges of this window are composed from segments of a raised sine function.

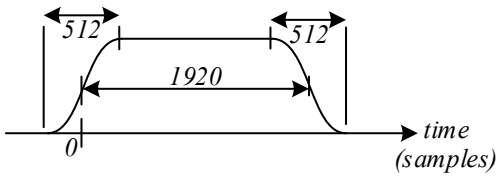

Fig. 2. Window function: $w(t)$.

In the following description of the SPAR encoder and decoder, metadata values and matrix coefficients will be variously referred to as functions of the block number $k \in \mathbb{Z}$ or as functions of time, $t \in \mathbb{R}$, where the meaning will be clear from the context, and where the time-based function may be computed according to the window function $w(t)$. For example, a metadata value such as $x_1(k)$ may be alternatively referred to as a function of time:

$$\underset{(\mathbb{R}\to\mathbb{R})}{x_1(t)} = \sum_{k=-\infty}^{\infty} w(t - kT) \underset{(\mathbb{Z}\to\mathbb{R})}{x_1(k)}. \quad (3)$$

In addition, short-term Fourier domain (STF) versions of the signals will be referred to using the capitalized variable name, so that, for time-block k, the STF version of the signal $\mathbf{x}(t)$ would be:

$$\mathbf{X}(k, \omega) = FFT\{w(t)\mathbf{x}(t - kT)\}. \quad (4)$$

4.1. SPAR Encoder Signal Path

The SPAR encoder begins by down-mixing the compact signal $\mathbf{x}(t)$ to produce the FOA signal $\mathbf{y}(t)$, where

$$\mathbf{y}(t) = [foa_W(t), foa_X(t), foa_Y(t), foa_Z(t)]^T \quad (5)$$
$$= \mathbf{M}(t) \times \mathbf{x}(t), \quad (6)$$

and the (time-varying) down-mix matrix is:

$$\mathbf{M}(t) = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ x_1(t) & x_2(t) & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ y_1(t) & y_2(t) & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ z_1(t) & z_2(t) & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \quad (7)$$

This down-mix matrix creates the FOA signal by panning the two objects according to their direction, and "truncating" the 9-channel HOA signal to four channels.

For each block k, the SPAR encoder forms a predictor in the frequency domain:

$$\mathbf{PRED}_X(k, \omega) = \mathbf{FOA}_W(k, \omega) P_X(k, \omega) \quad (8)$$
$$\mathbf{PRED}_Y(k, \omega) = \mathbf{FOA}_W(k, \omega) P_Y(k, \omega) \quad (9)$$
$$\mathbf{PRED}_Z(k, \omega) = \mathbf{FOA}_W(k, \omega) P_Z(k, \omega), \quad (10)$$

where the prediction frequency-responses ($P_X(k, \omega)$ et al.) are defined in terms of smooth functions for each block k. A "spatially

whitened" FOA signal $\mathbf{y}'(t)$ is then created by subtracting the predicted functions from the original FOA signals:

$$\mathbf{y}'(t) = [foa'_W(t), foa'_X(t), foa'_Y(t), foa'_Z(t)]^T \quad (11)$$

$$= \begin{bmatrix} foa_X(t) \\ foa_X(t) - pred_X(t) \\ foa_Y(t) - pred_Y(t) \\ foa_Z(t) - pred_Z(t) \end{bmatrix}. \quad (12)$$

This spatially whitened FOA signal, $\mathbf{y}'(t)$ is then passed to four independent EVS encoders, with the bit-rate of the $foa_W(t)$ signal being higher than the bit-rate of each of the "whitened" signals, as shown in Table 1.

4.2. SPAR Decoder Signal Path

The SPAR Decoder begins by decoding the four EVS streams to produce the post-EVS signals:

$$\mathbf{y}''(t) = [foa''_W(t), foa''_X(t), foa''_Y(t), foa''_Z(t)]^T. \quad (13)$$

The first channel, $foa''_W(t)$, is then filtered by D decorrelators to produce a D-channel decorrelated signal set:

$$\mathbf{dec}(t) = [dec_1(t), \ldots, dec_D(t)]^T. \quad (14)$$

The compact signals are then derived by mixing the $\mathbf{y}''(t)$ and $\mathbf{dec}(t)$ signals. This mixing is carried out in the frequency-domain:

$$\mathbf{X}'(k,\omega) = \mathbf{C}(k,\omega) \times \mathbf{Y}(k,\omega) + \mathbf{P}(k,\omega) \times \mathbf{D}(k,\omega), \quad (15)$$

where $\mathbf{C}(k,\omega)$ and $\mathbf{P}(k,\omega)$ are frequency-dependent matrices of size $[11 \times 4]$ and $[11 \times D]$, respectively. These matrices are derived by the SPAR metadata encoder (Section 4.3) and transmitted to the SPAR decoder in the form of matrix coefficients for B band center-frequencies. For the experiments described in this paper, the number of sub-bands was $B = 12$.

The final 11-channel time-domain output is:

$$\mathbf{x}'(t) = [obj'_1(t), obj'_2(t), hoa'_1(t), \cdots, hoa'_9(t)]^T \quad (16)$$

4.3. The SPAR Metadata Encoder

In parallel with the SPAR encoder signal path, the SPAR metadata encoder examines the $\mathbf{x}(t)$ and $\mathbf{y}'(t)$ signals, and derives the up-mix matrices $\mathbf{C}(k,\omega)$ and $\mathbf{P}(k,\omega)$, which are used in the SPAR decoder, as defined in Section 4.2.

The metadata is determined for a set of B frequency sub-bands. For each time-block k and sub-band $b \in \{1, \ldots, B\}$, the bandpass filtered signals $\mathbf{x}_{b,k}(t)$ and $\mathbf{y}'_{b,k}(t)$ are derived and the primary up-mixing matrix is determined:

$$\mathbf{C}_{b,k} = \underset{\mathbf{C}}{\operatorname{argmin}} \int_t \left|\left|\mathbf{C} \times \mathbf{y}'_{b,k}(t) - \mathbf{x}_{b,k}(t)\right|\right|_2 \quad (17)$$

Whilst the matrix $\mathbf{C}_{b,k}$ provides the least-squares optimum up-mix matrix for band b, poor correlation conditions may result in energy loss of the synthesized upmix signal. To remedy this, the covariance of the error signal, $\mathbf{C}_{b,k} \times \mathbf{y}'(t) - \mathbf{x}(t)$, may be partially replicated by the application of the matrix $\mathbf{P}(k,\omega)$ in the SPAR Decoder. This matrix is determined, for each sub-band, by principal component analysis [9].

The resulting sub-band matrices $\mathbf{C}_{b,k}$ and $\mathbf{P}_{b,k}$ are quantized and differentially Huffman encoded, to produce the metadata that is sent to the decoder along with the EVS-coded data.

Table 1. The compact format HOA order N and number of objects M, target overall bitrates for FOA and HIQ operating modes, individual EVS bitrates for the energy compacted WX'Y'Z' channels, and maximum metadata (MD) bitrate.

	Compact N	Compact M	Target (kbps)	EVS Bitrates (kbps) W	X'	Y'	Z'	MD (kbps)
FOA	1	0	81.2	32	16.4	16.4	16.4	4
	1	0	128	48	24.4	24.4	24.4	4
HIQ	2	2	128	48	24.4	24.4	24.4	40
	2	2	256	96	48	48	48	40
	2	4	384	128	64	96	64	50
	2	4	512	128	128	128	128	50

Table 2. Number of listeners used by each laboratory for each FOA and HIQ test (Dolby/Other).

	Test 1 - Loudspeaker	Test 2 - Headphone
FOA	10 / 9[a]	10 / 10[b]
HIQ	12 / 9[a]	10 / 12[c]

[a] Nokia. [b] Ericsson. [c] Philips.

5. AUDIO QUALITY EVALUATIONS

Listening test results are reported for experiments with the following operating modes:

- FOA mode, where the reference signal is a FOA scene,
- HIQ mode, where the reference signal is a complex spatial scene.

The precise compact formats and target bitrates tested for each operating mode are described in Table 1. These bitrates were divided among the four down-mix channels and the metadata. The target values exclude an additional bitrate margin of $< 10\%$. The specifics of each configuration were determined prior to the formal tests by preliminary informal listening, and are also shown in Table 1.

MUSHRA listening tests [11] were conducted to determine the basic audio quality in accordance with [3]. In line with requirements defined in [2], the test material comprised a selection of channel-, object, and scene-based (HOA) audio, and combinations thereof. Altogether, twenty spatial scenes were processed, comparing a hidden reference, and 3.5kHz and 7.0kHz anchors with the MAEC outputs at the above-mentioned bitrates. The reference conditions were the source test material items directly rendered through the same renderer as the test conditions.

The scenes were rendered to a 7.1.4 loudspeaker array in Test 1, and binaurally for headphone playback in Test 2. Corroborating test results from independent cross-checks are reproduced here with the permission of Ericsson, Nokia and Philips laboratories [10]. These parties were not involved in the development or submission of Dolby's MAEC solution, and did not submit a competing solution. Table 2 indicates the number of expert listeners used by each laboratory for each test. All subjects were pre- and post-screened according to ITU-R BS.1534-3 guidelines [11].

Figure 3 depicts the mean MUSHRA scores over the 20 test inputs and 95% confidence intervals for each test configuration, given a two-sided student's t-distribution. The results for all references and anchors are in line with ITU-R recommendation BS.1534-3. Results

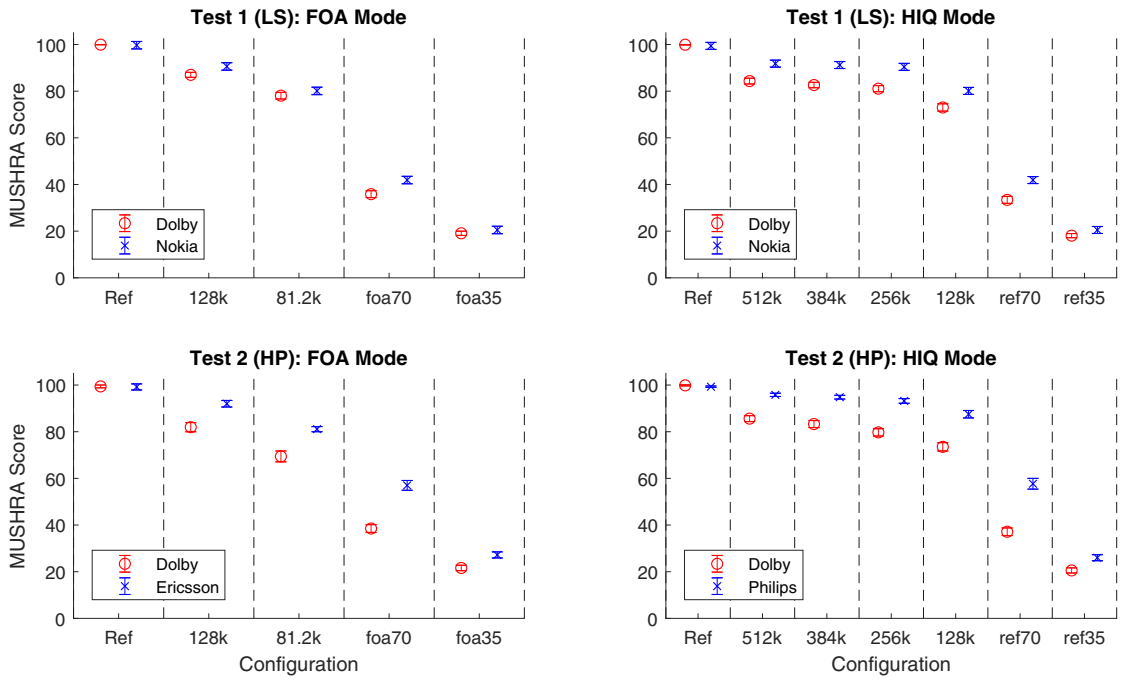

Fig. 3. Loudspeaker (LS) and headphone (HP) listening test results for FOA and HIQ modes from Dolby and cross-check laboratories, courtesy of Ericsson, Nokia and Philips.

for the FOA mode above 128kbps and HIQ mode above 256kbps lie within the excellent MUSHRA region (scores of 80 or above).

Examining the HIQ-mode subjective test results in more detail, a quality saturation effect at bitrates above 256 kbps is observed. Increasing the bitrate beyond that point does not lead to commensurate quality enhancements tending towards fully transparent quality. The authors see several potential causes for this behavior that should be investigated in the future. Potential reasons include imperfections in the conversion from a complex VR audio scene to the compact representation, a compact representation with potentially too few objects or too low HOA order, and limitations due to the 4-channel FOA down-mix in terms of both the down-mix matrix $\mathbf{M}(t)$ from equation (7) used and the number of down-mix channels.

It is also worth commenting on the apparent differences between the results of different test labs. The authors believe that this is an effect often experienced when comparing absolute ratings from different subjective listening tests done in different situational contexts. One specific reason for the observed deviations may be inter-cultural differences among the listener panels of the labs situated in different geographical regions.

6. CONCLUSIONS

MAEC has been presented as a solution for efficient coding of VR audio content, building on the 3GPP EVS codec. In-house quality evaluations following a 3GPP endorsed test plan, and verified by independent cross-checks, confirm that a representation with only 4 EVS encoded downmix channels plus spatial metadata enables excellent audio quality at or above 256kbps. This quality level is deemed adequate for 3GPP VR media services with mobile endpoint consumption of managed and third-party VR content. One essential MAEC component is the content ingestion engine that converts arbitrarily complex VR audio scenes, which may be composed of a very large number of audio elements, into a compact format with very few audio objects and a low-order HOA scene. Test results show that this conversion can effectively remove perceptually irrelevant spatial information prior to the actual coding stage.

7. ACKNOWLEDGMENTS

The authors wish to thank the independent cross-check labs at Ericsson, Nokia and Philips, who contributed to the test campaign of the MAEC system.

8. REFERENCES

[1] 3GPP. Virtual reality (VR) media services over 3GPP. Technical Report 26.918, 3rd Generation Partnership Project (3GPP), 03 2018. Version 15.2.0.

[2] 3GPP. Revision of DRAFT VRStream-2 - Submission Process for VRStream Audio Profiles. Technical Report AHVIC-139, 3rd Generation Partnership Project (3GPP), 04 2018. Version 0.3.

[3] 3GPP. Subjective test methodologies for the evaluation of immersive audio systems. Technical Specification (TS) 26.259, 3rd Generation Partnership Project (3GPP), 09 2018. Version 15.0.0.

[4] 3GPP. Codec for Enhanced Voice Services (EVS); General overview. Technical Specification (TS) 26.441, 3rd Generation Partnership Project (3GPP), 06 2018. Version 15.0.0.

[5] S. Bruhn, H. Pobloth, M. Schnell, B. Grill, J. Gibbs, L. Miao, K. Järvinen, L. Laaksonen, N. Harada, N. Naka, S. Ragot, S. Proust, T. Sanda, I. Varga, C. Greer, M. Jelínek, M. Xie, and P. Usai. Standardization of the new 3GPP EVS codec. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5703–5707, April 2015.

[6] C. Q. Robinson, S. Mehta, and N. Tsingos. Scalable format and tools to extend the possibilities of cinema audio. In *The 2012 Annual Technical Conference Exhibition*, pages 1–12, Oct 2012.

[7] Michael A. Gerzon. Periphony: With-height sound reproduction. *J. Audio Eng. Soc*, 21(1):2–10, 1973.

[8] H. Purnhagen, T. Hirvonen, L. Villemoes, J. Samuelsson, and J. Klejsa. Immersive audio delivery using joint object coding. In *Audio Engineering Society Convention 140*, May 2016.

[9] L. Villemoes, T. Hirvonen, and H. Purnhagen. Decorrelation for audio object coding. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 706–710, March 2017.

[10] 3GPP. Virtual Reality (VR) streaming audio; Characterization test results. Technical Specification (TS) 26.818, 3rd Generation Partnership Project (3GPP), 09 2018. Version 15.0.0.

[11] ITU-R. Method for the subjective assessment of intermediate quality levels of coding systems. Technical Report BS.1534-3, Radiocommunication Sector of International Telecommunications Union (ITU-R), 2015.