ETSI TR 126 918 V15.2.0 (2018-07)

Universal Mobile Telecommunications System (UMTS);
LTE;
Virtual Reality (VR) media services over 3GPP
(3GPP TR 26.918 version 15.2.0 Release 15)

Reference
RTR/TSGS-0426918vf20

Keywords
LTE,UMTS

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

Foreword

This Technical Report (TR) has been produced by ETSI 3rd Generation Partnership Project (3GPP).

The present document may refer to technical specifications or reports using their 3GPP identities, UMTS identities or GSM identities. These should be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between GSM, UMTS, 3GPP and ETSI identities can be found under http://webapp.etsi.org/key/queryform.asp.

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Contents

Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 or greater indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the document.

Introduction

Virtual Reality (VR) is the ability to be virtually present in a space created by the rendering of natural and/or synthetic image and sound correlated by the movements of the immersed user allowing interacting with that world.

The immersive multimedia experience has been an exploration topic for several years. With the recent progress made in rendering devices, such as Head mounted displays (HMD), a significant quality of experience can be offered.

Before any possible standardization, it is necessary to study the field:

- to understand how the equipment used for creating such an immersive experience works, e.g. by collecting information on the optical systems and audio rendering processes;
- to evaluate the relevance to Virtual Reality for 3GPP;
- to identify the possible points of interoperability, and hence potential standardization.

Use cases for Virtual Reality need to be listed and mapped to the already existing 3GPP services if applicable.

Media formats required for providing the immersive experience need to be identified and potentially evaluated subjectively so as to extract requirements on minimum device and network capabilities.

1 Scope

The scope of the present document is to investigate the relevance of Virtual Reality in the context of 3GPP. Virtual Reality is the ability to be virtually present in a non-physical world created by the rendering of natural and/or synthetic image and sound correlated by the movements of the immersed user allowing to interact with that world. With recent progress made in rendering devices, such as Head mounted displays (HMD), a significant quality of experience can be offered. By collecting comprehensive information on VR use cases, existing technologies and subjective quality, the report attempts to identify potential gaps and relevant interoperability points that may require further work and potential standardization in 3GPP in order to support VR use cases and experiences in 3GPP user services. The report primarily focuses on 360 degrees video and associated audio, which support three degrees of freedom (3DOF).

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[2] Savino, Peter J.; Danesh-Meyer, Helen V. (1 May 2012). Color Atlas and Synopsis of Clinical Ophthalmology -- Wills Eye Institute -- Neuro-Ophthalmology. Lippincott Williams & Wilkins. p. 12.

[3] Dagnelie, Gislin (21 February 2011). Visual Prosthetics: Physiology, Bioengineering, Rehabilitation. Springer Science & Business Media. p. 398.

[4] T.E. Boult and G. Wolberg, "Correcting chromatic aberrations using image warping", *Computer Vision and Pattern Recognition 1992. Proceedings CVPR '92. 1992 IEEE Computer Society Conference on*, pp. 684-687, 1992, ISSN 1063-6919.

[5] Levoy, Marc, and Pat Hanrahan. "Light field rendering". Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. ACM, 1996.

[6] J. P. Snyder, "Flattening the Earth: Two Thousand Years of Map Projections," University of Chicago Press, 1993.

[7] Skupin, R., Sanchez, Y., Hellge, C., & Schierl, T., "Tile Based HEVC Video for Head Mounted Displays", IEEE International Symposium on Multimedia (ISM), December 2016.

[8] Michael A. Gerzon, *Periphony: With-Height Sound Reproduction*. Journal of the Audio Engineering Society, 1973, 21(1):2-10.

[9] Recommendation ITU-R BS.2266-2 (04/2014): "Framework of future audio broadcasting systems".

[10] Recommendation ITU-R BS.2051-0 (02/2014): "Advanced sound system for programme production".

[11] Lee et al., "Scalable Multiband Binaural Renderer for MPEG-H 3D Audio", in *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 907-920, Aug. 2015.

[12] Breebaart, J., Nater, F., and Kohlrausch, A., "Parametric binaural synthesis: Background, applications and standards," *NAG/DAGA International Conference on Acoustics*, 2009.

[13] Brungart, D. S., Kordik, A.K., and Simpson, B.D - "Effects of headtracker latency in virtual audio displays", J. Audio Eng. Soc., Vol. 54, No. 1/2, 2006 January/February.

[14] Lindau, A. - "The Perception of System Latency in Dynamic Binaural Synthesis", In *Fortschritte der Akustik: Tagungsband der 35*, NAG/DAGA pp. 1063-1066 - Rotterdam - 2009.

[15] Lindau, A., et al., "A Spatial Audio Quality Inventory (SAQI)", Acta Acustica united with Acustica, 100(5), 2014.

[16] Silzle, A., "3D Audio Quality Evaluation: Theory and Practice", *ICSA*, 2014.

[17] C. Schissler, A. Nicholls & R. Mehra, "Efficient HRTF-based Spatial Audio for Area and Volumetric Sources", IEEE Trans on Visualization and Computer Graphics, 2016, Vol 22, No 4, pp 1356 - 1366.

[18] K. E. Bystrom, W. Barfield and C. Hendrix, "A Conceptual Model of the Sense of Presence in Virtual Environments," in Presence, vol. 8, no. 2, pp. 241-244, April 1999.

[19] Sheridan, Thomas B. "Musings on telepresence and virtual presence". Presence: Teleoperators & Virtual Environments 1, no. 1 (1992): 120-126.

[20] Sheridan, T., Zeltzer, D., & Slater, M. (1995). "Presence and performance within virtual environments", In W. Barfield and T. Furness III (Eds.), Virtual Environments and Advanced Interface Design. (pp. 473-513). Oxford University Press.

[21] Lindau, A., Weinzierl, S., (2011). "Assessing the plausibility of virtual acoustic environments", Forum Acusticum 2011, 27 June -1 July, Aalborg., European Acoustic Association, pp. 1187-1192.

[22] Recommendation ITU-R BS.1534-3 (10/2015): "Method for the subjective assessment of intermediate quality level of audio systems".

[23] Bertet, Stéphanie, et al. "Influence of Microphone and Loudspeaker Setup on Perceived Higher Order Ambisonics Sound Field", in Ambisonics Symposium 2009, June 25-27, Graz.

[24] Durlach N.I, Mavor A.S, "Virtual reality Scientific and Technological Challenges" National Academy Press, 1995.

[25] Draper M.H, "The adaptive effects of virtual interfaces: Vestibulo-ocular reflex and simulator sickness", PhD Thesis, University Washington, Sponsored by US Airforce Department, 1995.

[26] Di Girolamo S, Picciotti P, Sergi B, Di Nardo W, Paludetti G, Ottaviani F "Vestibulo-ocular reflex modification after virtual environment exposure.", Acta Oto-Laryngologica. 2001 Jan; Vol 121 Issue 2 pp 211-215.

[27] Jombı'k P, Bahy'l V, "Short latency disconjugate vestibulo-ocular responses to transient stimuli in the audio frequency range." Journal of Neurology, Neurosurgery, and Psychiatry; Vol 76 No 10, Oct 2005 pp 1398-1402.

[28] Amin M.S, "Vestibuloocular Reflex Testing" Medscape Article Number 1836134, Feb 10 2016.

[29] Jerald J, "Scene-Motion- and Latency-Perception Thresholds for Head-Mounted Displays" University of North Carolina PhD Dissertation, 2010.

[30] Brungart et al., "Effects of Headtracker Latency in Virtual Audio Displays", Proceedings of International Conference on Auditory Display; ICAD-05, July 2005.

[31] http://www.york.ac.uk/inst/mustech/3d_audio/higher_order_ambisonics.pdf.

[32] Recommendation ITU-R BT.1359-1 (1998): "Relative Timing of Sound and Vision for Broadcasting".

[33] EBU Technical Recommendation R37 (2007): "The relative timing of the sound and vision components of a television signal".

[34] 3GPP TS 26.116: "Television (TV) over *3GPP* services; Video profiles".

[35] ISO/IEC AWI 23090-2: "Coded representation of immersive media -- Part 2: Application Format for Omnidirectional Media" - OMAF.

[36] MPEG-H 3D Audio Verification Test Report, Geneva, 2017. http://mpeg.chiariglione.org/standards/mpeg-h/3d-audio/mpeg-h-3d-audio-verification-test-report.

[37] 3GPP TS 26.114: "IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction".

[38] 3GPP TS 26.223: "Telepresence using the IP Multimedia Subsystem (IMS); Media handling and interaction".

[39] 3GPP TS 26.247: "Transparent end-to-end Packet-switched, Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH)".

[40] Jerome Daniel, Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format, 23rd AES Conference, Copenhagen 2003, p. 13.

[41] Recommendation ITU-R BS.2076-0 (06/2015): "Audio Definition Model".

[42] ETSI TS 103 223 (V1.1.1): "MDA; Object-Based Audio Immersive Sound Metadata and Bitstream".

[43] C. Nachbar, F. Zotter, E. Deleflie & A. Sontacchi, "AmbiX - A Suggested Ambisonics Format" Procedings of Ambisonics Symposium 2011, June 2-3, 2011, Lexington, USA.

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

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

[46] ISO/IEC 23008-3:2015: "Information technology -- High efficiency coding and media delivery in heterogeneous environments -- Part 3: 3D audio".

[47] ISO/IEC 23008-3:2015/Amd 3:2017; "MPEG-H 3D Audio Phase 2".

[48] Wenzel, E.M. - "Analysis of the Role of Update rate and System Latency in Interactive Virtual Acoustic Environments", 103rd AES Convention, New York, 1997.

[49] Sandvad, J. - Dynamic Aspects of Auditory Virtual Environments, 100th AES Convention, Copenhagen, 1996 May 11-14.

[50] Michael Chapman et al., A standard for interchange of Ambisonic signal sets, Ambisonics Symposium, Graz 2009.

[51] Stitt, P., Hendrickx, E., Messonier, J.C., Katz, B. FG - "Effect of Head Tracking Latency on Binaural Rendering in Simple and Complex Sound Scenes", 140th AES Convention, 2016 June 4-7, Paris, France.

[52] Satoshi Yairi, Yukio Iwaya and Yoiti Suzuki - "Investigation of system latency detection threshold of virtual auditory display", Proceedings of the 12th International Conference on Auditory Display, London, UK June 20 - 23, 2006.

[53] Wenzel, E.M., "Effect of increasing system latency on localization of virtual sounds with short and long duration", Proceedings of the 7th International Conference on Auditory Display, pp.185-190, 2001.

[54] Recommendation ITU-R BT.2021-1 (02/2015): "Subjective methods for the assessment of stereoscopic 3DTV systems".

[55] Recommendation ITU-T P.915 (03/2016): "Subjective assessment methods for 3D video".

[56] Brungart, D. S., Kordik, A.K., and Simpson, B.D - "The Detectability of Headtracker Latency in Virtual Auditory Displays" In: *Proc. Of ICAD 2005 - 11th Meeting of the International Conference on Auditory Display*, Limerick, Ireland, July 6-9, 2005.

[57] Levitt,H. - "Transformed up-down methods in psycho-acoustics." J.Acoust Soc America,49(2), 467-477, 1970.

[58] Mackensen, P. - "Auditive Localization, Head Movements, an additional cue in Localization." PhD thesis, TU-Berlin, 2004.

[59] Klein, S. - "Measuring, estimating, and understanding the psychometric function: A commentary.", Perception & Psychophysics, 63(8) pp.1421-1455, 2001.

[60] Recommendation ITU-R BT.1788 (01/2007): "Methodology for the subjective assessment of video quality in multimedia applications".

[61] Recommendation ITU-R BT.500-13 (01/2012): "Methodology for the subjective assessment of the quality of television pictures".

[62] Recommendation ITU-T P.910 (04/2008): "Subjective video quality assessment methods for multimedia applications".

[63] E. Upenik, M. Rerabek and T. Ebrahimi. A testbed for subjective evaluation of omnidirectonal visual content, *32nd Picture Coding Symposium (PCS)*, 4-7 December 2016, Nuremberg, Germany.

[64] Purnhagen H., Hirvonen T., Villemoes L., Samuelsson J., Klejsa J., "Immersive Audio Delivery Using Joint Object Coding" AES 140th Convention 2016 June 4-7, Paris, France.

[65] ITU-T COM 12-C97-E, October 2009, “New subjective test method for evaluation of overall and spatial sound quality”.

[66] Recommendation ITU-T P.806 (02/2014): “A subjective quality test methodology using multiple rating scales”.

[67] 3GPP TS 26.401 V14.0.0 (2017-03) – General audio codec audio processing functions; Enhanced aacPlus general audio codec; General description (Release 14).

[68] 3GPP TS 26.441: "Codec for Enhanced Voice Services (EVS); General overview".

[69] 3GPP TR 26.952: “Codec for Enhanced Voice Services (EVS); Performance characterization”.

3 Definitions and abbreviations

3.1 Definitions

For the purposes of the present document, the terms and definitions given in 3GPP TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in 3GPP TR 21.905 [1].

diegetic: part of the VR scene and rendered according to HMD head-tracking information

non-diegetic: independent of the VR scene and rendered independently of HMD head-tracking information

3.2 Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in 3GPP TR 21.905 [1].

3DOF	3 Degrees of freedom
6DOF	6 Degrees of freedom
ACN	Ambisonic Channel Number
AOP	Acoustic Overload Point
BRIR	Binaural Room Impulse Response
CBA	Channel-Based Audio
CICP	Coding Independent Code Point

CMAF	Common Media Application Format
DASH	Dynamic Adaptive Streaming over HTTP
DAW	Digital Audio Workstation
EPG	Electronic Program Guide
ERP	Equirectangular projection
FOA	First Order Ambisonics
FOV	Field of view
HMD	Head Mounted Display
HOA	High Order Ambisonics
HRTF	Head-related transfer function
M2S	Motion to Sound
MBMS	Multimedia Broadcast Multicast Service
MDA	Multi-Dimensional Audio
mTSL	minimum Total System Latency
OBA	Object-Based Audio
OMAF	Omnidirectional MediA Format
QoE	Quality of experience
SBA	Scene-Based Audio
SNR	Signal to Noise Ratio
SRM	Similarity Ring Metric
TV	TeleVision
VR	Virtual Reality

4 Introduction to Virtual Reality

4.1 Definition

4.1.1 Virtual reality

Virtual reality is a rendered version of a delivered visual and audio scene. The rendering is designed to mimic the visual and audio sensory stimuli of the real world as naturally as possible to an observer or user as they move within the limits defined by the application.

Virtual reality usually, but not necessarily, requires a user to wear a head mounted display (HMD), to completely replace the user's field of view with a simulated visual component, and to wear headphones, to provide the user with the accompanying audio. Some form of head and motion tracking of the user in VR is usually also necessary to allow the simulated visual and audio components to be updated in order to ensure that, from the user's perspective, items and sound sources remain consistent with the user's movements. Additional means to interact with the virtual reality simulation may be provided but are not strictly necessary.

4.1.2 Augmented reality

Augmented reality is when a user is provided with additional information or artificially generated items or content overlaid upon their current environment. Such additional information or content will usually be visual and/or audible and their observation of their current environment may be direct, with no intermediate sensing, processing and rendering, or indirect, where their perception of their environment is relayed via sensors and may be enhanced or processed.

Augmented reality and virtual reality are related and may be very similar, especially if the augmented reality is presented indirectly to the user. However, Augmented reality is out of scope of this study.

4.1.3 Mixed reality

Mixed reality can be considered as an advanced form of augmented reality where some virtual elements are inserted into the physical scene with the intent to provide the illusion that these elements are part of the real scene. Mixed reality is also out of scope of the present document.

4.2 Video systems

4.2.1 Introduction

Virtual reality has the promise to place users into immersive worlds that interact with their head movements. At the video level, this is achieved by providing a video experience that covers as much of the field of view (FOV) as possible together with the synchronization of the viewing angle of the rendered video with the head movements. Although many different types of devices may be able to provide such an experience, head mounted displays (HMD) are the most popular. They rely either on dedicated screens integrated into the system and running with external computers (*Tethered*) or on a smartphone inserted into the HMD (*Untethered*). The first approach has the advantages of only requiring lightweight screens and benefiting from a high computing capacity compared to smartphone-based systems, which offer a higher mobility and are less expensive to produce. In both cases, the video experience is generated the same way thanks to lenses-based systems as described in the following clauses as well as some basic principles on the Human Visual System.

4.2.2 Field of view

4.2.2.1 Definition

The Human field of view (FOV) is defined as the area of vision at a given moment (with a fixed head). It is the angle of visible field expressed in degrees measured from the focal point. The monocular FOV is the angle of the visible field of one of the eyes whereas the binocular FOV is the combination (not addition) of the two eyes fields.

4.2.2.2 Horizontal FOV

The horizontal monocular FOV is the addition of the *nasal FOV* (from pupil to nose, 60°) and the *temporal FOV* (from pupil to the side of the head, 100-110°) [2]. The monocular horizontal FOV is around 170° in average. The binocular horizontal FOV is around 200-220° degrees [3].

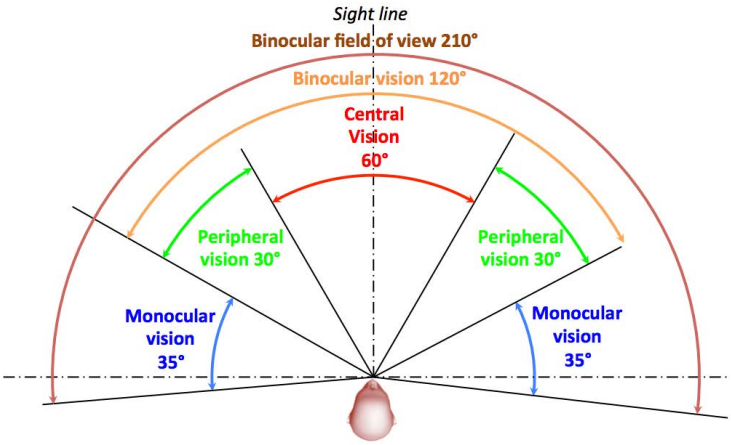


Figure 4.1: Horizontal human field of view

The central vision is also called the comfort zone where sensibility to details is the most important even if the maximum acuity is only a few degrees (3-5°) around the focal point (called the fovea zone). Although less sensible to definition, the peripheral vision is more receptive to movements. The common area covered by both monocular FOV is the area where depth perception is possible (binocular vision: 120°). Figure 4.1 summarizes the different viewing angles composing the horizontal FOV.

4.2.2.3 Vertical FOV

The vertical FOV is composed of the central vision area (60°) and the upper and lower peripheral visions (30° and 45° respectively) as illustrated in Figure 4.2.

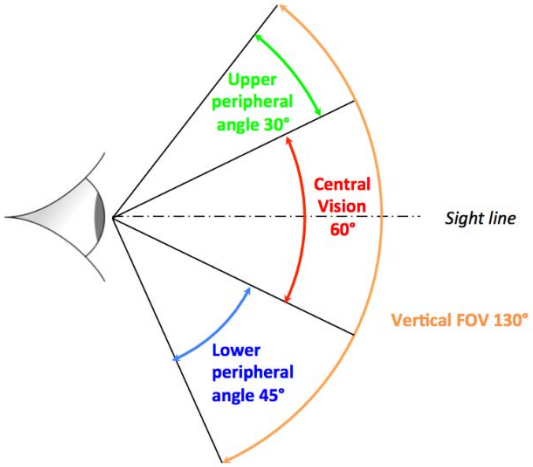


Figure 4.2: Vertical human field of view

The vertical FOV is typically around 135° [2]. For both eyes the combined visual field is 130-135° vertical and 200° horizontal [3].

4.2.3 Lenses for a wider FOV

In order to ensure an immersive experience, a large enough FOV is required. Due to the limited size of the screens, lenses are used in between the eyes and the screens in order to fill up the human field of view as much as possible. Figure 4.3 describes the principle.

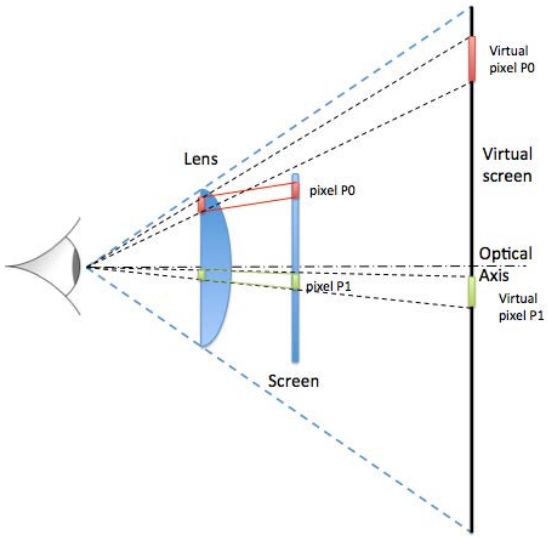


Figure 4.3: Use of lenses for VR

Due to the properties of the lenses, the perception of looking at a much larger scene is achieved. Each light ray is expanded through the lens.

4.2.4 Optical aberrations

4.2.4.1 Introduction

Although lenses are used so as to increase the field of view, the downside is that aberrations are introduced that need to be corrected or compensated so as to offer a good quality of experience. There are mainly two types of aberrations created by the lenses: The lens distortion and the chromatic aberration.

4.2.4.2 Lens distortion

When light crosses the lens, it is deviated from its original direction (refraction) proportionally to its distance from the axis of the lens (rays crossing the lens at its axis are not deviated) as illustrated in Figure 4.3. In such a case, images are spherically distorted; their corners stretch outwards and the lines start to curve. In Figure 4.3, pixel P0 is further from the axis than pixel P1. The projected pixel P0 on the virtual screen is then perceived larger than the projected pixel P1. This distortion is called the *Pincushion distortion*.

Pincushion distortion is a lens effect that causes images to become pinched to the centre. The Pincushion distortion effect increases with the object distance from the optical axis of the lens as shown in Figure 4.4.

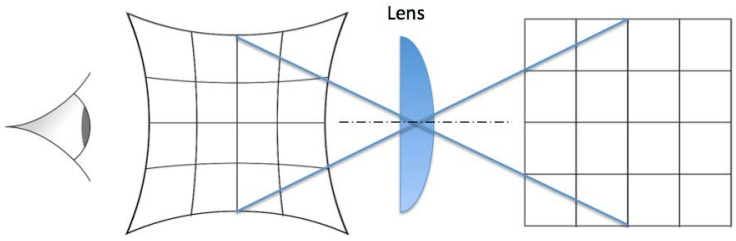


Figure 4.4: Illustration of the optical Pincushion effect

In order to compensate such an effect and remove the apparent distortion, the opposite distortion is applied on the display. This is called the *Barrel distortion*. The Barrel distortion is the diminution of the image magnification with the radial distance of every point from the optical axis (the further is a point from the centre, the higher its distance from it is reduced). Figure 4.5 depicts the resulted image when the Barrel distortion is applied on a screen and its result after a lens Pincushion on it.

The direct correlation of the field of view with the amount of distortion of the image is that the wider the field of view is, the more distorted the image is.

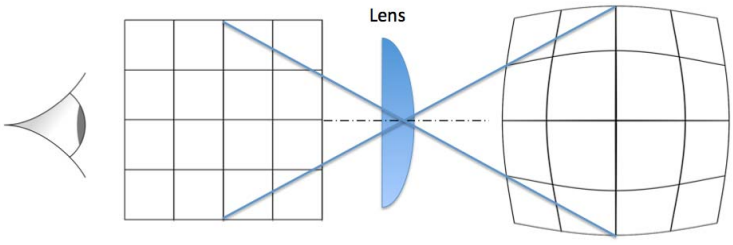


Figure 4.5: Barrel distortion for correcting the pincushion effect

Even if with such a process the apparent distortion is removed, there are mainly to issues introduced:

First, applying the Barrel/Pincushion combination implies a reduction of pixel density. Even if it barely remains the same at the centre, the reduction is particularly important at the edges of the picture thus loosing fidelity in those regions. However, this process is considered acceptable because the assumption is made that, most of the time, the user looks straight ahead and rather turn his head instead of his eyes. Moreover, as explained in Clause 4.4.2.2, the peripheral vision is much less sensible to resolution than the central vision.

The second issue of applying a Barrel/Pincushion combination is that the image size is reduced. Edge areas of the original picture are lost, thus reducing the FOV. Such a limited FOV with visible black areas all around the picture is called a tunnel effect, which makes the immersion feeling to be lost as shown on Figure 4.6.

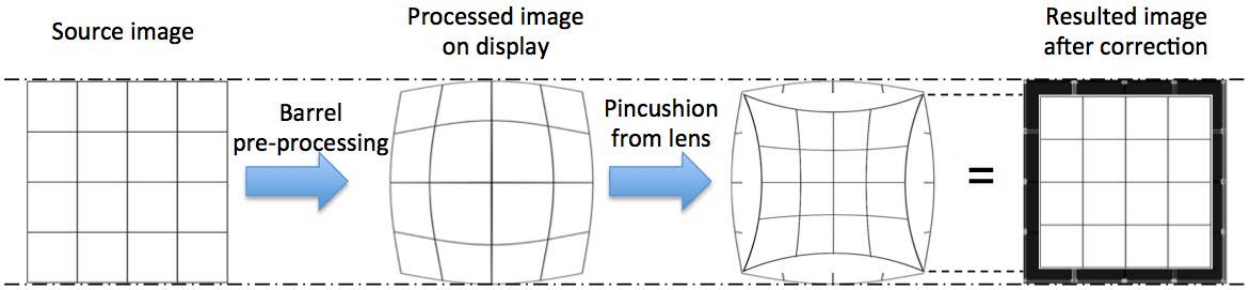


Figure 4.6: Consequence of the Barrel/Pincushion effect on image size

One solution for solving this size reduction is to use a higher resolution image at the source so as to get the desired resolution after correction. This would require video decoders and graphic-buffers capabilities to be higher than what the display can render.

4.2.4.3 Chromatic aberration

A common problem with the use of lenses occurs when not all the wavelengths of colours originated from the same location converge to the same focal plane as shown in Figure 4.7. This is called the chromatic aberration. It is a colour separation effect where the light of different wavelength refracts differently on the glass of a lens (also called colour fringing). This chromatic aberration can be corrected by separately resampling the colour channels of an image [4].

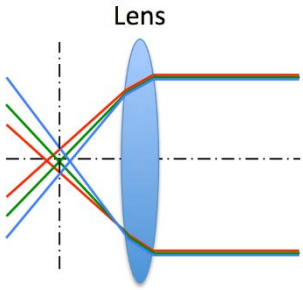


Figure 4.7: Illustration of the chromatic aberration

4.2.5 VR Video systems

4.2.5.1 Introduction

Figure 4.8 illustrates a typical functional workflow describing the functional steps for video over an end-to-end delivery chain. This workflow considers the DASH delivery over MBMS as an example and other approaches may differ while considering other types of service (e.g. conversational services) or other types of delivery mechanisms.

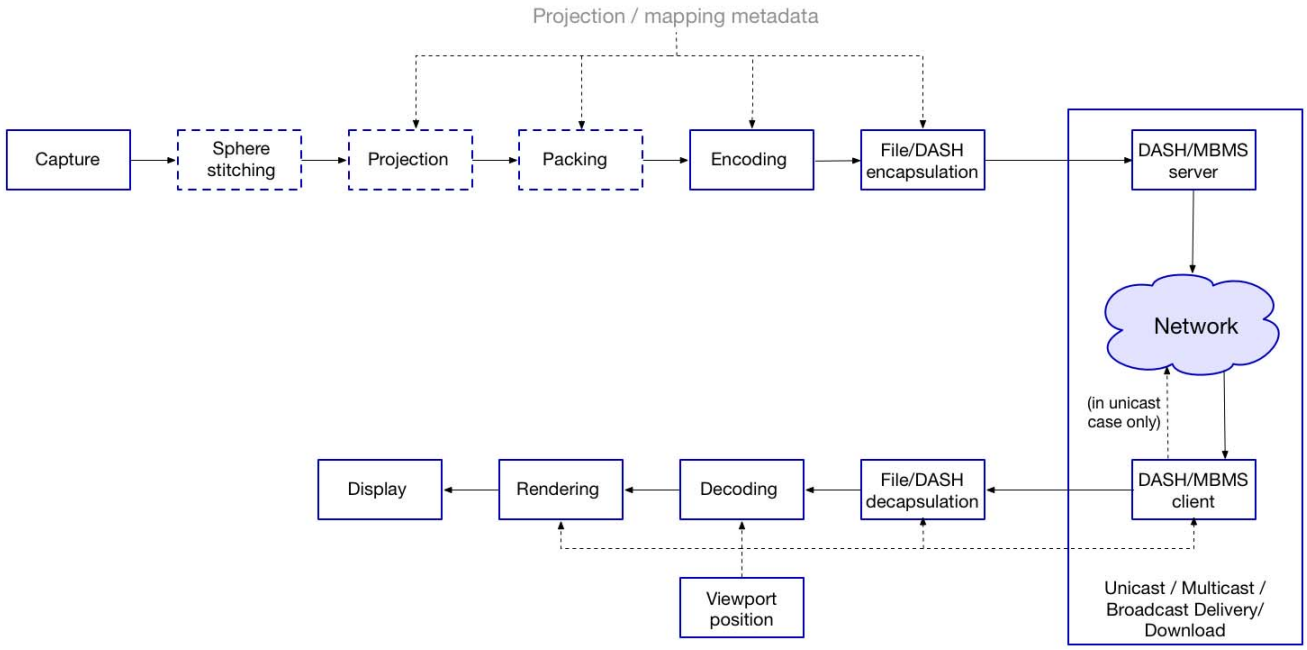


Figure 4.8: Overview of the VR video processing chain

NOTE: Clauses 4.2.5.7 and 4.2.5.8 do not cover conversational services. Conversational services will be treated separately.

4.2.5.2 Capture

VR content may be represented in different formats, e.g. panoramas or spheres depending on the capabilities of the capture systems. Many systems capture spherical videos covering the full 360°x180° sphere. Capturing of such content is typically done by multiple cameras. Various camera configurations can be used for recording 2D and 3D content.

For 2D content, cameras can be mounted on a ring to capture horizontal panoramas or mounted on a sphere to capture spherical (360°) video. Multi-camera arrangements for capturing panoramic videos require focal points of all camera views to coincide at a common point so that stitching can be performed with minimum parallax effect.

3D content can be captured by stereo camera pairs with a relatively small overlap arranged in a star configuration ("segmented stereo"). However, such camera systems may suffer from parallax errors. On the other hand, mirror-based systems can capture 3D images using camera pairs reducing parallax errors. Another option is "stereo by extreme overlap" (as shown on Figure 4.9) in which stereoscopic content is created from overlapping images captured by either fish-eye or wide-angle lenses, or by clusters of cameras. During processing, each image sensor is split into a left and right section and the corresponding left and right panoramas are stitched from these sections.

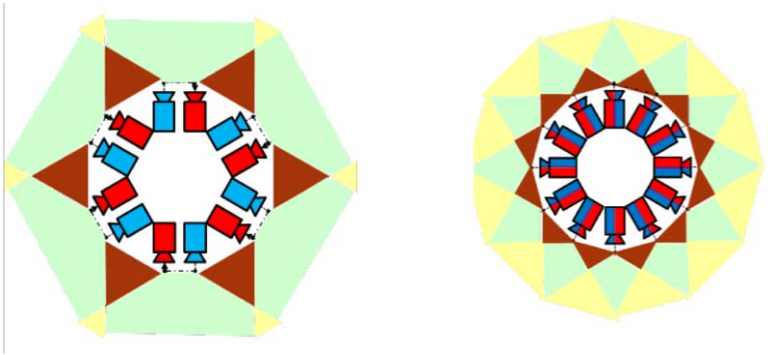


Figure 4.9: (Left) Segmented stereo (right) Stereo by extreme overlap

Light field rendering is another promising approach for creating 3D content [5]. However, light-field rendering requires dense camera grids. Hence, existing approaches use depth-based rendering to generate the intermediate camera views and reduce the required number of cameras (Figure 4.10). The disadvantage of this approach is that very accurate depth maps and sophisticated depth-based processing are needed which increases the required computational power and makes the approach error-prone.

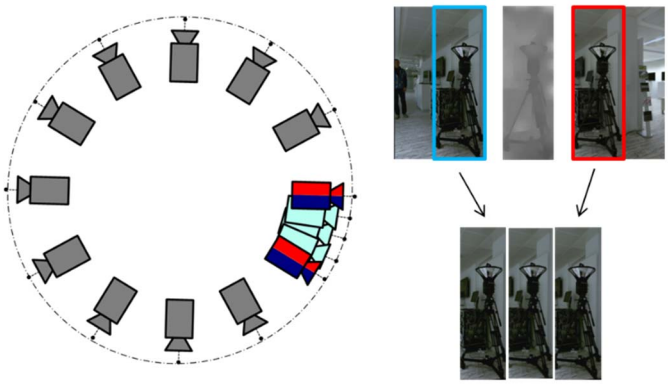


Figure 4.10: (Left) Depth-based light field processing (Right) Generation of intermediate views

4.2.5.3 Sphere stitching

Captured views from each camera are stitched together so as to combine the individual views of the omnidirectional camera systems into a single panorama or sphere. The stitching process should avoid parallax errors and visible transitions between the single views. Parallax errors occur because cameras do not have a common optical center when arranged in a star configuration. This results in blind regions (gaps) between the camera views. Also, in the overlap areas, objects recorded from different viewpoints appear on different positions in the single views. Blind regions and artefacts in overlap areas impede the stitching process as illustrated on Figure 4.11. Stitching captured camera views parallax-free is less complex for camera arrangements with a common focal-point, e.g. mirror-based systems. Such systems can reduce parallax down to objects within 1 m.

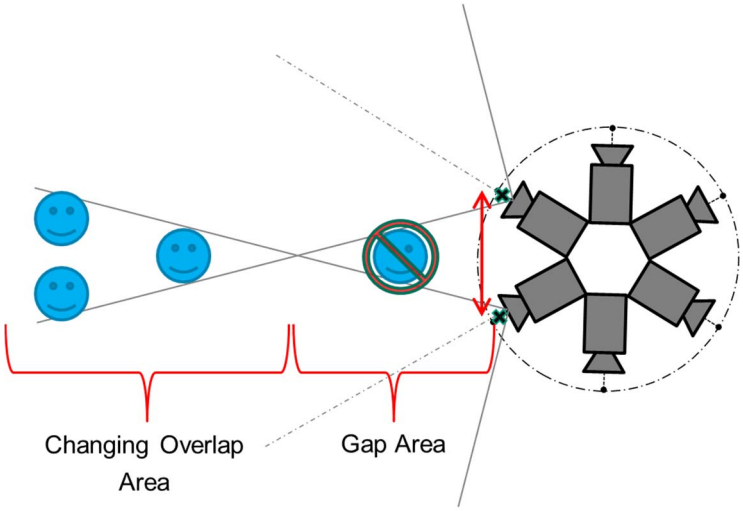


Figure 4.11: Illustration showing how parallax errors occur

Stitching can be done offline during the post-production stage or in real-time. For live transmission, a real time stitching process is required which should be able to process a large amount of data from multiple cameras and provide a high quality, error free panorama or sphere. Real time stitching is highly dependent on the omnidirectional camera system and is still a big challenge especially in terms of avoiding the parallax errors.

Taking into account the camera arrangements and stitching techniques, existing camera systems can roughly be classified into: mirror-based systems (direct stitching), systems with depth-aware stitching (segmented stereo and stereo

by extreme overlap), and systems with depth-enabled light field rendering. Table 4.1 gives a summary of the advantages and disadvantages of such camera systems, respectively:

Table 4.1: Comparison of existing VR capture/stitching systems

System	Pros	Cons
Direct stitching by mirror-based systems	- Parallax-free setup - Easy stitching - Almost no overlap - High resolution (>10k, 60fps) - Full lens control - Real-time 2D&3D processing - Capable of live transmission	- Bulky system - Calibration needed - Sensitive to damages
Depth-aware stitching (Segmented stereo, stereo by extreme overlap)	- Small form factor, light weight - Robust and compact systems - Easy handling - No calibration at setup - Existing stitching software tools - Established post-production	- Usually closed system, no lens control - Restricted real-time processing, limited use for live events - Reduced resolutions (due to overlap) - Anisotropic resolution of wide-angle and fisheye lenses causing sometimes extreme distortions in the resulting panoramas (for stereo by extreme overlap)
Depth-enabled light field rendering	- Parallax-free rendering - Enables producing novel views of the scene thus moving freely inside the scene	- Complex computing - Supervised post-production - Still an error-prone process - No real-time capabilities yet

4.2.5.4 Projection

Modern video coding standards are not designed to handle spherical content. Therefore, in VR systems, projection is used for conversion of a spherical (or 360°) video into a two-dimensional rectangular video before the encoding stage. A sphere can be projected onto a plane in various ways [6]. However, no projection method can be distortion free. The distortion caused by the conversion from spherical to planar domain is referred to as "sampling distortion". The final reconstruction quality of a spherical video is a function of both sampling and coding distortions.

The most commonly used projection method is the equirectangular projection (ERP), in which the horizontal and vertical coordinates simply correspond to longitude and latitude, respectively, with no transformation or scaling applied. However, equirectangular projected images have large redundancy near the poles because they are stretched in latitude direction. This causes a redundant number of bits to be spent to encode the poles of image (relative to the actual information content).

A projection that reduces the sampling distortion (compared to equirectangular) is the cubic projection in which a portion or whole of the sphere is projected to planar images. The images are arranged as the faces of a cube each of which has a 90°x90° FoV. Cube mapping is a sub-case of rectilinear (gnomonic) projections in which straight lines in real 3D space are mapped to straight lines in the projected space. Hence, each cube face retains straight lines. Both types of projection are illustrated in Figure 4.12.

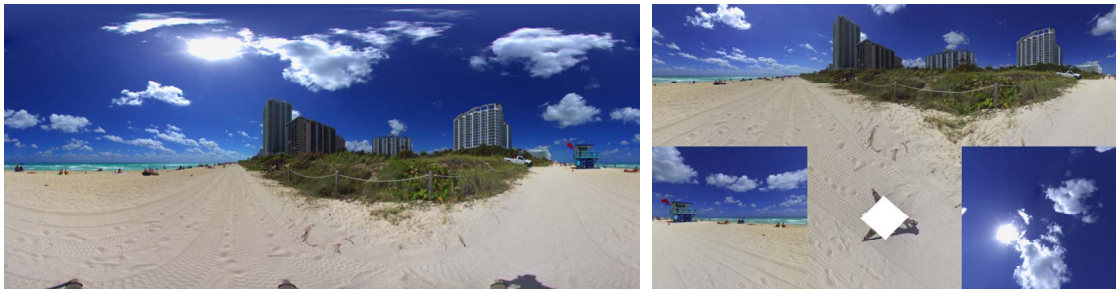


Figure 4.12: Equirectangular projection (left) vs Cubic mapping (right) of the same spherical content. Cube maps provide an approximately equal-area projection whereas equirectangular projection contains redundant samples towards poles of the image

Some other projection types (classified according to the type of geometry used for rendering) are listed below:

- Sphere.
- Squished Sphere.
- Cylinder.
- Platonic Solid:
  - Cube (6 surfaces).
  - Octahedron (8 surfaces).
  - Icosahedrons (20 surfaces).
- Truncated Pyramid.
- Segmented Sphere.
- Direct Fisheye.

4.2.5.5 Region-wise mapping (packing)

After projection, the obtained two-dimensional rectangular image can be partitioned into regions that can be rearranged to generate "packed" frames.

The operations to produce packed frames from projected frames (denoted as "packing" or "region-wise mapping") might include translation, scaling, rotation, padding, affine transform, etc. Reasons to perform region-wise mapping include increasing coding efficiency or viewport dependent stream arrangement (as detailed with the multi-stream approach in Clause 4.2.5.6.3 and illustrated on Figure 4.13).

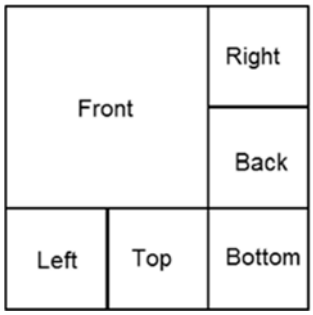


Figure 4.13: Example of a multi-resolution cube map. Some of the rectangular areas in the projected frame can be downsampled to construct the packed frame

If region-wise mapping is not used, the packed VR frame is identical to the projected frame.

4.2.5.6 Encoding & Decoding

4.2.5.6.1 Introduction

Current 360 video services offer a limited user experience since the resolution in the user viewport and hence the visual quality are not on par with traditional video services. Multiple times UHD resolution is needed to cover the full-360 surroundings in a visually sufficient resolution. This poses a major challenge to the established video processing chain and to the available end devices.

There are mainly three approaches that can be considered for 360 video delivery. All solutions can be grouped into:

- Single stream approach.
- Multi-stream approach.
- Tiled stream approach.

Additionally, a 360 video may contain one or more "highlight region(s)" described in the spatial domain. Such regions correspond to spatial areas which may be associated with a specific intent and/or process, e.g. serving specific artistic intent, defining default viewport modes or any practical purpose for optimizing the delivery of VR services. Such highlight regions are content properties and are independent from any 2D projection map used.

4.2.5.6.2 Single stream approach

For HMDs, one straightforward approach would be to encode an exact or over-provisioned viewport for each user, i.e. crop the interesting part (e.g. viewport) for the user at the server side and encode it. However, although this approach minimizes the number of non-viewport samples to be decoded, it comes at the cost of an encoding overhead when considered for large-scale deployments. Another option that is considered as a single stream approach is to encode the full 360 video, transmit it to the receiver and decode the full 360 video while showing only the viewport.

Therefore, solutions that lie within this group have the drawback that either they may not be scalable or they may impose a big challenge in terms of required network resources (high bitrate of high resolution video) and required processing at the client side (decode a very high resolution video).

Mobile devices typically contain hardware video decoders tailored to resolutions used in traditional video services (HD or UHD). Therefore, it is important to limit the overall resolution to be transmitted and decoded in the mobile devices.

Using single stream approach, the receiver decodes the entire video (corresponding to either the viewport (exact or over-provisioned) or the full 360 video).

4.2.5.6.3 Multi-stream approach

The multi-stream approach consists of encoding several streams, each of them emphasizing a given viewport and making them available for the receiver, so that the receiver decides which stream is delivered at each time instance. The number of the streams made available can vary and be optimized; with a larger number of streams, a better match to the users' viewport can be obtained. However, this requires more storage capacity at the server side. Even though multiple streams are encoded and made available, only a single stream needs to be decoded depending on the users' viewport.

There are two ways of generating viewport-dependent video bitstreams for the multi-stream approach:

- Projection/Mapping based: A viewport dependent projection (e.g. Truncated Pyramid) or a projection (e.g. cubic) plus a viewport dependent mapping/packing (e.g. multi-resolution cubemap shown on Figure 4.13) is used so that the number of samples is higher at the viewport and lower at surrounding areas. The encoding is done as usual, i.e. viewport unaware.

- Encoding based: The encoder is configured so that the samples of the viewport are encoded at a higher quality, e.g. with a lower quantization parameter (QP).

Using multi-stream approach, the receiver decodes the corresponding entire video which results in different resolution or different quality areas. Mobile devices typically contain hardware video decoders tailored to resolutions used in traditional video services (HD or UHD). Therefore, it is important to limit the overall resolution to be transmitted and decoded in the mobile devices.

4.2.5.6.4 Tiled stream approach

Another approach is to use HEVC tiles or separate video streams for 360 video delivery. It allows emphasizing the current user viewport through transmitting non-viewport samples with decreased resolution, i.e. selecting the tiles from the viewport at a high-resolution version and the tiles that do not belong to the viewport at a lower resolution [7]. Hence, the full 360° surroundings are always available on the end device but the number of samples that lie outside the user FoV is reduced. For this purpose, at the encoder side the full 360 video is projected into a frame, mapped (e.g. cube map) and encoded into several tiles at different resolutions.

Fallback schemes comprising low resolution tiles and other hybrid tiling approaches combining low and high resolution tiles may be considered.

When tiled stream approach is followed each of the tiles can be encoded as motion-constrained HEVC tiles or as separate video streams.

Using motion-constrained tile HEVC streams, the varying spatial resolution in video picture can be achieved by merging motion-constrained HEVC tiles of different resolutions into a single common bitstream and therefore this approach allows to use a single decoder.

Using separately encoded video streams, several decoders are required at the receiver side, as many as the number of the video streams the receiver chooses to decode.

Using tiled-streaming approach, the receiver can choose to decode only a subset of the received video stream depending on the current viewport position and/or device capabilities (e.g. video decoder capabilities).

4.2.5.7 File/DASH encapsulation/decapsulation

If DASH is used for the delivery of 360 video additional signalling may be necessary. For instance, projection and mapping formats might be required to be signalled at the MPD so that client can request the appropriate Representations and/or Adaptation Sets.

File/DASH encapsulation is then performed differently depending on the type of the considered solution (single-stream, multi-stream, tiled stream).

The receiver can choose to decapsulate only a subset of the received video stream depending on the current viewport position and/or device capabilities (e.g. video decoder capabilities).

4.2.5.8 Delivery

A panoramic or 360 video can be delivered in unicast, multicast or broadcast mode. In all of these modes, the delivery can be realized in the form of download or streaming and in real-time or non-real time.

For unicast streaming delivery, DASH can be used. For multicast or broadcast delivery, DASH over MBMS can be used.

For unicast delivery, all three approaches mentioned in Clause 4.2.5.6 (single stream, multi-stream, tiled stream) could be used.

For unicast and MBMS delivery, the DASH client requests appropriate segments depending on the viewport position, available network throughput, device capabilities and service requirements. E.g. for multi-stream approach, the DASH client requests the stream (representation) that matches best to the expected viewport position (subject to network latency and user movement).

4.2.5.9 Rendering

Rendering includes different steps. Typically, after a sequence of 2D images are decoded, one or more of the following steps are included:

- Sphere Mapping.
- Field-of-View generation.
- Region-wise unpacking.
- Generation of one view per eye for stereoscopic content.
- Restricted coverage rendering (e.g. if the content does not include full 360 video).
- Seamless switching across different field-of views and different resolutions of the sequence of 2D images.
- Other regular 2D operations such as bar data removal, tone-mapping, etc.

Considering the limited FOV of the existing VR displays, only a part of the video frame needs to be displayed to the user. For instance, in 2016, the FOV of the most currently available HMDs ranged between 100°-110°.

Since the actual rendered video will cover only a portion of the scene, the renderer benefits from receiving metadata related to the current user viewport. The renderer also uses the video characteristics such as projection and mapping formats.

4.3 Audio systems

4.3.1 Introduction

This clause describes Audio Systems for use with Virtual Reality. This clause provides a high-level overview of requirements for audio VR, Clause 4.3.2 contains an overview of audio capture systems for VR, Clause 4.3.3 describes Content Production Workflows for VR and Clause 4.3.4 describes audio rendering systems for VR.

In general, Virtual Reality requires Audio Systems that provide the consumer with the perception of being immersed in a virtual environment. However, immersion itself is not a sufficient condition for successful commercial deployment of virtual reality multimedia services. To be successful commercially, audio systems for VR should also provide content creation tools, workflows, distribution and rendering systems that are practical to use and economically viable to the content creator and consumer.

Whether a VR system is practical to use and economically viable for successful commercial deployment, depends on the use case and the level of sophistication expected for both the production and consumption of the use case. For example, uploading a short spherical video of a vacation trip uploaded to a social media website may not afford the same level of production complexity or immersive experience found in a blockbuster cinematic production. Similarly, users may have very different expectations. For example, when compared to the casual users, intensive gamers may have very different tolerance levels to both the quality and the degree of the immersion being delivered as well as the equipment necessary to achieve such an experience.

In addition to considerations of immersion, production, distribution and rendering complexity, VR audio use cases that support two-way communication have further constraints including, for example, a sufficient low mouth to ear latency requirements for conversational quality. As illustrated on Figure 4.14, careful consideration of the following aspects is therefore required:

1) the Audio Capture System.
2) the Content Production Workflow.
3) the Audio Production Format (see Clause 4.3.4).
4) the Audio Storage Format.
5) the Audio Distribution Format.
6) the Audio Rendering System (see Clause 4.3.5).

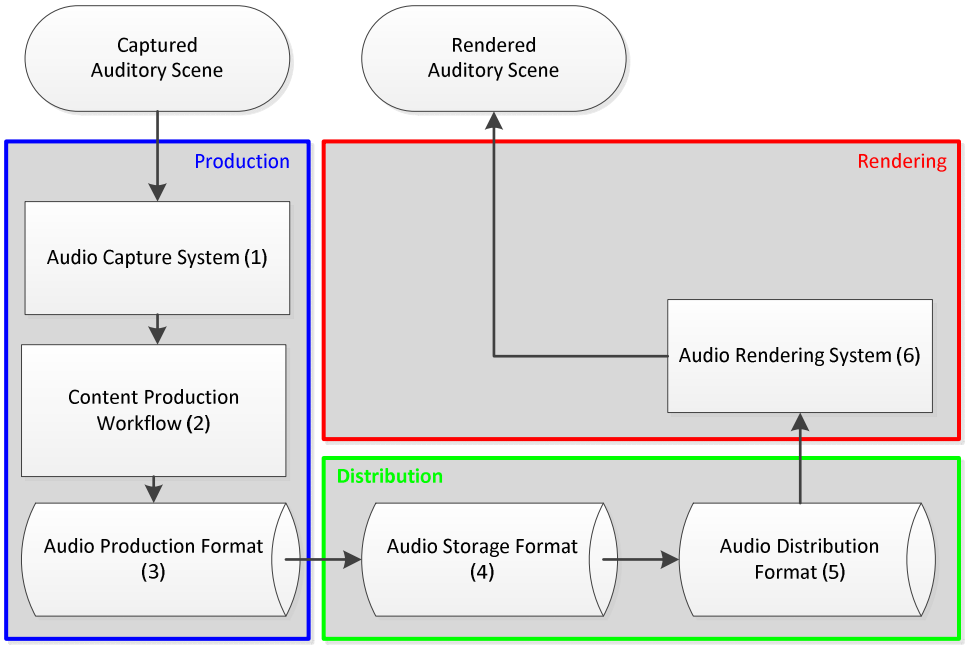


Figure 4.14: Audio System Components for VR

4.3.2 Audio capture system

4.3.2.1 Introduction

The audio capture system design for VR is dependent on the choice of the audio content production format, which itself hinges on the considerations described in Clause 4.3.1. For example, applications requiring the use of scene-based audio may support a single spherical microphone array to capture the auditory scene directly. Applications that make use of channel or object-based formats may choose to use one or more microphones optimized and arranged for the specific sound source being recorded.

4.3.2.2 Audio capture system for scene-based audio representation

Scene-based Audio is one of the three immersive audio representation systems defined, e.g. in Recommendation ITU-R BS.2266-2 [9]. Scene-based Audio is a scalable and loudspeaker-independent sound-field representation based upon a set of orthogonal basis functions such as spherical harmonics. Examples of Scene-Based audio formats commercially deployed for VR include B-Format First Order Ambisonics (FOA) and the more accurate Higher-Order Ambisonics (HOA).

Ambisonics is a periphonic audio system [8], i.e. in addition to the horizontal plane, it covers sound sources above and below the listener. An auditory scene for Ambisonics can be captured through the use of a first or higher-order Ambisonics microphone. Alternatively, separate monophonic sources can be captured and panned, or transformed, to a set of desired locations.

B-format Microphones

The "B-Format", or First Order Ambisonics (FOA) format uses the first four low-order spherical harmonics to represent a three-dimensional sound field using four signals:

W: the omnidirectional sound pressure.

X: the front/back sound pressure gradient at the capture position.

Y: the left/right sound pressure gradient at the capture position.

Z: the up/down sound pressure gradient at the capture position.

The four signals can be generated by processing the raw microphone signals of the so-called "Tetrahedral" microphone, which consists of four capsules, in a Left-Front-Up (LFU), Right Front Down (RFD), Left-Back-Down (LBD) and Right-Back-Up (RBU) configuration, as illustrated on Figure 4.15.

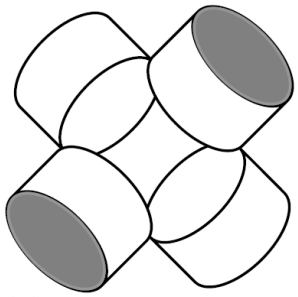


Figure 4.15: The "tetrahedral" microphone

Horizontal only B-format microphones

Other microphone array configurations can be deployed for portable spherical audio and video capture devices, with real time processing of the raw microphone signal components to derive the *W, X, Y* and *Z* components. Some configurations may support a *Horizontal-only B-format* in which only the W, X, and Y components are captured. In contrast with the 3D audio capability of FOA and HOA, a horizontal-only B-format foregoes the additional sense of immersion provided by the height information.

Higher-Order Ambisonics microphones

The spatial resolution and listening area of First Order Ambisonics can be greatly augmented by enhancing the number of directional audio components. These are *second*, *third*, *fourth* and Higher Order Ambisonics systems (collectively termed HOA). The number of signal components needed for a three-dimensional Ambisonics system of order *N* is given by $(N+1)^2$ and illustrated on Figure 4.16.

Figure 4.17 shows an example Higher Order Ambisonics capable microphone.

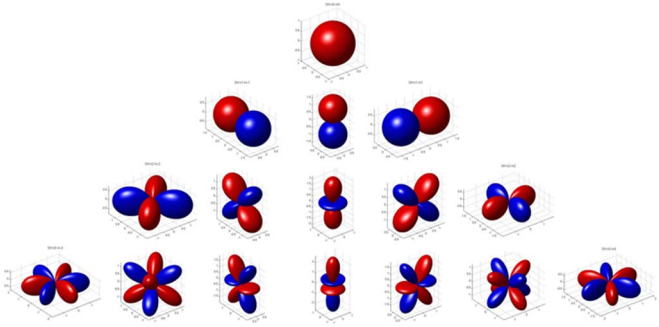


Figure 4.16: Spherical harmonics from order N=0 (top row) to order N=3 (bottom row)

Figure 4.17: A Higher Order Ambisonics capable microphone

Several formats exist for Higher Order Ambisonics data exchange. The component ordering, normalization and polarity should be properly defined and further details are provided in Clause 6.2.

General considerations for scene-based audio capture systems

Some general considerations of the audio capture system that affect the perception of immersion include:

- Signal to Noise ratio (SNR): Noise sources that are not part of the audio scene detract from the feeling of realism and immersion. Therefore, the audio capture system should have a low enough noise floor such that it is properly masked by the recorded content and not perceptible during reproduction.
- Acoustic Overload Point (AOP): Non-linear behaviour of the audio capture system may detract from the feeling of realism. The microphones should have a sufficiently high acoustic overload point to avoid saturation for the types of audio scenes and use cases of interest.
- Microphone frequency response: Microphones should have a frequency response that is generally flat along the audio frequency range.
- Wind Noise Protection: Wind noise may cause non-linear audio behaviour that detracts from the sense of realism.
- Microphone element spacing, crosstalk, gain- and directivity matching: These aspects ultimately enhance or detract of the spatial accuracy of the scene-based audio reproduction.
- Latency: If two-way communication is required, the mouth to ear latency should be low enough to allow a natural conversational experience.

4.3.2.3 Audio capture system for channel-based audio representation

Audio capturing using microphones and post-processing techniques for channel-based representation are well known in the industry, as they have been the standard for decades.

Multiple microphones are used to capture sounds from different directions; either coincident or spaced microphone arrays are used. Depending on the number and arrangement of microphones different channel-based formats are created, as e.g. stereo from XY mic pairs, or 5.1 by main microphone techniques or 8.0 by using microphone arrays. Alternatively, microphones built into VR cameras can be used to create channel-based audio representations.

Microphone post-processing allows different formats; products implementing post-processing of raw microphone recordings exist, e.g. a professional camera for VR with built-in microphones delivers 4.0, 5.0, 7.0 or 8.0 audio output channels. Consumer cameras provide e.g. 5.1 channel-based audio output.

For cinematic VR the microphone signals are modified and mixed further in post-production as described in the following Clauses.

4.3.2.4 Audio capture system for object-based audio representation

Object-based representations represent a complex auditory scene as a collection of single audio elements, each comprising an audio waveform and a set of associated parameters or metadata. The metadata embody the artistic intent by specifying the transformation of each of the audio elements to playback by the final reproduction system.

Sound objects generally use monophonic audio tracks that have been recorded or synthesized through a process of sound design. These sound elements can be further manipulated, e.g. in a Digital Audio Workstation (DAW), so as to be positioned in a horizontal plane around the listener, or in full three-dimensional space using positional metadata. An audio object can therefore be thought of as a "track" in a DAW.

The spatial accuracy of object-based audio elements is dependent on the metadata and the rendering system. It is not directly tied to the number of delivered channels.

Experiences provided by object-based audio typically go beyond that which is possible with single-point audio capture collocated with the camera resulting in audio which is more likely in order to satisfy the artistic intent of the producer.

4.3.3 Content production workflow

4.3.3.1 Introduction

Figure 4.18 depicts a basic workflow for content creation and delivery for VR today.

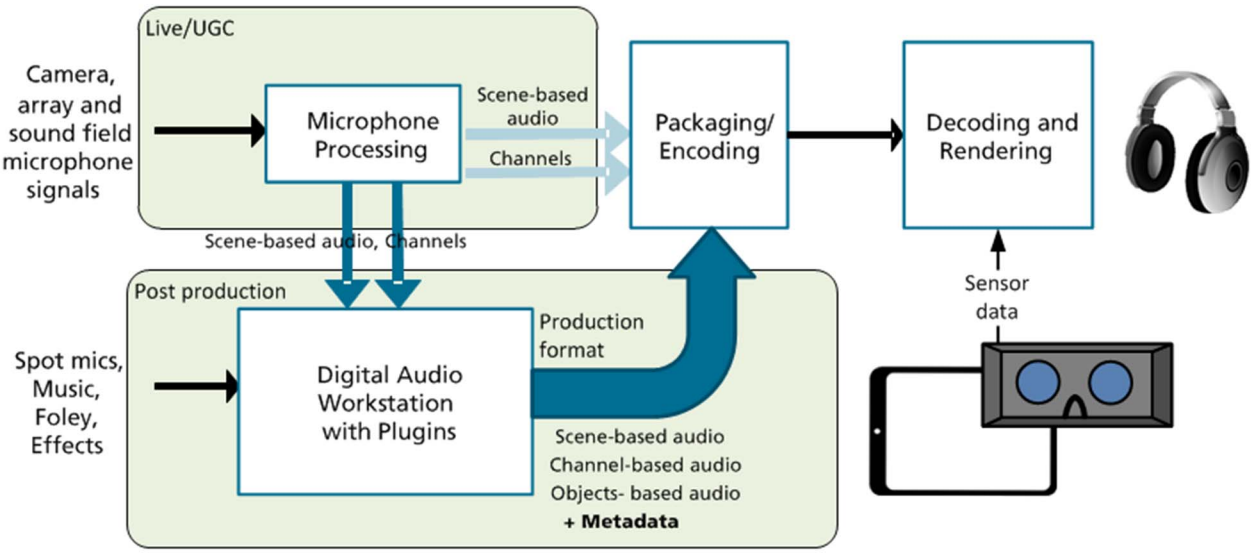


Figure 4.18: Event Multicast to VR enable user equipment

The workflow depicted in Figure 4.18 shows a complete VR audio system involving scene-based, object-based and channel-based audio representations. Not all of the components depicted need to be used for any particular use case.

4.3.3.2 Content production workflow for Scene-Based Audio

In applications involving real-time streaming of user generated content, the use of a live mixing engineer or post-production may not always be possible or desirable. In such applications, the content production workflow can be

simplified to one of direct transmission of a scene-based audio stream. The sound field is captured with an appropriate microphone array configuration and transmitted to the far-end in a format such as Ambisonics or Higher-Order Ambisonics.

Even though no user manipulation of the content is performed, it is still possible for the content transmitted to be modified as part of the transmission chain. In particular, Ambisonics and Higher-Order Ambisonics are built on a layered approach that makes the technology suitable for scalability. Scalability may be desirable depending on e.g. the bandwidth available, the computational resources on the renderer side, etc.

For example, the content could be scaled according to link budget conditions. An example of such an approach is depicted in Figure 4.19, where a Higher-Order Ambisonics content being transmitted adapts, based on the link conditions. When link conditions are adequate, higher spatial audio resolution can be transmitted, enhancing the sense of immersion and spatial localization. Conversely, if link conditions degrade, the network may to use lower spatial audio resolution, saving bit rate to improve link margin.

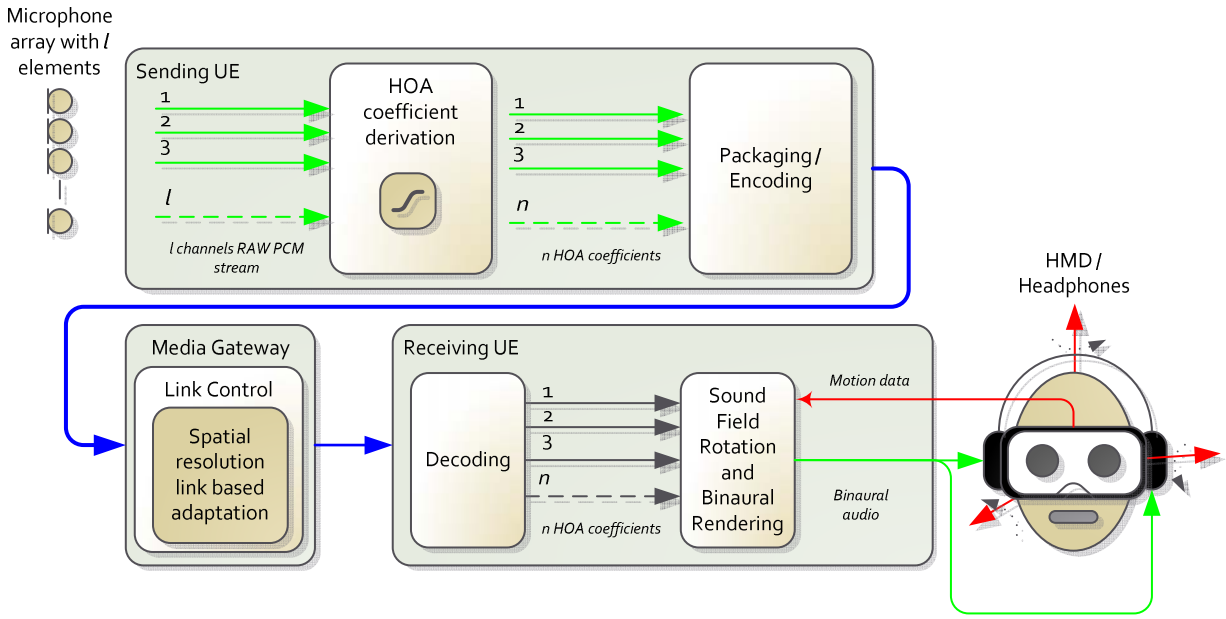


Figure 4.19: Example of a pure Scene-Based Audio workflow showing link adaptation

4.3.3.3 Channel-based content production workflow

In most cases the microphone signals, effects, foley and music are mixed in post-production, either in a live mixing console or a DAW as shown in Figure 4.18. Tools and workflows for channel-based audio production are well established in the industry. Most DAWs support immersive audio formats or can host plugins to support mixing for immersive output formats such as 9.1, 11.1, or 13.1.

In the context of VR, 5.1 surround represents a popular audio format enabling sounds from all directions on the horizontal plane to be reproduced. This format has been adopted by some existing VR platforms for mobile consumption. In order to provide truly immersive sound, a 3D format is required and there is a trade-off between the spatial resolution and the number of audio channels which can be made.

4.3.3.4 Production and post production for Object-Based audio representation

The process of production and post-production for linear VR experiences is similar to traditional cinematic content. The Figure 4.20 provides an example of a content creation process for live content. A set of audio elements obtained via spatial/sound-field recordings as well as spot microphones reach an audio mixing console or digital audio workstation where an audio engineer crafts an audio mix suitable for binaural reproduction over headphones. This creative process is paramount to deliver a high-quality, hyper-real experience. Hyper-realism is used to describe forms of compositional aesthetic where certain sounds that are present in the real environment are handled in a way so that they are either removed or somehow exaggerated.

An essential component of VR mixing for cinematic content as well as for live content is the positioning of the different audio elements of the mix in space (i.e. panning) so that they match the video reference. For the mentioned linear

content use-cases, this may be achieved through a user interface where the sound objects are positioned into a room-model in reference to a viewer. For VR, the mixing interfaces should be adapted to account for the fact that the video/visuals can encompass the entire sphere or even an entire 3D region around the nominal listening position.

For live produced content, generally the same principle applies. Different sound sources obtained via spatial/sound-field recordings, typically captured at different camera locations, as well as spot microphones reach a live audio mixing console. There an audio engineer assisted by automated components developed for live VR production crafts an audio mix suitable for binaural reproduction over headphones. The most obvious difference to the process for cinematic content is that the mix needs to be created in real-time.

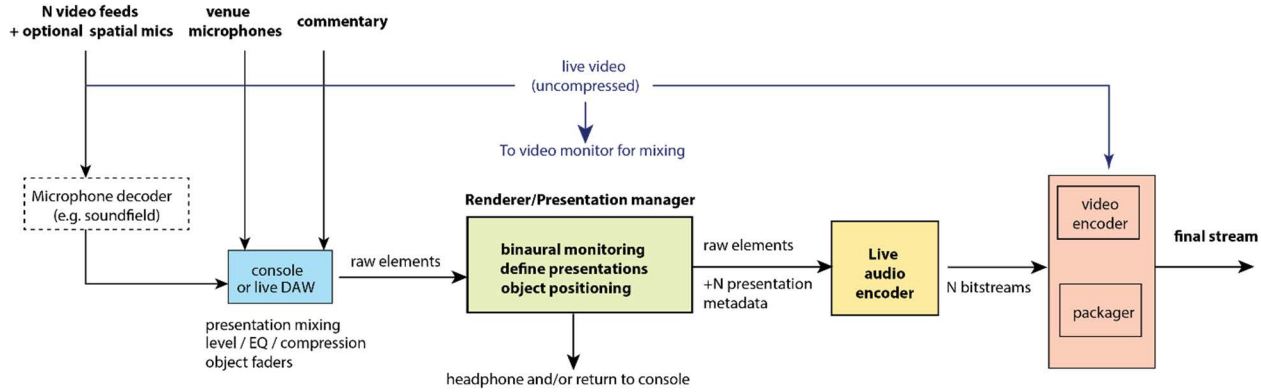


Figure 4.20: Example of a content creation process for live content

A key component in live workflows is the renderer/presentation manager that generates multiple mixes from a set of input audio elements, e.g. for the different camera viewpoints. This ability to output customized mixes for the different camera viewpoints is more critical for VR than for traditional broadcast use cases as a tight audio-video consistency is a requirement for immersion.

Object-based audio is well suited to creating the hyper-realistic mixes required by professional live VR applications, where the creation of a soundscape that matches the camera viewpoint often needs be compromised in order to craft an interesting and compelling audio mix. This may require enhancing far away elements that cannot be captured solely from the camera location but are nonetheless important in order to permit the user to follow the action. A solution is to author audio objects in 'world-space', i.e. the stadium or venue and let the presentation manager transform their position to match multiple viewpoints (e.g. for different cameras). In addition, an increasing number of systems are appearing which can be used to tie audio to real-time tracking systems in order to define dynamic audio objects.

4.3.3.5 Object metadata and controls for VR

Cinematic VR mixing often requires some audio elements to have specific playback-time behaviour. For instance, some non-diegetic background elements or music should preferably be kept 'head-referenced' i.e. non-head-tracked, while the diegetic sound effects or dialog should be 'scene-referenced' (i.e. head-tracked). Similarly, it may be desirable for some audio elements to be rendered with higher timbral fidelity by bypassing binaural processing at playback time.

Another category of controls for VR applications determines the environmental model. For spherical videos / 3DOF content, the environmental model is often captured and included as part of the audio elements themselves. This means that the reverberation, distance, source directivity and other room characteristics will be static, allowing the end-user to consume the content from a fixed position, but be unable to freely move within the given space.

Finally, a last category of VR specific controls relates to gaze-based interaction where the end-user can emphasize or even mute/unmute some of the elements in the mix by looking at specific points or directions.

These specific behaviours or properties can be easily authored and attached to the audio elements as object metadata.

4.3.4 VR audio production formats

4.3.4.1 Introduction

Recommendation ITU-R BS.2266-2 [9] presents a framework of future audio representation systems and the need for a production exchange file format. The framework recognizes channel, object and scene-based audio representations. Table 4.2 describes these different audio representations.

Table 4.2: Audio production formats

Signal type	Examples
Channel-based audio Mixes or mic array recordings for a specific loudspeaker layout e.g. Stereo, 5.1, 7.1+4	e.g. Full Mix, Music
Object-based audio Audio elements with positional metadata Rendered to target speaker layout or headphones	e.g. Dialogues, Helicopter
Scene-based audio B-Format (First-order Ambisonics) Higher-Order Ambisonics (HOA)	e.g. Crowd in Sports, Ambience

Note, that all signal types from Table 4.2 can describe 3-dimensional audio as necessary for an immersive VR experience. All signal types require audio metadata for control of the rendering e.g.:

- Channel configuration.
- Type of Scene-Based normalization scheme and Ambisonics coefficient ordering.
- Object configuration and properties, e.g. position in space.
- Diegesis, i.e. change upon head-tracking or steady with respect to the head - examples below:
  - Non-diegetic: A narrator who is not visible in the scene can be rendered independently from the head-tracking.
  - Diegetic: An actor's dialogue is rendered according to his position in the scene, taking the head-tracking information into account.

A range of file formats and metadata supporting VR video streaming and playback is in active use today. However, no commonly agreed mechanism to exchange content between these multiple VR file formats exists. This can make the user's production work frustrating as one may need to re-render the material for each platform where the video will be consumed. VR audio content, including object, channel and scene-based audio, is accompanied by metadata that allows the receiving party to properly interpret the audio data being transmitted. To facilitate production and distribution of content, it is desirable that the accompanying metadata can be interpreted by the different file formats used. This clause surveys the existing VR file formats and how essential metadata is handled.

4.3.4.2 Survey of existing spatial audio formats

VR streaming portal - Vendor 1

Vendor 1 describes an open metadata scheme to allow .mp4 containers to accommodate spatial (scene-based) and non-diegetic audio. The Spatial Audio Box (SA3D) contains information such as Ambisonics type, order, channel order and normalization. The Non-Diegetic Audio Box (SAND) is used to indicate audio that should remain unchanged by listener head rotation (e.g. commentary, stereo music, etc.).

At the time of this writing, the channel ordering is based on the Ambisonic Channel Number (ACN), and the normalization is Schmidt semi-normalization (SN3D).

VR streaming portal - Vendor 2

At the time of this writing, the service offered by vendor 2 service accepts videos in .mp4 or .mov containers. Audio is in AAC format using AAC-LC profile. The following formats are supported:

- 1 Channel, Mono.
- 2 Channel, Stereo (Left, Right).
- 4 Channel, Ambisonic (1st Order, ACN channel ordering, SN3D normalization).
- 6 Channel, 5.1 Surround (Left, Right, Center, LFE, Left Surround, Right Surround).

Additionally, vendor 2 supports "Binaural audio" and "Quadraphonic audio" through the use of four mono or stereo audio tracks. These audio tracks correspond to the four 90 degree cardinal directions matching the video (0deg, 90deg, 180deg, and 270deg). An open source audio and video conversion tool (FFmpeg) is suggested to build mp4 files with the "binaural" or "quadraphonic" audio formats.

VR streaming portal - Vendor 3

Vendor 3 also supports spatial audio using a proprietary 8 channel audio output format or first order Ambisonics with either ACN ordering and SN3D normalization (AmbiX) or Furse-Malham (FuMa) ordering. In addition, non-diegetic audio support can be enabled through content production tools provided by the vendor. Files are saved under an .mp4 container. Vendor 3 also supports playback of audio and video generated using the metadata scheme of Vendor 2.

Recommendation ITU-R BS.2076-0 [41] - Audio Definition Model (ADM)

The Audio Definition Model (ADM) is an open standard that seeks compatibility across object, channel and scene-based audio systems using XML representation. It aims to provide for a way to describe audio metadata such that each individual track within a file or stream is correctly rendered, processed or distributed.

The model is divided into a content part and a format part. The content part describes what is contained in the audio such as dialogue language and loudness. The format part contains technical information necessary for the audio to be decoded or rendered correctly, such as the position coordinate for a sound object and the order of an HOA component.

Recommendation ITU-R BS.2076-0 [41] provides for a series of ADM elements such as audioTrackFormat (describing what format the data is in), audioTrackUID (uniquely identifying a track or asset with a recording of an audio scene), audioPackFormat (grouping audio channels), etc. Guidelines for the use of IDs, coordinate systems and object-based parameter descriptions are also provided. Finally, a series of examples of ADM usage for channels, object and scene-based audio, including XML sample code and UML diagrams, concludes the recommendation.

Metadata that is specific to virtual reality, such as indication of whether a content is diegetic or non-diegetic, was not part of the recommendation at the time of writing.

Currently, ADM is more of a production format rather than a format that is conducive to streaming applications, but this may change in the future. For streaming, a container format that allows both audio and video packets (compressed and uncompressed) along with ADM metadata would be desirable.

ETSI TS 103 223 [42] - Multi-Dimensional Audio (MDA)

The Multi-Dimensional Audio (MDA) is a metadata model and bitstream representation of an object-based sound-field for linear content. The target applications for the standard are cinema and broadcast.

The metadata, which is part of the standard, can indicate when an object occurs on the program timeline and where it is positioned within the sound-field. The reference renderer is based on Vector Base Amplitude Panning [Clause 4.1] and renders objects to their desired sound locations. A bitstream representation is defined and mappings between common URI values and shorter Label values are provided to minimize overhead.

A broadcast extension (Clause 7) is defined in the specification, introducing support for Higher Order Ambisonics within the MDA stream. The channel numbering follows the ACN convention. The default normalization is N3D but other HOA normalization types are also supported (FuMa, SN3D, N2D, SN2D). Additional broadcast extensions are provided for Loudness and dynamic range compression management.

Metadata that is specific to virtual reality such as indication of whether a content is diegetic or non-diegetic was not part of the specification at the time of this writing.

AmbiX [43]

AmbiX supports HOA scene-based audio content. AmbiX files contain linear PCM data with word lengths of 16, 24, or 32 bit fixed point, or 32 bit float, at any sample rate valid for `.caf` (Apple's Core Audio Format).

AmbiX adopts ACN ordering and SN3D normalisation and can provide for full-sphere and mixed-order Ambisonics support. The ACN+SN3D convention for full-sphere Ambisonics is adopted by Vendors 1, 2 and 3 listed above. This convention is gaining rapid traction as a popular format for the exchange of Ambisonics content.

Additional metadata to indicate non-diegetic content is missing in this scheme.

ETSI TS 103 190-1 [44] - ETSI TS 103 190-2 [45]

The Digital Audio Compression system defined in ETSI TS 103 190-2 [45] also known as AC-4, is an audio codec which enables the delivery of immersive experiences for a wide range of use cases including broadcast, OTT and Virtual Reality. AC-4 allows the flexible use of objects, channels and scene-based audio. AC-4 does not contain a renderer; it provides audio for further rendering into different playback environments (e.g. binaural, headphones, speakers). For VR use cases AC-4 natively supports coding of audio to enable movement along 3 degrees of freedom, and as well 6 degrees of freedom when object audio is used.

ISO/IEC 23008-3 MPEG-H 3D Audio [46], [47]

The ISO/IEC 23008-3 MPEG-H 3D Audio System provides the possibility to carry audio channels, audio objects and scene-based-audio signals (FOA/HOA) and metadata inside a single audio bitstream. This capability allows content producers to choose a combination of audio elements depending on their creative intent and target distribution platform. For VR-specific content creation, audio elements can be authored as being diegetic and non-diegetic. To provide an immersive audio experience on mobile and tethered VR platforms, the MPEG-H decoder includes loudness management and 3DOF rendering capabilities for loudspeakers and binaural headphones. MPEG-H 3D Audio specifies a normative interface for the orientation of the user's head with respect to the ground, as Pitch, Yaw, Roll, and 3D Audio technology permits rendering of the audio scene to any user orientation. State-of-the-art coding tools enable high-quality immersive audio on low complexity decoders (MPEG-H Low Complexity profile) and at bitrates that have been traditionally used for 5.1 surround sound.

The MPEG-H verification test report [36] provides details on four listening tests that were conducted to assess the performance of the Low Complexity (LC) Profile of MPEG-H 3D Audio. These tests were conducted in the context of Broadcast.

MPEG OMAF 3D Audio Baseline profile (OMAF) is considering MPEG-H 3D Audio LC Profile. This media profile is being specified in MPEG-I OMAF [ISO/IEC 23090-2: Omnidirectional Media File Format] and is the only MPEG audio media profile that fulfils all the requirements for MPEG-I Phase 1a [http://mpeg.chiariglione.org/standards/mpeg-i/application-format-omnidirectional-media/requirements-omnidirectional-media-format] for omnidirectional media.

4.3.4.3 Observations

For scene-based audio, existing commercial solutions seem to be converging towards the use of ACN ordering and SN3D normalization (with a trivial conversion between N3D and SN3D). However, such metadata has not yet been standardized - existing just as extensions to the .mp4 container.

For the simultaneous handling of objects, channels and scene-based audio, e.g. Recommendation ITU-R BS.2076-0 ADM [41] provides a platform to exchange broadcast content that could possibly be expanded to handle aspects specific to VR (e.g. support for mixed diegetic/non-diegetic content, possibly even including a reference renderer, streaming, etc.).

To reduce fragmentation in the industry, a common production, metadata stamping and rendering workflow would be of interest for VR audio (similar to other broadcast efforts undertaken in ITU-R and elsewhere). Figure 4.21 displays an example processing chain where universal VR audio metadata can be added to the content during production and retrieved at the rendering side, regardless of the presence of possible spatial audio compression.

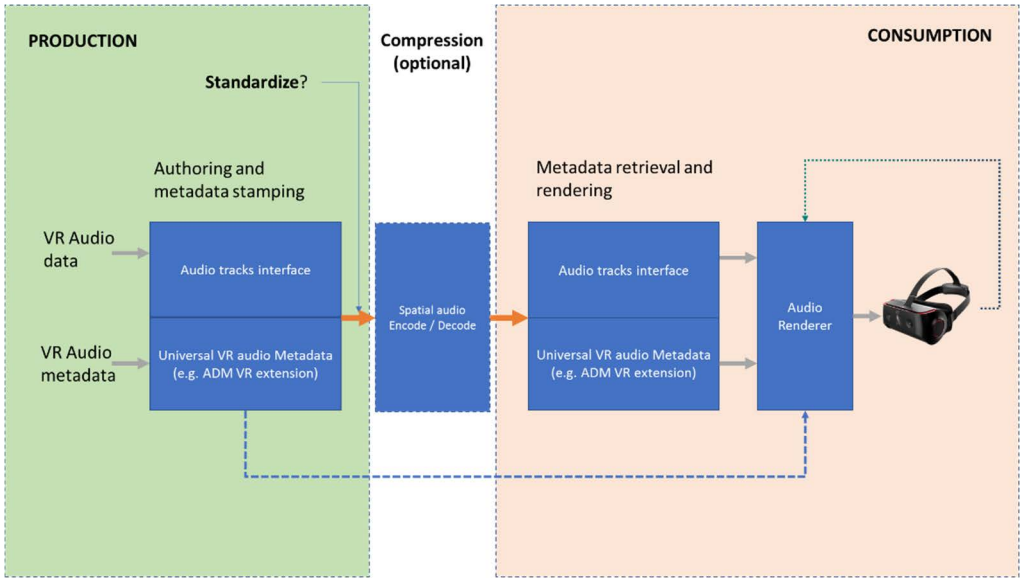


Figure 4.21: Example processing chain with universal VR audio metadata

4.3.5 Audio rendering system

4.3.5.1 Audio rendering system for Scene-Based Audio

In Scene-Based Audio (SBA), the rendering system is independent of the sound scene capture or creation. The sound scene rendering is typically performed on the receiving device and can produce real or virtual loudspeaker signals. The vector of loudspeaker signals $S=[S_1 \ldots S_n]^T$ can be created by:

$$S=D.B$$

Where B is the vector of SBA signals $B=[B_{(0,0)} \ldots B_{(n,m)}]^T$ and D is the rendering matrix for the target loudspeaker system.

More typically, for VR applications, the presentation of the audio scene is done binaurally over headphones. The binaural signals can be obtained through a convolution of the virtual loudspeaker signals S and a matrix of binaural impulse responses of the loudspeaker positions IR_{BIN}:

$$S_{BIN}=(D.B)*IR_{BIN}$$

In VR applications, it is desired to rotate the sound-field relative to head-movement. Such a rotation can be applied through a multiplication of a matrix F to the SBA signals:

$$B' = F.B$$

$$F = D.L$$

Where the matrix F is produced by multiplying the rendering matrix D with a matrix L. The matrix L consists of spherical harmonics of the loudspeaker points rotated by the amount of the head movement.

For efficiency, the binauralized rendering matrix $D_{IR} =(D.F)*IR_{BIN}$ can be computed ahead of real time processing.

4.3.5.2 Audio Rendering system for Channel-Based Audio

Channel-based formats are most well known in conventional audio productions. Each channel is associated with a corresponding loudspeaker. Loudspeaker positions are standardized in e.g. ITU-R BS.2051 [10] or MPEG CICP. In the context of VR, each loudspeaker channel is rendered for headphones as a virtual sound source; i.e. the audio signal of each channel is rendered from the correct location in a virtual listening room. The most straightforward way of doing this is to filter the audio signal of each virtual source with a measured response from a reference listening room. The acoustic responses can be measured using microphones in the ears of a human or artificial head. They are referred to as binaural room impulse responses (BRIR).

Such an approach is considered to give high audio quality and accurate localization. The downside of this approach is the high computational complexity, especially for a higher number of channels to be rendered and long BRIRs. Therefore, alternative methods have been developed to lower the complexity, while maintaining audio quality. Typically, these alternative methods involve parametric models for the BRIR, e.g. by applying sparse filters or recursive filters [11], [12].

4.3.5.3 Audio Rendering System for Object-Based Audio

In object-based audio rendering, sound sources are presented individually with metadata that describes the spatial properties of each sound source, such as position, orientation, width, etc. Sound sources are rendered separately in the 3D audio space around a listener, making use of these properties.

The rendering is performed either onto a specific loudspeaker configuration or onto a headset. Loudspeaker rendering uses different types of loudspeaker panning methods, while rendering to a headset can be performed in one of two different ways. The first headset rendering method is direct binaural rendering of the individual sound sources using HRTF filtering. The second is by indirect rendering that first uses panning methods to render the sources onto a virtual loudspeaker configuration and then by performing a binaural rendering of the individual virtual loudspeakers.

For rendering to headphones, the direct binaural rendering of individual sources usually provides better quality as it will, when properly implemented, provide more accurate positioning and more distinctive perception of the sound sources.

4.3.6 Audio rendering system

Independent of the format, audio signals need to be rendered to form the appropriate audio output signal for headphone playback. This rendering needs to take into account the sensor data reflecting the user orientation. The rendering process typically involves binaural rendering technology to render virtual speakers, scene-based audio or sound objects in the auditory space around the listener.

For mobile devices, computational complexity is an important factor in the characterization of a rendering system. Another highly-relevant performance metric is motion-to-sound event latency, i.e. the time lag between the motion of the users' head and the update of the sound position arriving at the user's ears. The threshold of noticeability for the head-tracking latency reported in literature is 60-85 ms [13]. Other sources claim a mean latency detection threshold at ~108 ms with a standard deviation of 30 ms [14]. Those results are limited to the conditions of the study and may not be directly related to all aspects of the user experience (e.g. motion sickness). The applicability of these results needs to be verified in the context of VR.

4.3.7 Audio quality evaluation of a VR delivery system

4.3.7.1 Audio signal flow

In Clause 5, several use cases for VR are outlined. Independent of the particular case a general structure for the signal flow can be drawn as the one on Figure 4.22.

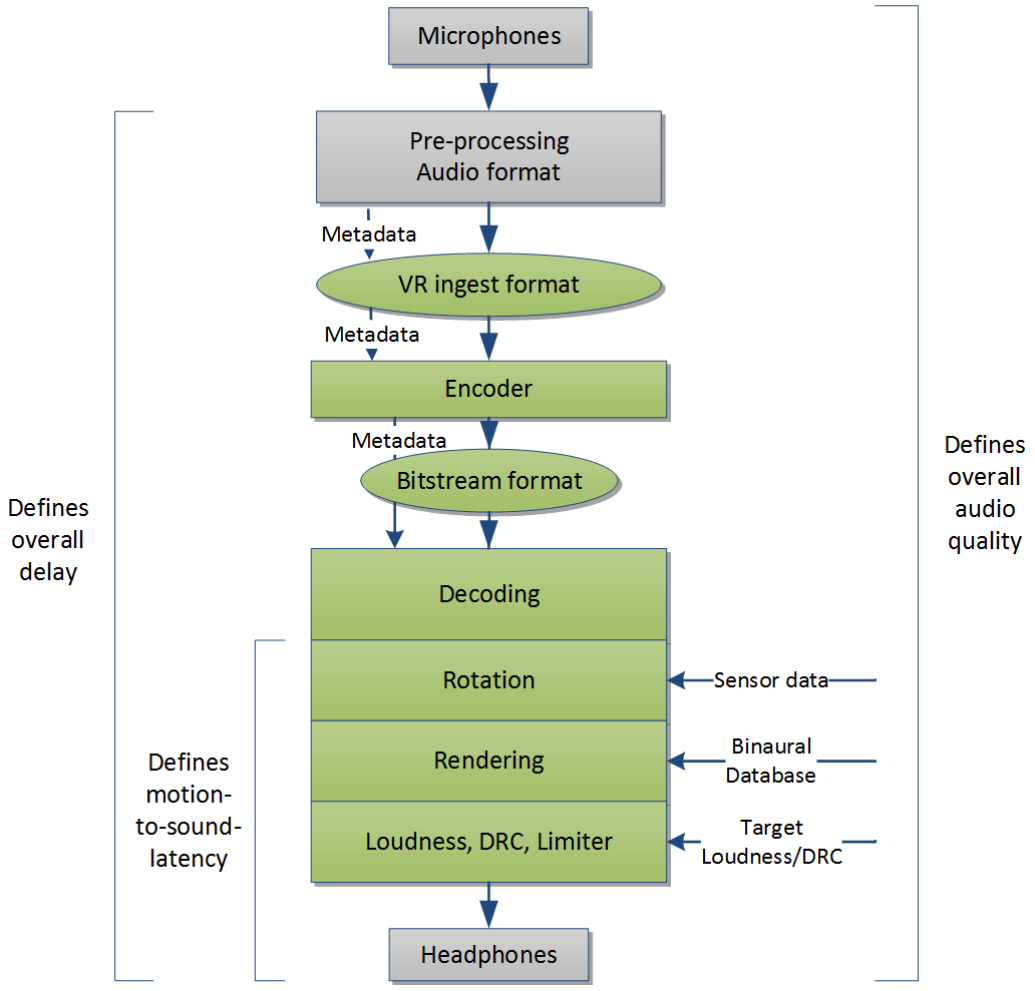


Figure 4.22: Audio processing steps in a one-way VR end-to-end scenario.

4.3.7.2 Factors influencing the QoE

Multiple factors influence the Quality of Experience [15].

In a spatial audio reproduction system, the following quality features contributing to the QoE were identified in [16], as shown in Figure 4.23:

- Localization accuracy (azimuth, elevation, distance)
- Timbre or coloration
- Auditory spaciousness (width of an auditory event)
- Artefacts (such as noise, clicks, gaps, parts not belonging to the content)
- Reverberance or envelopment + engulfment (perception of the room)
- Dynamic accuracy (for moving objects)

Quality Features	1,5	1,7	2,0	2,2	2,4	2,6	average
Localisation accuracy	1,4	3,2	1,8	3,2	3,0	3,5	2,7
Timbre	1,8	0,4	2,3	4,0	2,8	2,8	2,3
Auditory spaciousness	1,6	2,0	1,8	2,3	3,0	3,3	2,3
Artefacts	1,2	2,4	3,5	1,8	0,8	2,7	2,0
Reverberance	2,6	0,8	1,5	0,8	2,5	2,0	1,7
Dynamic accuracy	0,5	1,4	1,5	1,3	2,3	1,7	1,4
Quality Elements	Content	Head tracking	Audio processing, coding	HRTFs	Recording technique, mic setup	Binaural Rendering	

Figure 4.23: Correlation between quality features and technical parameters. no relation = 0, very important relation = 4 [16]

In addition, for an interactive and communication application, the following parameters are important:

- Audio latency (correlated to mouth-to-ear delay)
- Spatial sluggishness/blurriness (correlated to motion-to-sound latency)

4.3.7.3 Recommended objectives for quality assessment and assurance

From a service perspective, the goal should be that the overall QoE is maximized within the given constraints and that a minimum QoE is ensured. These quality features listed here are determined by the all coloured blocks in Figure 4.22. Therefore, the following is needed:

1) A technical evaluation of the subjective quality of the relevant parts of end-to-end system needs to be performed to assess the overall QoE.
2) A full technical description and either an implementation of these blocks and/or performance requirements are required to ensure the tested QoE.

4.3.8 Data exchange for audio formats

4.3.8.1 Introduction

VR audio systems require the delivery of a feeling of immersion in the virtual environment. Audio formats supporting VR need to render a spatial audio impression that corresponds to the virtual environment where the listener is immersed. Possible audio formats to deliver spatial audio include channel-based audio (CBA), object-based audio (OBA) and scene-based audio formats (SBA). A combination of those formats (e.g. scene-based audio augmented with objects) may also be of interest for certain use cases.

4.3.8.2 Data exchange for Scene-Based Audio formats

For a successful transmission and reception of Scene-Based audio content, the sending and receiving parties need to be aware of the format being used for the data exchange. In case of Ambisonics and Higher-Order Ambisonics, different possibilities exist for choices of Spherical Harmonics Component Ordering and Normalization.

Spherical Component Ordering Conventions

- Furse-Malham higher-order format (FuMa) [31]
- Single Index Designation (SID) [40]
- Ambisonic Channel Number (ACN) (used in AmbiX) [50]

Spherical Component Normalization Conventions

- maxN [31]

- SN3D (used in AmbiX) [43]
- N3D

4.3.9 Analysis of Ambisonics

Ambisonics is being deployed in commercial, over-the-top VR streaming services. For many of the available commercial services, First Order Ambisonics (FOA) using 4 channels of audio is employed, but also, Higher Order Ambisonics adoption is increasing. Ambisonics is a relevant VR audio format for consideration by 3GPP.

There is the question whether FOA would be a suitable VR audio representation for 3GPP VR applications. Four listening tests were conducted as part of the study (see Clause 6.1.3, 6.1.5 - 6.1.7). The tests were performed with specific choices of renderer and different results might be seen for other renderers or reproduction environments.

The two studies presented in Clause 6.1.3 and 6.1.5 assessed the localization quality.

The test results in Clause 6.1.3 show that significant localization quality can be gained when going from FOA to HOA, with the choice of renderer used in that test. Those tests used binaural rendering with generic HRTFs which may limit the ability to correctly localize and externalize sound sources for certain subjects. The same also applies for the reference conditions of the listening tests.

In Clause 6.1.5 the overall quality and spatial localization (direction and localization blur) for 1st, 2nd, and 3rd order Ambisonics by rendering using a circular array of 16 loudspeakers in two levels was evaluated. Anchors in form of 0th order Ambisonics (rendered from a mono channel), stereo and first-order Ambisonics with attenuated gradient signals were used. A reference was provided in form of a written description of the spatial sound scene. The results indicate an increased overall and localization quality going from first-order Ambisonics to higher orders, but also statistically significant improvements of first-order Ambisonics over the mono and stereo conditions, i.e. over what is achievable using existing 3GPP speech and audio services. Higher order Ambisonics (e.g. 4th or 5th order) were not assessed in this test.

Two other studies, presented in Clauses 6.1.6 and 6.1.7, assess coding of FOA B-Format signal representations with 3GPP eAAC+ and 3GPP EVS codecs, respectively. These are comparative studies assessing the quality of coded FOA representations compared to uncoded FOA.

The study presented in Clause 6.1.6 suggests that two 3GPP eAAC+ stereo streams can appropriately carry B-format audio that was derived from 7.1.4 channel audio, after a conversion to A-format. Renderered to a 7.1.4 loudspeaker array as well as binauralized to headphones with generic HRTFs and without head-tracking MUSHRA listening test results indicate that the “Good” to “Excellent” quality range is achievable at bitrates from 64kbps (2x32kbps) to 192kbps (2x96kbps). In that range a quality increase is observed that is commensurate with increased bitrate. While with loudspeaker rendering the quality assessement of FOA may be affected by the loudspeaker configuration, the evaluation with binaural rendering using generic HRTFs may be limited regarding the ability to correctly localize and externalize sound sources for certain subjects. The same also applies for the reference conditions of the listening tests.

Clause 6.1.7 shows that the 3GPP EVS codec can be used to encode super-wideband FOA B-format representations obtaining MUSHRA scores in the “Good”/”Excellent” quality range with increasing quality from 4x24.4 kbit/s to 4x96 kbit/s compared to uncoded FOA.

4.4 Example service architectures

4.4.1 Introduction

This clause introduces different service architectures that permit the distribution of VR services. Different architectures are introduced in order to identify commonalities and differences of different service scenarios. The architectures the interfaces presented do not imply that they create any normative functions or interfaces for 3GPP standardization, but are expected to be used in the implementation and gap analysis discussion for the use cases.

4.4.2 Streaming architecture

The architecture introduced in this clause addresses service scenarios for the distribution of VR content in file or segment based download and streaming services, including PSS HTTP-based unicast (as for example defined in 3GPP TS 26.247 [39]) and MBMS Download Delivery including DASH-based services.

Figure 4.24 considers a functional architecture for such scenarios. VR Content is acquired and the content is uploaded for preparation using interfaces B, split in audio B_a and video/image in B_v. Along with the content preparation, metadata is available that may be used in the encoding and in the file format encapsulation. After media encoding, the content is made available to file format encapsulation engine as elementary streams E and the file format may generate a complete file for delivery or segmented content in individual tracks for DASH delivery over interface F. Content may be made available in different viewpoints, so the same content may be encoded in multiple versions.

At the receiving end, there is an expectation for the availability of a VR application that communicates with the different functional blocks in the receivers' VR service platform, namely, the delivery client, the file format encapsulation, the media decoding, the rendering environment and the viewport sensors. The reverse operations are performed. The communication is expected to be dynamic, especially taking into account the dynamics of sensor metadata in the different stages of the receiver. The delivery client communicates with the file format engine, and different media receivers decode the information and provide also information to the rendering.

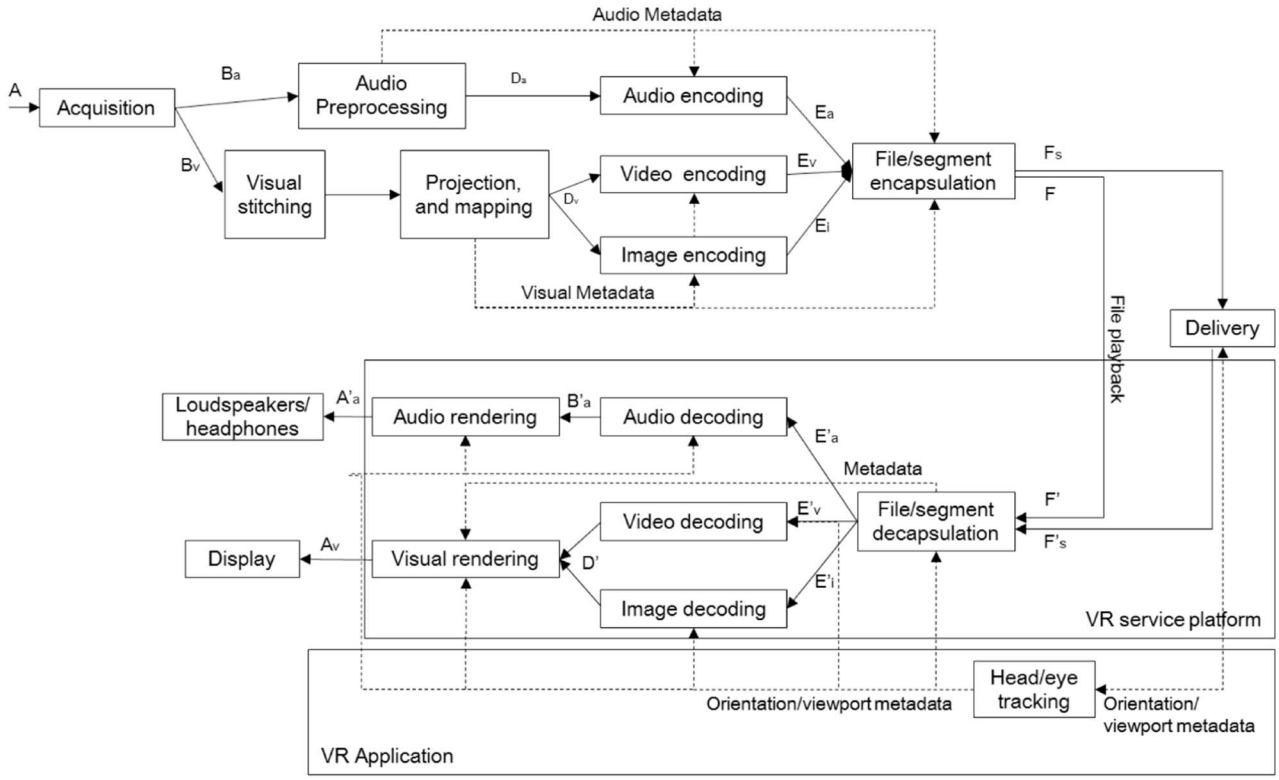


Figure 4.24: Example architecture for VR streaming services

5 Use cases for Virtual Reality

5.1 General overview

5.1.1 Introduction

Considering the VR application space it can be divided by the freedom or opportunity with which users can interact with the "*non-physical world"*:

1) A single observation point (or a series of single points under producer/user control)

2) Continuously varying observation point under user control

5.1.2 Single observation point

As a minimum, VR users should be able to look around from a single observation point in 3D space defined by either:

a) a producer (in the case of movie/entertainment content); or

b) the position of a capturing device(s) (in the case of live content).

This ability to look around and listen from a point in 3D space is usually described in terms of the 3 degrees of freedom; pitch, roll and yaw but it is worth noting that this point in space is not necessarily static - it may be moving. Users or producers may also select from a few different observational points but each observation point in 3D space only permits the user the 3 degrees of freedom.

360 degree video content captured in this way is widely available and when combined with simultaneously captured audio, binaurally rendered with an appropriate Binaural Room Impulse Response (BRIR), many and varied interesting VR applications can be enabled.

Most currently available movie content also features 360 degree view video with the audio component making use of multi-channel audio combined with binaural rendering.

5.1.3 Continuously variable observation point

This represents the ultimate goal for VR Systems where users are able to look around from observation points in 3D space and to move within that space under their own control with *apparently* full freedom. To achieve this apparent full freedom of movement it is necessary to add translational movement at least in a horizontal plane to the 3 degrees of freedom; pitch, roll and yaw above.

Such translational movement through the virtual space is very likely to be, but does not necessarily need to be, constrained to be consistent with conventional movement (walking in any direction in a horizontal plane, bending down etc.) - although for it to appear convincing to the user such consistency may be desirable. As a consequence, it is likely that not all of any 3D space is going to be reachable as an observation point by the user, as would be the case in the real world i.e. between ~0,15 m and ~1,8 m from the floor on which the user is standing.

Technology exists that can provide a basis for the video component of a truly variable observation point VR but, although audio component solutions exist, a harmonized combination of the two which is capable of realistically matching the audio to that image is still a fruitful research topic [17]. For example, the complexities of matching the Doppler of the sources with any movement of the observer and adaptively varying the BRIR acoustics to remain consistent with the environment around the user and those of the path between the user and the source, for each and every source, including shadowing when mobile objects in the scene come between source and listener seem far off. Based upon the above, it therefore seems likely that in the near-term such full movement will be constrained and under the control of a content producer who is able to take care of these acoustic effects.

5.1.4 Background to the use cases

In surveying the available technologies for VR in compiling the present document, it seems most appropriate to concentrate on those applications and use cases which allowing VR users to be able to look around [and listen] from a single observation point, or set of observation points, in 3D space.

5.1.5 Simultaneous diegetic and non-diegetic rendering

In extension of the below use cases, some media elements of a scene (certain sounds or parts of the video) are rendered as non-diegetic i.e. independently from head-tracking, and other elements of the scene are rendered as diegetic i.e. based on head-tracking.

5.2 Event broadcast/multicast use cases

5.2.1 Introduction

Combining delivery of spherical video, head tracking, and 3D audio capability, VR delivers immersive video consumption experiences for a variety of events. Events may take the form of sport events, concerts, game shows, TV shows, etc. Broadcast / Multicast of events is a relevant use case for VR, enabling new content delivery business models for the industry.

5.2.2 "Infinite seating" content delivery via multicast

At the event capture side, omnidirectional camera and microphones rigs can be placed at certain seating/viewing locations. Each audio and video capture rig delivers a distinct experience, corresponding to unique seating locations at the event side.

In "pay-per-view" applications, the content provider may market these locations at different price tiers or the viewer may be able to switch between them.

The "infinite seating" audio experience can be further augmented with additional audio sources (e.g. commentator microphones, ball-kick microphones, additional sources from the crowd and field). The "infinite seating" video experience can be further augmented with techniques such as action replay, cuts from different vantage points, etc.

The experience/sound of the scene adapts continuously as the viewer changes position, providing an optimized experience of the audio/video scene, according to the individual viewer's actual position in the environment.

Figure 5.1 illustrates such a use case.

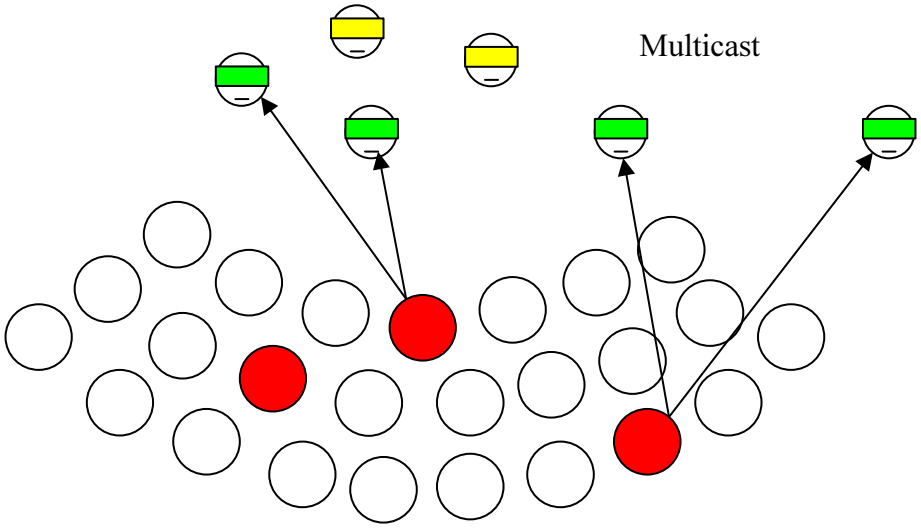


Figure 5.1: "Infinite Seating" multicast to VR enabled user equipment

5.2.3 Event multicast to VR receivers

The content provider may deliver a version of the event as an omnidirectional VR program to its viewers for consumption via VR equipment. The video and audio of the VR program is delivered via multicast to many VR viewers simultaneously as depicted on Figure 5.2.

The experience/sound of the scene adapts continuously as the viewer moves, providing an optimized experience of the audio/video scene, according to the individual viewer's actual position in the environment.

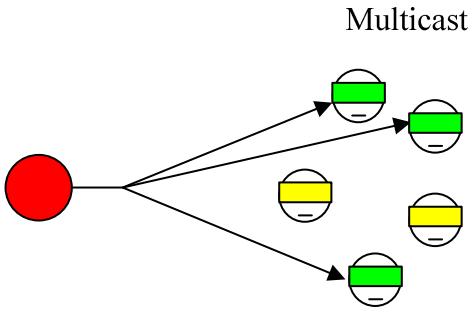


Figure 5.2: Event Multicast to VR enable user equipment

5.3 VR streaming

5.3.1 Use case

A user watches a VR on-demand video content with a HMD. The content is streamed over the 3GPP network using unicast. The user can navigate within the 360 degree video by moving his head and watch different fields of view within the video.

5.3.2 Areas of investigation

Some areas are to be studied as part of VR video streaming, such as:

1) Use of adaptive streaming formats for 360 video delivery, with content adaptation and/or selection affected by, among others, the following factors:

   a) Changing network conditions.

   b) Device capabilities.

   c) Field of view in response to user interaction, e.g. head movement.

2) 360 video content provisioning and streaming techniques toward improving the network bandwidth efficiency without compromising user experience and interactivity.

5.4 Distributing 360 A/V content library in 3GPP

5.4.1 Introduction: content library and target devices

A Service provider has access to a library of 360 A/V content. The library is a mixture of content formats from user generated content, documentaries, promotional videos, as well as highlights of sports events. The latter content is replaced and updated daily. The content enables to change the field-of-view based on user interaction. The Service provider wants to create a portal to distribute the content to mobile devices. The service provider wants to target two types of applications:

- Primarily, view with an HMD with head motion tracking. The Service provider expects different types of consumption and rendering devices with different capabilities in terms of decoding and rendering capabilities.

- As a by-product, the content provider may permit viewing on a screen with the field-of-view for the content adjusted by manual interaction (e.g. mouse input or finger swipe).

The Service provider has access to the original footage of the content and may encode and transcode it to appropriate formats. The footage includes different types of VR content, including:

- For video:

  - The video is spherical and may cover the full 360 sphere or a subset of the 360 sphere.

  - Monoscopic video, i.e. a spherical video without real depth perception with Equirectangular Projection (ERP).

  - Stereoscopic video, i.e. a spherical video using a separate input for each eye with ERP.

  - Original content:

    - Original content, either in on original uncompressed domain or in a high-quality mezzanine format.

    - Basic VR content: as low as 4k x 2k (ERP), 8 or 10bit, BT.709, as low as 30fps

    - High-quality: up to 8k x 4k (ERP), 10 bit, possibly advanced transfer characteristics and colour transforms, sufficiently high frame rates, etc.

    - Sufficient metadata is provided to appropriately describe the A/V content.

- For audio:

  - Spatial audio content for immersive experiences, provided in the following formats:

    - channel-based audio;

    - object-based audio;

    - scene-based audio; or

    - combination of the above.

- Sufficient metadata for encoding, decoding and rendering the spatial audio scene permitting dynamic interaction with the content. The metadata may include additional metadata that is also used in regular TV applications, such as for loudness management.

- Diegetic and non-diegetic audio content.

The service provider wants to reach as many devices as possible and wants to minimize the number of different formats that need to be produced while ensuring that the content is presented in highest quality on the different devices.

In particular, the service provider wants to reach devices that are already in the market or emerging mobile devices.

5.4.2 Downloading content

The service provider wants to enable that a certain amount of the content can be downloaded to devices through HTTP and is played back on the device after downloading. The service provider wants to ensure that a device downloads only content that it can decode and render while providing the best user experience for the device capabilities.

5.4.3 Non-real-time MBMS distribution of VR content

Many of the devices support MBMS file delivery services. The service provider agrees with an MNO that certain content is pre-cached on mobile devices for offline consumption using MBMS file delivery services. The service provider wants to ensure that MBMS delivered content can be consumed by as many devices as possible.

5.4.4 Streaming content

For certain contents, the service provider wants to ensure that content is rendered instantaneously after selection, so a DASH-based streaming is considered. The service provider wants to ensure that a device accesses only content that it can decode and render while providing the best user experience for the device capabilities. The service provider also wants to ensure that the available bandwidth for the user is used such that the rendered content for the user is shown in the highest quality possible.

5.5 Live services consumed on HMD

5.5.1 Introduction

Roger has subscribed to a regular live TV service from a service provider that is distributed over 3GPP networks. The service provider shows a tennis match of a Grand Slam tennis tournament and Roger is watching the TV program on his mobile device using a HMD. The TV program is rendered on a virtual screen in front of him. During the regular break while regular subscribers get ad breaks, the premium subscriber is offered the highlights of the recent points in an immersive environment providing a 360-degree view from different seats in the stadium. Roger can now consume during the ad break the full immersive experience for the highlights from a stadium viewpoint.

5.5.2 Streaming content

The service provider offers the first rounds of the tournament as a unicast DASH-based live streaming service. The service provider wants to ensure that a device accesses only content that it can decode and render while providing the best user experience for the device capabilities. The service provider also wants to ensure that the available bandwidth for the user is used such that the rendered content for the user is shown in the highest possible quality and with an end-to-end latency that matches TV services.

5.5.3 MBMS delivery of content

With the success of service and the tournament heading towards the finals, the service provider decides to offer the live TV service over MBMS to address scalability issues. Both the regular TV content as well as the HMD-dedicated content are delivered over MBMS. The service provider wants to ensure that MBMS delivered content can be consumed by as many devices as possible.

5.5.4 Mixed delivery of content

With even more success of the service at the next tournament, the service provider decides to offer the live TV service over MBMS. The regular TV including regular ads are delivered over MBMS, but different 360 experiences are created and provided over unicast streaming and the user can select different views.

5.6 Social TV and VR

5.6.1 Base use case

Alex and his friends have signed up to a service provider for a new live football experience. Alex watches a live football match on his HMD. His friends do the same, each at different places. All of them wear a headset with microphone to communicate. They connect socially to at least communicate verbally with each other. Everyone sees the football match in such a way that they can comment and discuss the actions of the match live as if they would sit in front of the same TV. They experience at least the sound communication as if the friends are sitting next to each other at different fixed positioned seats.

5.6.2 360-degree experience

In an extended version, each of them are placed at a virtual stadium seat and watch the match while being able to follow the action by head movement.

5.7 Cinematic VR use cases

5.7.1 Introduction

Both interactive and linear cinematic experiences are possible with VR. In interactive VR, the viewer can experience/interact with the story as an actual observer in the middle of the action. Although this has been realized with traditional video techniques, the unique immersive experience of VR offers new appeal for this use case.

At content creation, additional post-production and related technical considerations are necessary, if e.g. revoicing the cinematic VR experience to different languages need to happen. These considerations are applied before encoding for delivery.

5.7.2 Linear cinematic VR

A viewer enjoys a VR movie from a fixed point in the scene. The viewer of the movie can freely turn and move his head to observe other details in the scene and may follow the actors by listening and watching. The viewer can change his view to explore the details in the scene, looking around freely. At the same time the user experiences an accurate binaural representation of the scene, so that he can follow the story without the necessity to (visually) search the virtual scene to find the main action. The accurate audio rendering supports the natural behaviour of the viewer to look around, and then is able to quickly re-focus on a sudden event by 'hearing' accurately where it happens.

As an alternative to binaural rendering, the audio can be rendered through loudspeakers in a listening environment using a multichannel loudspeaker setup.

5.7.3 Interactive cinematic VR

Interactive Cinematic VR poses some paradigm shifts for movie production. The viewer can look anywhere within the scene and may make decisions as to how the story would branch, removing some of the control that directors have over framing, audio effects, etc. In the case of interactive stories, a number of video clips corresponding to the user choices in the conduction of the story are required.

5.8 Learning application use cases

5.8.1 Introduction

VR can enable a number of educational applications with varying degrees of interactivity. At a basic level, the ability to watch a pre-recorded class remotely with spherical video and audio enables a higher level of immersion and assists learning. The student is able to view other students that are physically present in the class, detect which student is making a question, etc.

If immersive and interactive environments are offered to a student, psychological processes similar to when people construct knowledge through interaction with objects take place. Virtual environments created for learning tasks in this context are therefore expected to be highly interactive and can benefit from both object and scene-based 3D audio capture and rendering.

Examples of interactive learning applications include surgeon's training, interactive learning games for K-12 students, etc.

5.8.2 Remote class participation

In a Remote Class Participation situation, the remote student can be assigned to a "virtual seat" in the class. The "virtual seat" corresponds to the location where a 3D camera and microphone are positioned in the classroom, allowing the student to participate in the class remotely.

Remote Class Participation can be realized through streaming of the 3D audio/video content. In a streaming situation, the student would be able to follow the class remotely and view/listen other students physically present in class, leveraging the spatial audio cues delivered by the Virtual Reality system to know which student is talking. This application is analogous to the Event Broadcasting application described in Clause 5.2.

In a more interactive scenario, the student participates in real-time with the class, making questions to the professor or interacting with colleagues. This would have strict latency constraints so that the class flow is not disrupted.

5.9 VR calls use cases

5.9.1 Spherical video based calls

In a Spherical Video based call, two parties communicate with each other. The parties experience the feeling of being immersed in the far-end party's environment. For example, one call participant may be calling home from a beach. The call participant at home is able to have a spherical video and audio experience of the beach location.

On the capture side, the Spherical Video Call can be accomplished through the use of an omnidirectional camera and 3D audio capture microphone arrays, possibly augmented through the use of certain speech processing techniques.

On the rendering side, the Spherical Video Call can be accomplished through the use of a Head Mounted Display (HMD) and spatial audio rendering techniques, including binauralization of audio presented over headphones.

Few opportunities exist for audio manipulation outside of real time speech processing due to the latency constraints of two-way full duplex communication. Techniques that can capture the 3D sound field/audio objects and render it in real time with minimal latency would be useful to enable such an experience.

5.9.2 Videoconferencing with 360 video

Anne is holding a video conference call with her team from her home office. The team is located in a meeting room around a 360 camera A microphone array is located on top of the camera to capture spatial sound. The conferencing application on her tablet shows a section of the 360°-view of the conferencing room. On voice activity, the camera shows the person currently talking. Anne can swipe the image to look into any direction in the conference room.

Anne is wearing a stereo headset. She can clearly localize the voices of all participants. The sound scene rotates as she swipes the screen keeping the sounds and images spatially aligned.

When Anne talks to her team, the speech signal is played back with low latency from an integrated speaker located below the camera.

5.9.3 Gap analysis for VR calls use cases

The following gaps are observed within the related 3GPP services, namely MTSI in TS 26.114 [37] and IMS-based Telepresence in TS 26.223 [38] from a media handling perspective:

To support VR calls use cases with spatial audio and 360 degree video, it may be necessary to define SDP-based mechanisms for the negotiation of VR capabilities across MTSI / TP senders and receivers during both call setup and mid-call. VR capabilities here may include the related codecs, formats and media handling mechanisms for delivery of spatial audio and 360 degree videos.

The more detailed gap analysis in Clause 9.3.1 is also expected to apply for this use case. The identification of additional gaps is for further study.

5.10 User generated VR use cases

5.10.1 User generated 360 video and audio recording

Anne is attending a rehearsal of her friend Fred's band and he asks her to share her experience with their friends. Anne uses a handheld 360° camera that captures 360° video and spatial audio. Anne uploads the content to a media store and shares a link with her friends.

Her friend Ben downloads and plays the recording on his VR glasses, connected to the smartphone using headphones. He can turn his head to see all musicians in the rehearsal room. He hears the sound from the band and the crowd from the corresponding locations and also Anne's voice commenting the show.

5.10.2 User generated live streaming - "See what I see"

Anne is attending a street parade and likes to share her experience with her friend Ben. Anne uses a smartphone with an integrated 360° camera. The phone uses multiple microphones with post-processing for immersive audio.

Ben accepts the incoming stream and consumes the parade on his smartphone using headphones. He can turn the phone to see the crowd or marching band. He hears the sound from the band and the crowd from the corresponding locations and also Anne's voice commenting the show.

5.11 Virtual world communication

5.11.1 In-Game Communications

In in-game communications, players in an online virtual game are able to hear and talk to the players within their immediate vicinity inside the virtual game. The 3D audio scene rendered for each player will be in full agreement with the player's 3D visual scene. The sound from player B at relative position P_{AB} from player A, will be rendered in the audio scene of player A such that player A will hear it coming from the relative position P_{AB}. Similarly, the sound from player A at relative position P_{BA} from player B, will be rendered in the audio scene of player B such that player B will hear it coming from the relative position P_{BA}. By using a Head Mounted Display (HMD), the rendering can be controlled by head tracking so that the positions of other players are updated as the user turns his/her head.

Each player will obtain the audio signals of its nearest players (up to a maximum limit) and a local audio rendering client will render those signals at the relative positions provided by the game engine. The rendering of other game sounds, such as environmental sounds or event sounds, may be handled separately by the game client software or they are treated as additional virtual sound sources in the same rendering as the communication. The game engine may be local (client-based) or remote (cloud-based).

In the envisioned use case the service may be initiated by the game that connects to VR communication through some APIs. The individual users may be identified by the underlying communication service, but the positioning of the audio sources in the game is done by the gaming engine. The local gaming client may receive or render all or a subset of the audio sources, for example depending on their proximity. The VR scene is defined by the gaming engine, of which the communication service is unaware.

Another use case example is that the players, through a VR conference server, firstly initiate a VR communication. The conference server then creates a virtual scene in connection with a gaming server. The conference bridge may in this case expose certain communication data, e.g. speaker identities and other metadata, to the gaming server. That server controls the virtual positions of the players in response to the provided data and the game context.

Since in-game communication is a full duplex service, a sufficiently low end-to-end delay is crucial in order provide a good user experience.

5.11.2 Virtual Meeting Place

The main idea here is to create a virtual world where people can meet and interact anonymously through their avatars with other people. A user would be able to move freely in the virtual world (6 DOF) and mingle with different groups of people depending for example on the discussion they are having. In this scenario, the user would be able to speak to other users in his/her immediate proximity and obtain a spatial rendering of what the other users in his/her immediate proximity are saying and would hear them from the same relative positions they have to him/her in the virtual world.

5.12 HMD-based legacy content consumption

5.12.1 Introduction

Unlike use cases based on 360-degrees video and audio formats, the immersive experience is not created by the delivered media itself but rather by the application-generated scene inside which the media is presented.

5.12.2 Private TV

Bob is at home and wants to watch a nature documentary from the 3GPP TV bouquet but his family has decided to rather watch a song contest, "*The best 3GPP Voice"*, on TV. Bob then decides to use his HMD for watching his documentary while his family watches their favourite program on TV.

Bob receives the TV programs usually distributed over 3GPP services to his smartphone. They are mapped by the device into a 360 environment as shown in Figure 5.3. Bob has multiple choices for selecting the virtual environment such as a movie theatre or a living room, etc.

The TV channel is rendered in front of him by default in a virtual screen but may be fixed, meaning that head movement tracking is not active.

Figure 5.3: TV channel mapping into a 360 environment

5.12.3 Floating mosaic

Bob browses TV contents on his HMD. He receives multiple streams in thumbnail version that are displayed around him on small virtual screens, as illustrated in Figure 5.4. He can quickly monitor a channel by turning his head in the direction of the desired channel. When pointing to a mosaic channel he can hear the associated audio.

Figure 5.4: Mosaic representation in a 360 environment

When interested by a specific channel, Bob selects it and the mosaic channel becomes the main channel with a higher resolution and is centered in front of the user.

5.12.4 Enriched TV

Bob is watching a TV program on his HMD. The TV program is rendered on a virtual screen in front of him. A 360 scene covers the background of the spherical environment.

Bob can activate the contextual menu so as to display additional information related to the current program such as the EPG (electronic program guide), a selection of different viewing angles, etc. Figure 5.5 illustrates such a use case.

Figure 5.5: enriched TV representation in a 360 environment

5.13 Use cases for Highlight Region(s) in VR video

5.13.1 Initial view point for on-demand content

Bob starts an on-demand VR session. The scene around an initial view point needs is projected for him. Instead of letting Bob tuning in at a random view point of the spherical video, an initial view point that his HMD can display is provided, so that he can start his VR experience from a meaningful viewpoint.

5.13.2 View point for random tuning in

For cases where the user tunes in somewhere in the timeline of the content, e.g. when trick-play mode is enabled for on-demand content or when tuning in at a random moment in a live content, starting with a randomly provided view point may degrade the experience for the user. Hence, it would be helpful to signal, to the VR application, the recommended view point to start with at any point in time of the media timeline.

5.13.3 Director view from artistic intent

When producing 360 content, the author may want to convey his/her view on what a director wants the audience to focus on. This information may be used by different entities along the media chain from the production to the consumption end. For instance, the 360 content can be automatically navigated by generating a cut out view of the sphere according to the director's intention when the user does not have control of the viewing orientation or has released control of the viewing orientation.

5.13.4 Region-of-Interest (RoI)-driven content consumption

Bob selects a 360 content on a non-HMD (e.g. a mobile or tablet), the VR application offers to Bob the possibility to switch from one predetermined view to another. In this case, the application detects the number of highlight regions signalled in the 360 content for this purpose, and proposes to Bob different viewing directions.

Viewers using an HMD may also benefit from this approach, where the 360 content consumption may be offered both in an full-immersion mode, where the user can move around the 360 content, or in a "theatre mode", that enables a more traditional and relaxed watching experience, or the possibility to share the VR experience with other participants inside or outside the virtual space. Additionally, the ability to transition from these two modes may also be offered to the user.

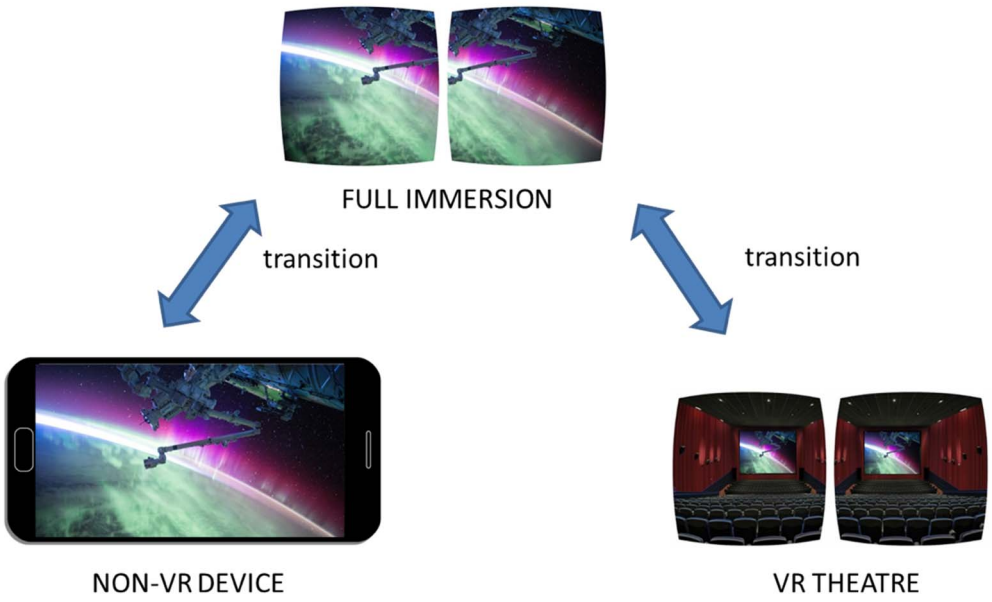


Figure 5.6: RoI driven 360 content consumption

Figure 5.6, shows both use cases where the watching experience is in full-immersion mode (top) and driven by the RoI, e.g. watching the content on a tablet/phone or in a virtual Theatre like mode in VR (bottom). If transition between the two modes is enabled, it should be smooth.

6 Audio quality evaluation

6.1 Audio quality evaluation of scene-based formats

6.1.1 Introduction

For a successful deployment of VR, audio formats should deliver a high Quality of Experience. One goal of these systems is to bring a sense of *presence* to the consumers [18]. *Presence* is defined as a cognitive state, the (psychological) feeling of being in the virtual environment [19], [20]. The physical interfaces involved in VR, and extreme demands on video system performance, often pose practical limitations for the evaluation of *presence*. Therefore, other criteria have also been proposed in the literature. For example, *plausibility* has been defined as "a simulation in agreement with the listener's expectation towards an equivalent real acoustic event." [21].

Regardless of the definition one chooses to adopt, from an audio perspective, either *presence* or *plausibility* demand systems that can deliver, at a minimum (adapted from [21]):

1) a low motion to sound latency, below the detection threshold with seamless adaptation of the audio scene as a function of the user's head motion for at least 3DOF;

2) binaural sound reproduction with undistorted magnitude and phase frequency responses, ideally compensated for the individual listener;

3) stable and accurate sound localization properties in an omnidirectional space; and

4) transparent or near-transparent audio quality.

By taking advantage of low-level psycho-physical characteristics, VR systems and omnidirectional videos can make the user believe that they are somewhere else and/or somebody else. This is achieved by presenting audio-visual realities and allowing the user to naturally interact with it. If the technology is not capable of delivering a neurologically convincing audio-visual experience, the sense of presence and immersion breaks down and sensory sickness may even occur.

To better understand the impact of the four quality prongs identified above, and other quality aspects of VR scene-based audio within the 3GPP context, this sub-clause collects results of listening tests and other perceptual experiments.

6.1.2 Report of one ITU-R BS.1534-3 binaural listening test for *basic audio* quality of encoded scene-based audio content with non-individualized HRTF and non-equalized headphones

Introduction and Test Method

ITU-R recommends that the "testing, evaluation and reporting procedures given in Recommendation ITU-R BS.1534-3 [22] be used for the subjective assessment of intermediate audio quality". Recommendation ITU-R BS.1534-3 [22] has also been previously used in other standardization activities pertaining to spatial audio coding such as the MPEG-H standardization. To provide an understanding of what quality levels can be achieved for scene-based audio content at reduced bit-rates, a test per Recommendation ITU-R BS.1534-3 [22] was conducted with a few scene-based audio encoding solutions.

In this experiment, 3 different coding solutions at various bit-rates were tested with 13 different HOA test materials of HOA order 3 (i.e. 16 coefficient channels). One of the solutions is the HE-AACv2 codec used to encode stereo pairs, a 2nd solution is the MPEG-H Low Complexity profile, and a third coding solution was included in the test to diversify the MUSHRA experiment (not relevant for 3GPP analysis purposes). These solutions are scored against the HOA reference signal (PCM).

The experimental design was such that all listeners listened to all test conditions for each system under test.

The testing methodology introduces distortions and localization errors associated with:

1) the limited speaker layout channel count in the renderer;

2) the differences between the single HRTF used and the assessor's individual HRTF; and

3) distortions in the headphone frequency response. Especially in the practically relevant case when using non-individualized HRTF and non-equalized headphones, spectral colouring distortions and more or less severe localization errors are frequently observed. Note that the testing methodology does not include any measure of absolute spatial accuracy. The test conditions were simply compared to the reference condition and all used the same non-individualised HRTF, with no indication of the intended spatial position. A testing methodology that would alleviate these problems should be ffs and may include presentation over loudspeakers and/or use of individualized HRTFs.

Test Materials

The test materials included thirteen different signals covering a range of audio content typical of broadcasts such as vocals, orchestra music, nature sounds, etc. Each of the five different coding solutions was tested with each of the thirteen different signals.

Listening conditions

For presentation, all test material was rendered to a 22.2 speaker layout (CICP 13) and binauralized. The binauralized stimuli were presented over headphones to the participants. No individualized HRTF or headphone compensation were used in this test and this may affect the absolute quality of the samples. No head-tracking was used in this test.

Presentation interface

The ARL STEP software was used for presentation of the samples and collection of results.

Attributes

Participants were asked to consider all perceptual differences between the systems under test and the reference signal when scoring the *basic audio quality*.

Assessors

The participants were all members of Qualcomm Advanced Tech R&D group and familiar with critical listening and audio quality evaluation. Post-screening of assessors was per ITU-R BS.1534-3 [22], section 4.1.2 and all assessors passed post-screening.

Size of listening panel

The listening panel contained eight assessors.

Systems under test

Higher Order Ambisonics test items truncated to 3rd Order Ambisonics (i.e. 16 HOA coefficients) were used in the test. Relevant systems under test are listed on Table 6.1.

Table 6.1: Systems under test

System under Test	Bitrate / condition
Hidden Reference	original test items
LP35	Anchor 3,5 kHz lowpass
System A (undisclosed)	600 kbps
MPEG-H LC profile	256 kbps
HE-AACv2 (FDK-AAC)	256 kbps

- The Hidden Reference provides a baseline for the *basic audio quality* without the encoding/decoding process.

- For a baseline of the audio quality level obtained by using an existing 3GPP stereo audio coder (eAAC+) in stereo pairs, the HE-AAC_v2 encoder (FDK-AAC) was configured as 8 stereo encoding instances using channel-pair elements (CPE) each with a target bitrate of 32 kbps (256/8 kbps). At this bit rate the HE-AACv2 codec insufficiently encodes phase relationships of signal pairs and, therefore, this configuration is unsuitable for coding of HOA audio. Note that the FDK-AAC encoder may be different than the eAAC+ encoder version in 3GPP, and different quality trade-offs may be obtained by reducing the number of CPE and increasing the target bitrate per stereo pair.

- Finally, for a baseline of the performance of an existing spatial audio coding solution at similar bit-rates, the MPEG-H Low complexity profile at 256 kbps was used to encode the HOA coefficients.

- System A (undisclosed) is a third coding solution that was included in the test for diversifying the MUSHRA test but not relevant for the 3GPP context.

- Note that these systems are not designed for conversational services and have delays that may be too high for certain 3GPP VR use cases.

All HOA coefficients were assigned to the encoder in the order of the Ambisonics Channel Numbering (ACN) scheme. To avoid potential clipping in the FDK-AAC encoder, the HOA coefficients were normalized per the SN3D normalization scheme prior encoding and rescaled after decoding.

Results

Figure 6.1 visualizes the absolute scores per test item and system under test. Table 6.2 further summarizes the results across all test items.

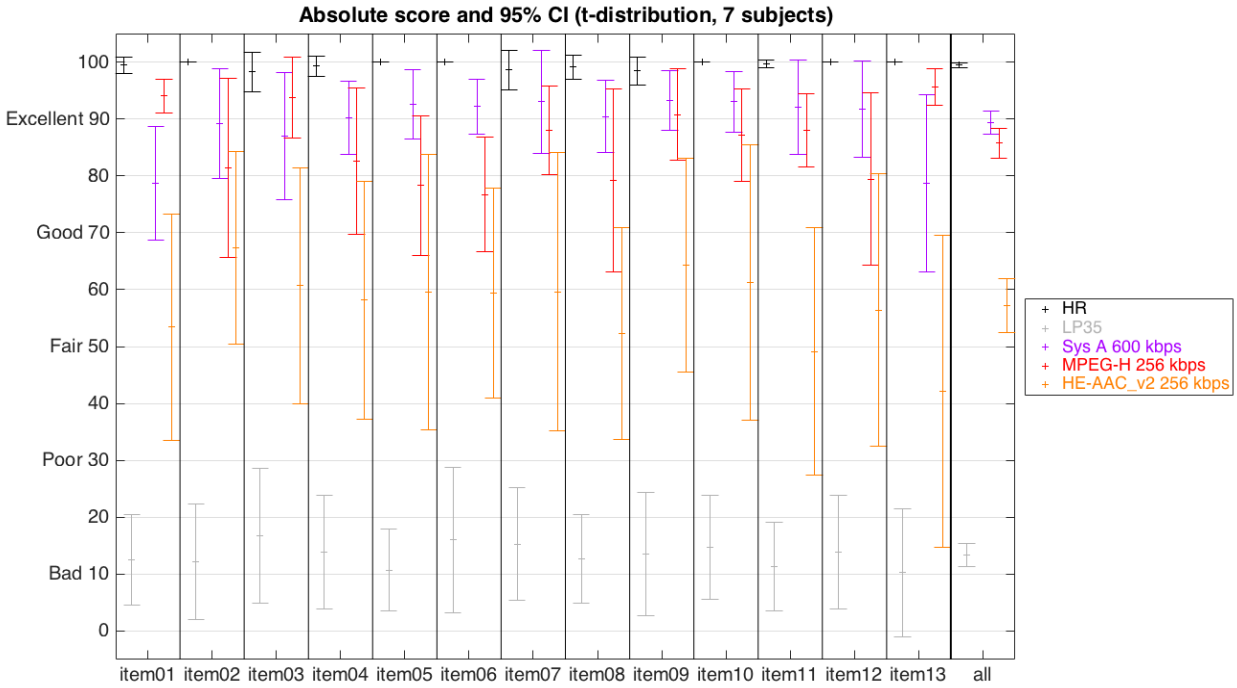


Figure 6.1: Absolute score and 95 % CI of basic audio listening test

Table 6.2: Summary of Average Scores

System	High	Low	Mean
Sys A @ 600 kbps	91,46	87,29	89,37
MPEG-H @ 256 kbps	88,33	83,17	85,75
HE-AAC_v2 @ 256 kbps	61,96	52,45	57,21
LP35	15,47	11,25	13,36
Hidden Reference	99,84	99,06	99,45

Observations

The following observations are made from the results:

1) For 3rd order Ambisonics encoded with HE-AACv2 at 256 kbps, the *basic audio quality* scores averaged across all test conditions is within the *Fair* range (40-60 MUSHRA points). Coding HOA natively with HE-AACv2 is not going to deliver good results at such low bit rates.

2) For 3rd order Ambisonics encoded with MPEG-H LC profile at 256 kbps, the *basic audio quality* scores averaged across all test conditions is within the *Excellent* range (80-100 MUSHRA points). Broadcast audio quality is met at 80 MUSHRA points which should also represent the minimum level of performance needed for VR broadcast applications.

3) A statistically significant difference exists between MPEG-H at 256kbps and HE-AACv2 at 256 kbps when used to encode 8 stereo pairs. This result highlights the importance of leveraging spatial audio coding techniques to reduce bit rate in scene-based audio coding applications.

4) The observed *quality scores* are only applicable to the specific rendering and binauralisation processes used and the conclusiveness of the results is limited by the distortions or localization errors associated with these processes.

6.1.3 Report of one ITU-R BS.1534-3 binaural listening test for *Localization quality* of synthetic scene-based audio content with non-individualized HRTF and non-equalized headphones

Introduction and Test Method

ITU-R recommends that the "testing, evaluation and reporting procedures given in ITU-R BS.1534-3 [22] be used for the subjective assessment of intermediate audio quality". ITU-R BS.1534-3 [22] has also been previously used in other standardization activities pertaining to spatial audio coding such as the MPEG-H standardization. To provide an understanding of what *localization quality* levels can be achieved for scene-based audio content at different Ambisonics orders, a test per ITU-R BS.1534-3 [22] was conducted comparing 1st, 2nd, 3rd and 6th order Ambisonics.

In this experiment, six different synthetic materials (i.e. created through spatially mixing monaural sound recordings into Higher Order Ambisonics coefficients) were used. The reference signals were 6th order Ambisonics contents. Lower order Ambisonics test material were created by truncating the spherical harmonics sound field decomposition into 3rd order, 2nd order and 1st order Ambisonics. All test material was PCM 32bit floating point 48 kHz.

The experimental design was such that all listeners listened to all test conditions for each system under test.

It is important to note that this procedure for generating content represents a condition that is quite distinct from the more common practical scenario of capturing Ambisonics content with a microphone array. Separate tests are needed to cover that space and different results can be expected.

The testing methodology introduces distortions and localization errors associated with:

1) the limited speaker layout channel count in the renderer;

2) the differences between the single HRTF used and the assessor's individual HRTF; and

3) distortions in the headphone frequency response. Especially in the practically relevant case when using non-individualized HRTF, spectral colouring distortions and more or less severe localization errors are frequently observed. Note that the testing methodology does not include any measure of absolute spatial accuracy. The test conditions were simply compared to the reference condition and all used the same non-individualized HRTF, with no indication of the intended spatial position. A testing methodology that would alleviate these problems should be for further study.

Test Materials

The test materials included six different sound scenes. Four of the sound scenes are from Bertet et al. [23], including three and four-party audio meetings, a kitchen and a classroom. These sound scenes are described in [23]. The two other sources include nature sounds and a movie trailer. Each of the five different Ambisonics solutions were tested with each of the six different sound scenes.

Listening conditions

For presentation, all test material was rendered to a 22.2 speaker layout (CICP 13) and binauralized. The binauralized stimuli were presented over headphones to the participants. No individualized HRTF, headphone compensation or head-tracking were used in this test.

Note that the lack of individualized HRTF and headphone compensation may have impaired the localization capability. This aspect is to be covered in further studies.

Presentation interface

The ARL STEP software was used for presentation of the samples and collection of results.

Attributes

Participants were asked to consider the localization of all directional sound sources by comparing the systems under test and the reference signal when scoring the *localization quality*.

Assessors

The participants were all members of Qualcomm Advanced Tech R&D group and familiar with critical listening and audio quality evaluation. Post-screening of assessors was per ITU-R BS.1534-3 [22], section 4.1.2 and two of the assessors were removed following post-screening.

Size of listening panel

The listening panel contained eleven assessors. Two assessors were removed following post-screening procedures.

Systems under test

Higher Order Ambisonics test items truncated to 3rd order (i.e. 16 HOA coefficients), 2nd order (i.e. 9 HOA coefficients) and 1st order (i.e. 4 HOA coefficients) were used in the test. In addition, 6th order (i.e. 49 HOA coefficients) test items were used as both reference and hidden reference signals. Relevant systems under test are listed on Table 6.3.

Table 6.3: Systems under test

System under Test	Bitrate / condition
Hidden Reference	Binauralized PCM 32-bit 48 kHz original test items (6th order HOA)
LP35	Anchor 3,5 kHz lowpass
FOA	Binauralized original test items truncated to 1st order Ambisonics
2nd order	Binauralized original test items truncated to 2nd order Ambisonics
3rd order	Binauralized original test items truncated to 3rd order Ambisonics

Results

Figure 6.2 visualizes the absolute scores per test item and system under test. Table 6.4 further summarizes the results across all test items.

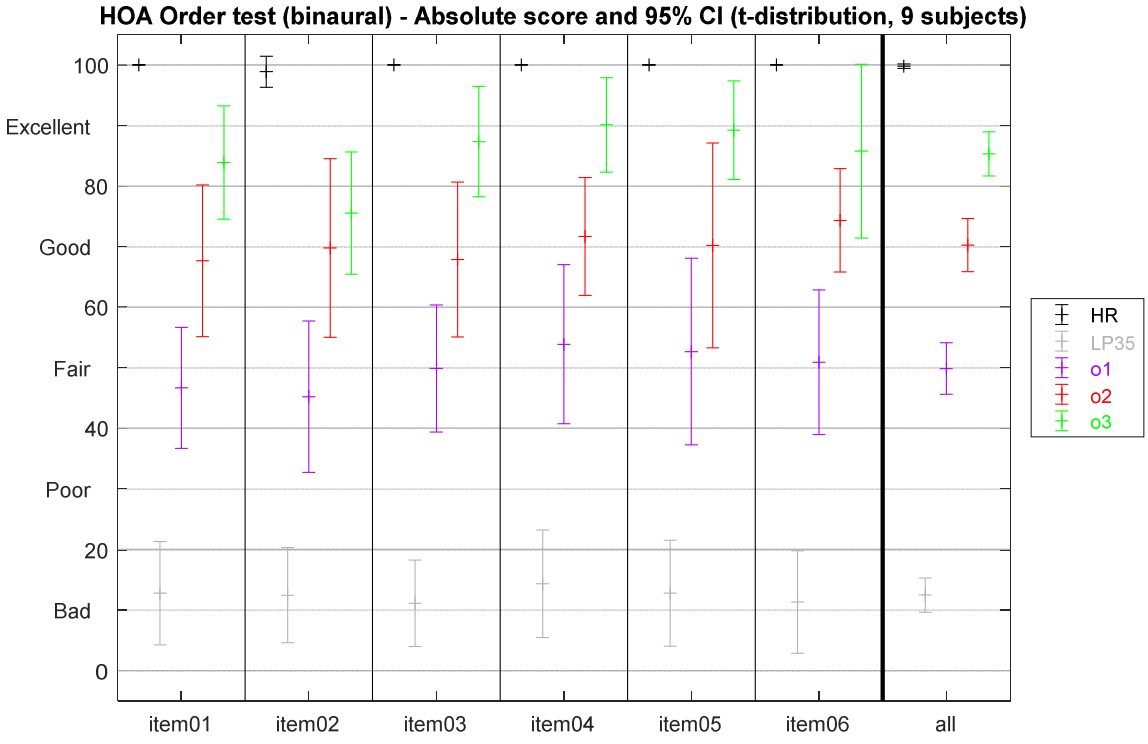


Figure 6.2: Absolute score and 95 % CI of localization listening test

Table 6.4: Summary of Average Scores

System	High	Low	Mean
3rd order	90,11	75,55	85,31
2nd order	74,33	67,67	70,26
1st order	53,89	45,22	49,87
LP35	14,33	11,11	12,47
Hidden Reference	100	98,89	99,81

Observations

The following observations are made from the results:

1) For 3rd order Ambisonics, the *localization quality* scores averaged across all test conditions is within the *Excellent* range (80-100 MUSHRA points). However, for one of the test items (item 04 - kitchen), the mean *localization quality* is within the *Good* range (60-80 MUSHRA points).

2) For 2nd order Ambisonics, the *localization quality* scores averaged across all test conditions and the mean of each test condition is within the *Good* range (60-80 MUSHRA points).

3) For 1st order Ambisonics, the *localization quality* scores averaged across all test conditions and the mean of each test condition is within the *Fair* range (40-60 MUSHRA points).

4) There is a statistically significant difference between the hidden reference (6th order) and the 3rd order HOA contents, indicating that benefits for synthetic content can still be achieved beyond 3rd order.

5) These *localization quality scores* are only applicable to the specific rendering and binauralization processes used and the conclusiveness of the results is limited by the distortions or localization errors associated with these processes. In addition, the scores are for uncoded audio material and do not consider additional artifacts introduced by coding. For practical 3GPP applications, audio bit-rate compression will be necessary and the absolute quality scores will differ.

6.1.4 Test of the ISO/IEC 23008-3 MPEG-H 3D Audio scene-based coding scheme

The MPEG-H verification test report [36] provides details on four large-scale listening tests that were conducted to assess the performance of the Low Complexity (LC) Profile of MPEG-H 3D Audio. Test material was either channel-based, channel plus objects, or scene-based, as Higher Order Ambisonics (HOA) of a designated order, possibly also including objects. Three tests were conducted over loudspeakers and one test over headphones binaurally rendered.

6.1.5 Listening test for synthetic scene-based audio content with loudspeaker rendering assessing overall and localization quality with written audio scene descriptions as reference

6.1.5.1 Introduction

In this experiment a listening test comparing the perceived overall and localization quality of different orders of Ambisonics was done. The test compared the performance of 1st, 2nd and 3rd order Ambisonics with synthetically created audio scenes with one or two point sources. Three anchor conditions were included in order to span the quality scale as evenly as possible. This test did not include higher orders of Ambisonics (> 3rd order), which are shown to provide a statistically significant better localization quality than 3rd order in clause 6.1.3.

6.1.5.2 Objectives

The main objective was to compare different Ambisonics orders. In order to capture not only the difference between the different orders of Ambisonics but also assess the suitability of the Ambisonics format compared to existing formats by 3GPP speech and audio services, two of the anchors were using a single (mono) respectively pair (stereo) of audio channels.

This particular test is focusing on the overall quality and accuracy of the perceived localization of point sources in virtual audio scenes, but did not cover all aspects of spatial reproduction, such as immersiveness etc. that are necessary to determine whether a particular Ambisonics format is suitable for an immersive VR experience.

6.1.5.3 Test methodology

In this test, a method inspired by the ITU-R BS.1534-3 [22] test methodology - but without explicit auditory references and post-screening of assessors - was used. In addition, in contrast to ITU-R BS.1534-3, there were two simultaneous ratings for each sound example, one for overall quality and one for localization accuracy, where the listeners were instructed to consider the direction of the sound and the localization blur. The use of more than one scale in the same trial have been tested successfully before in [65], and is also used in ITU-T P.806 [66] where 6 + 2 test questions are rated in the same trial.

References in form of written descriptions of the audio scenes were used. Written descriptions as references of spatial audio scenes have proven useful in earlier Ericsson internal tests, where no audio reference was available. According to the test subjects, the comparison of the perceived sound localization to the written description of the audio scene was a clear and achievable task. However, asking the listener to compare the audio direction with a textual reference has the drawback of focusing the listener's attention to the sound source direction as opposed to other spatial quality aspects (source width, height, etc.). It is noted that audio sources played out directly through the loudspeakers would be another type of reference, which was not considered in this assessment.

6.1.5.4 Physical test setup

In order to avoid the effect of HRTF filtering the test was carried out with a 16 speaker setup arranged as two circles, one at 0° elevation and one at +30° elevation. There were only two elevation angles, which might impact the generality of the conclusions. The two circles were offset 22.5° to distribute the speakers as evenly as possible. The speaker directions (elevation, azimuth) were known by the subjects as guidance for the assessment of the localization of the sound sources compared to the reference audio scene description. All speakers were positioned 2.0 m from the listening position as shown in Figure 6.3. The speakers chosen for this test, M-Audio AV-40, are compact 2-way speakers that, due to their size are limited in bass response but, on the other hand, provide a more defined acoustic center than bigger multi-driver speakers. The material used in the test was not relying on a response under 85Hz and therefore no subwoofer system was used.

The room used for the test is a well sound proofed and acoustically treated audio lab, rated as NR10, with short reverberation time.

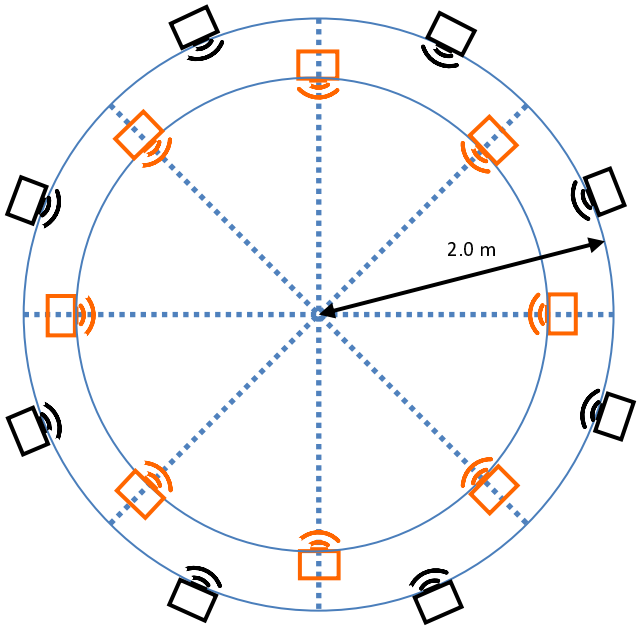


Figure 6.3: Loudspeaker setup. The black speakers are placed at an elevation of 0 degrees, and the orange speakers are placed at an elevation of 30 degrees. All speakers are placed at a distance of 2.0 m from the center.

The test subjects were instructed to keep their head straight ahead when listening. The testing software provided a function to guide the user into the exact sweet spot. This was done by playing pink noise through all the speakers and finding the right position by minimizing the panning and the phasing effects that can be heard when moving out of the sweet spot.

6.1.5.5 Test material

As the accuracy of the spatial localization was assessed, the audio material consisted of sound sources that would be perceived as typical point sources originating from certain points in space. The recordings used were all mono recordings with very little background noise and reverberation. The sources consisted of male and female voices and the sound of screws rattling in a glass.

10 different scenes were created by rendering one or two sound sources at different angles and mixing moving sources with stationary sources. The number of audio sources used was small, which might impact the generality of the conclusions, but allows for a better comparison between different audio scenes. Table 6.5 lists the scenes and the corresponding reference audio scene desriptions shown to the listeners. Scenes with more than one voice included partly overlapping talkers. The scenes circle_screws_high and two_speakers were used in a pre-test and the results for these are not presented as part of the main test results. Testing only synthetic content limits the generality of the conclusions.

Table 6.5: Audio scenes used in the test

Scene name	**Scene description**
circle_screws	The sound of rattling screws in a glass,moves 360 degrees in azimuth, in an even counter clockwize circular movement, starting from straight ahead (0 degrees), at fixed elevation 0 degrees
circle_screws_high	The sound of rattling screws in a glass, moves 360 degrees in azimuth, in an even counter clockwise circular movement, starting from straight ahead (0 degrees), at fixed elevation 30 degrees
circle_female	A sound of a female voice, moves 360 degrees in azimuth, in an even counter clockwise circular movement, starting from straight ahead (0 degrees) at fixed elevation 0 degrees
elevation_30	The sound of rattling screws in a glass, at fixed azimuth straight ahead (0 degrees), moves +30 degrees in elevation, in an even upwards movement, starting from an elevation of 0 degrees
scene_two_speakers_close	Two voices at fixed elevation of 0 degrees are heard from different azimuth angles. A female voice is heard from azimuth -30 degrees (front right) and a male voice is heard from azimuth -50 degrees (further to the right).
front_back	The sound of rattling screws in a glass at fixed elevation of 0 degrees is heard from two azimuth angles, first from azimuth -30 degrees (front right) and then from azimuth -150 (back right).
front_back_speech	Two voices at fixed elevation of 0 degrees are heard from different azimuth angles. A female voice is heard from azimuth +30 degrees (front left) and a male voice from azimuth +150 (back left).
two_speakers	Two voices at fixed elevation of 0 degrees are heard from different azimuth angles. A female voice is heard from azimuth +30 degrees (front left) and a male voice is heard from azimuth -150 degrees (right back).
two_speakers_one_moving	Two voices at fixed elevation of 0 degrees are heard, one female voice from azimuth -45 degrees (front right) and one male voice that moves clockwise from azimuth 0 degrees (straight ahead) to azimuth -180 degrees (straight back).
down_up	Two voices at fixed azimuth of 0 degrees are heard from different elevation angles. A female voice is heard from elevation 0 degrees (straight ahead) and a male voice is heard from elevation +30 degrees.

6.1.5.6 Test conditions

In addition to the Ambisonics of orders 1, 2 and 3, three additional conditions were evaluated to span the quality scale better. The conditions are described in Table 6.6.

Table 6.6: Conditions evaluated in the test

Short name	Description	Rendering details
AO0	0th order Ambisonics (Mono)	0th order Ambisonics using all 16 speakers.
AO1S	FOA (Stereo)	Rendered from FOA to two loudspeakers at +/-67.5 azimuth.
AO1A	FOA with attenuated harmonics	Rendered as ordinary FOA but the harmonic components, except *W*, were attenuated 6dB in order to provide an anchor point below FOA.
AO1	FOA	1st order Ambisonics using all 16 speakers.
AO2	2nd order Ambisonics	2nd order Ambisonics using all 16 speakers.
AO3	3rd order Ambisonics	3rd order Ambisonics using all 16 speakers.

The rendering of the loudspeaker signals was done using a basic encoding-decoding scheme illustrated in Figure 6.4.

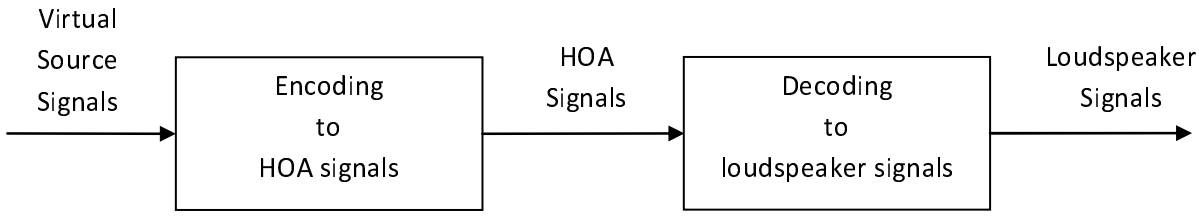


Figure 6.4: The Encoding-Decoding scheme used in rendering the different audio scenes

The encoding of the virtual sources to HOA signals was done by multiplying each sample of a source signal with the spherical harmonic transform vector Y_e mapping the current position of the source to the HOA beams. For moving sources the Y_e matrix was updated at every sample.

The decoding of the HOA signals to loudspeaker signals was done by first evaluating the spherical harmonic transform matrix Y_d that maps the loudspeaker positions to the HOA beams, then evaluating the decoding matix D as the pseudo inverse of Y_d , and then multiplying each HOA sample vector with decoding matrix D.

The ACN ambisonics channel order and the SN3D spherical harmonic normalization were used.

In the AO1S rendering, the first order HOA signals were mapped onto a loudspeaker configuration of only two loudspeakers at elevation 0 degrees and azimuth angles +/-67.5 degrees.

For the AO1A rendering the YZX components of the first order HOA signals were attenuated by a factor of 0.5 before being multiplied by the decoding matix.

The 16 loudspeaker signals were jointly normalized to an RMS level of -30 dBov, while the stereo channels were normalized to an RMS level of -39 dBov to be perceived similarily loud in average. The subjects were able to adjust the playback volume in a range of +/-4 dB, but were instructed not to change this setting while comparing the test samples.

6.1.5.7 Listening panel

The listening panel consisted of 9 experienced listeners of the Audio technology section at Ericsson Research. No post-screening of the subjects was made.

6.1.5.8 Software

The user interface used during the test was based on a typical MUSHRA test interface, but with no reference signal and with the addition of a second rating scale and a scene description text at the top that served as the reference in the test, see Figure 6.5. Using a scene description as reference limits the generality of the conclusions.

Figure 6.5: Software GUI used in the test

6.1.5.9 Test results

The test results are shown in Figure 6.6-6.9.

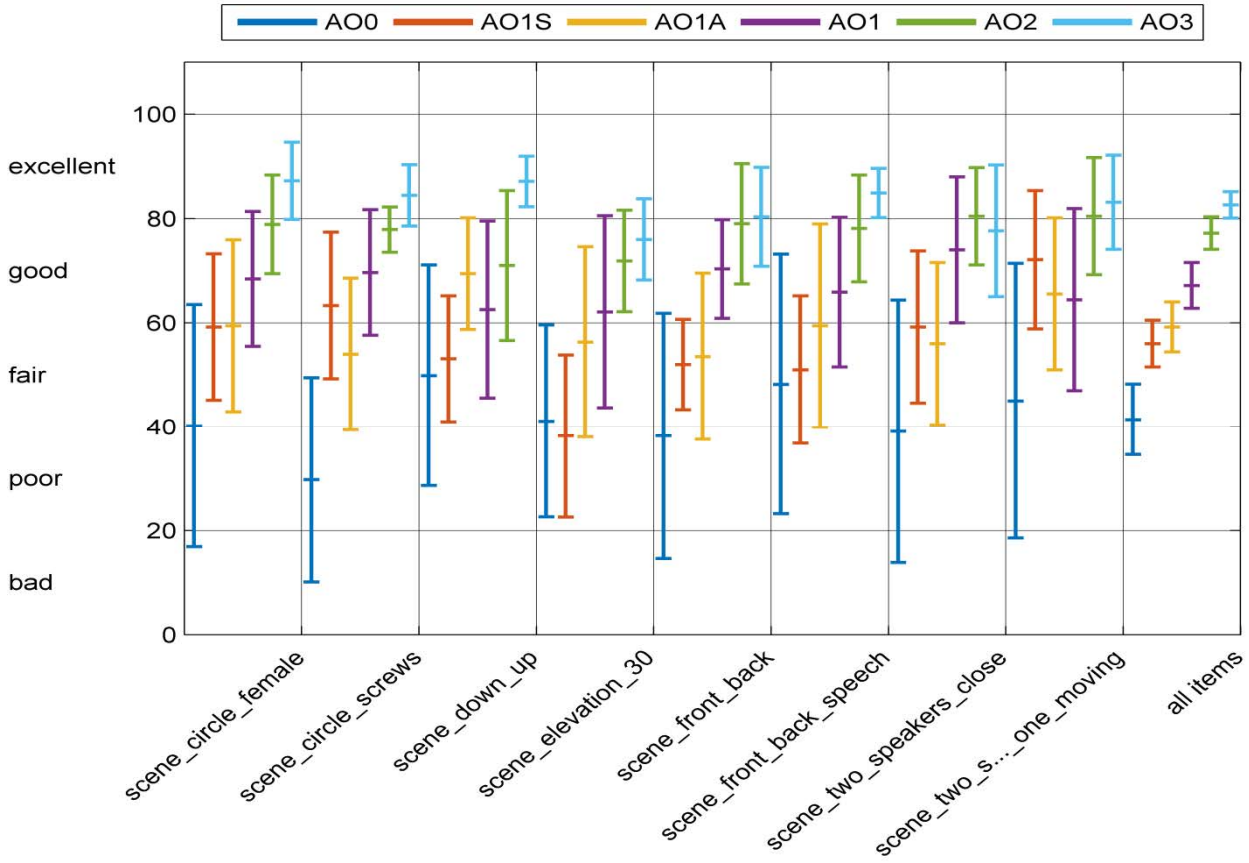


Figure 6.6: Absolute overall scores, with 95% CI.

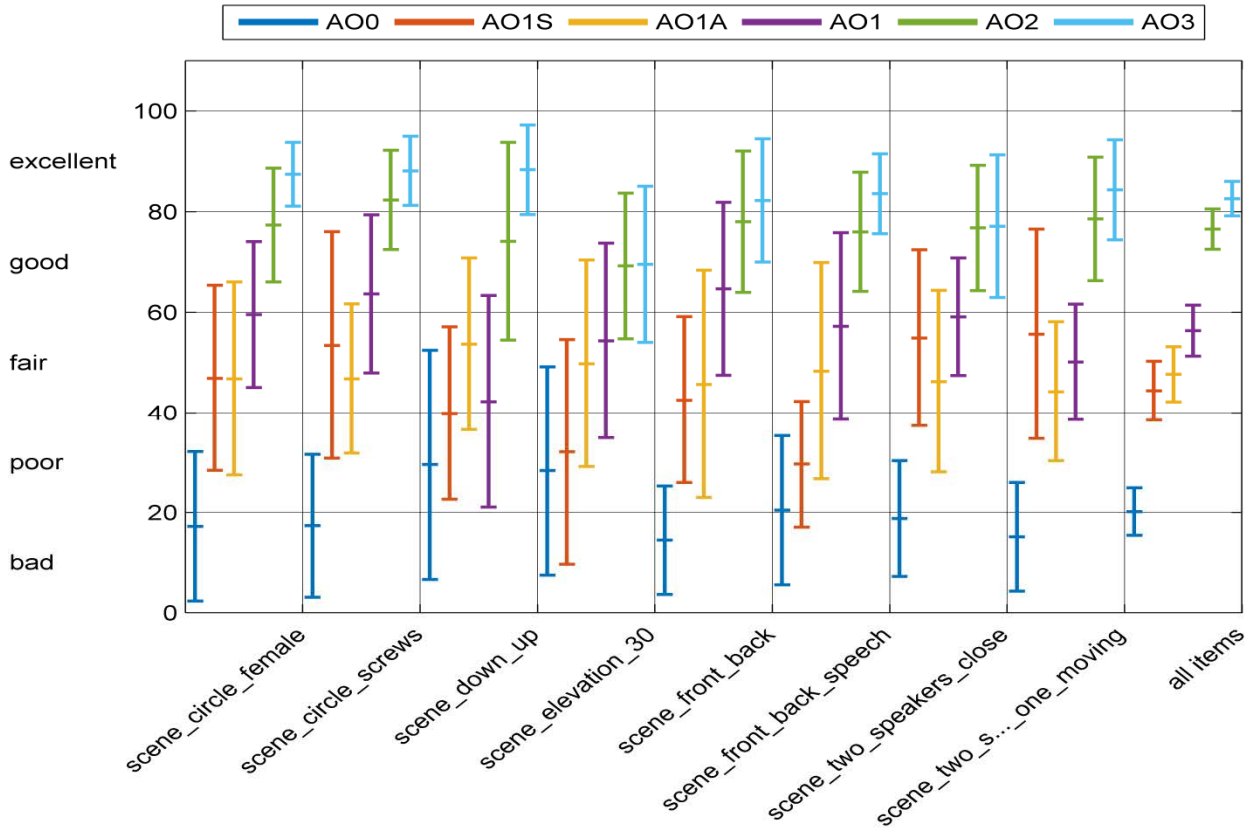


Figure 6.7: Absolute localization scores, with 95% CI.

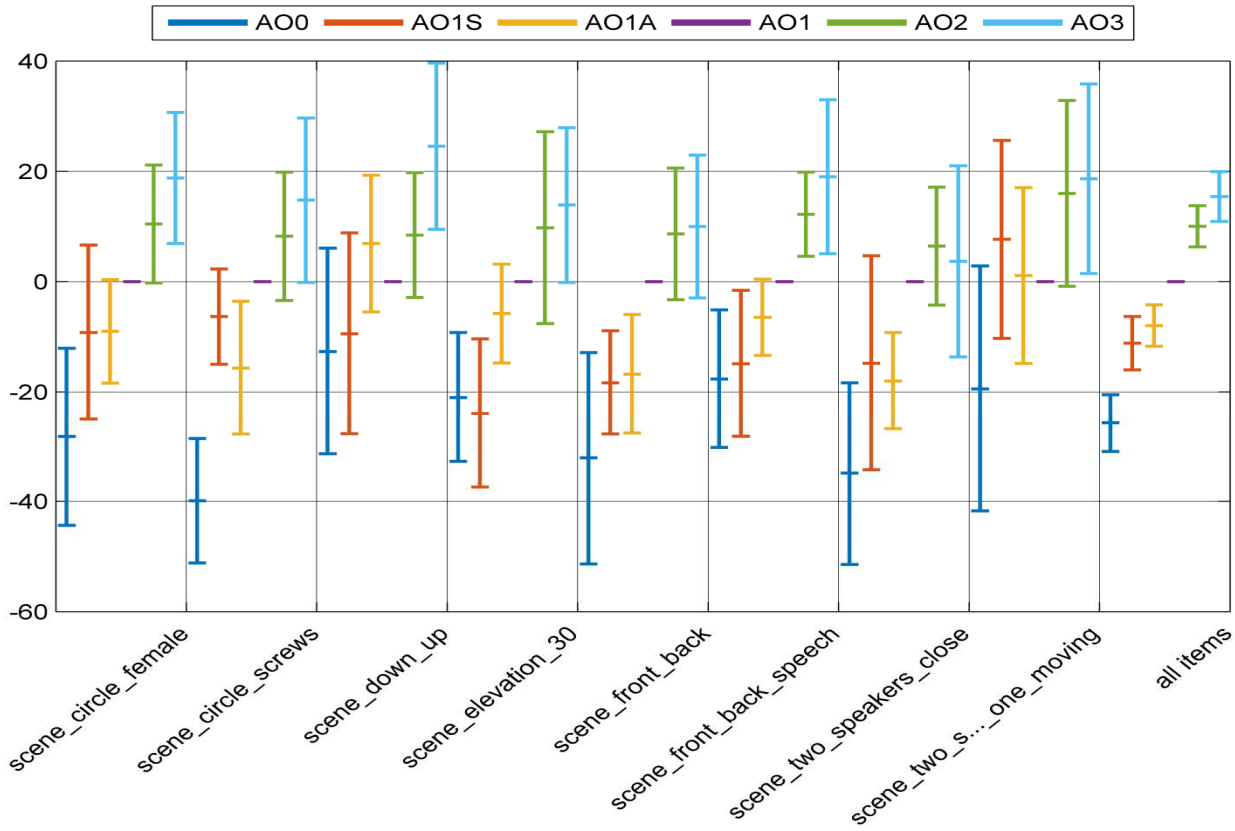


Figure 6.8: Difference overall scores relative to AO1, with 95% CI.

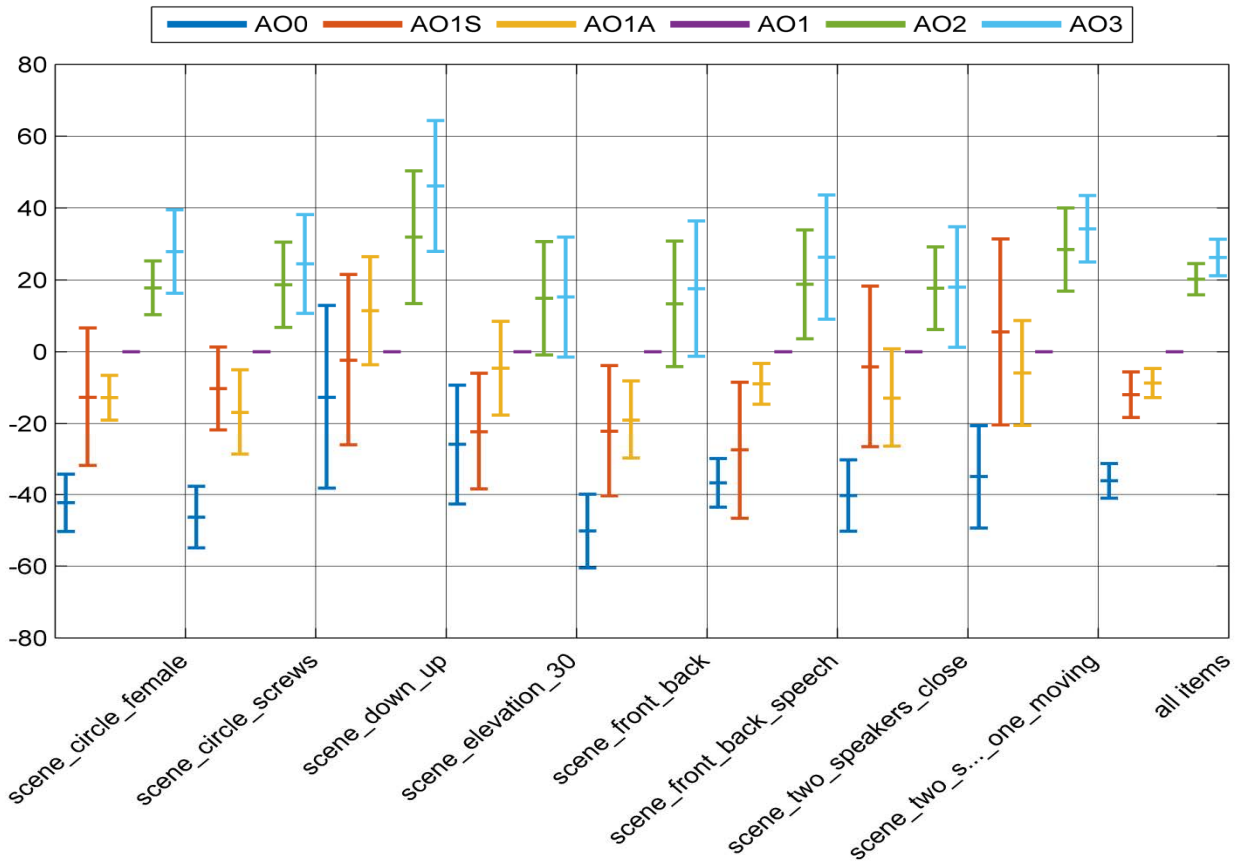


Figure 6.9: Difference localization scores relative to AO1, with 95% CI.

6.1.5.10 Conclusions

The results of the listening test indicate an increased overall quality and spatial localization accuracy with increasing orders of Ambisonics. Higher Order Ambisonics (HOA) with orders 2 and 3 both perform statistically significantly better than First Order Ambisonics (FOA). Further, Ambisonics orders 1, 2 and 3 all perform statistically significantly better than mono and stereo, i.e. what is achievable using existing 3GPP speech and audio services.

The overall audio quality measure seems well aligned with the spatial localization accuracy scores although the difference between FOA and HOA tends to be smaller.

This test did not include higher orders of Ambisonics (> 3rd order), which are shown to provide a statistically significant better localization quality than 3rd order in the test performed in clause 6.1.3. The absence of these higher quality conditions, or an explicit audio reference, may result in an overestimation of the scores for 1st, 2nd and 3rd orders.

6.1.6 Report of one test on encoding First-Order Ambisonics with 3GPP enhanced AAC+ with loudspeakers and with non-individualized HRTF and non-equalized headphones

6.1.6.1 Introduction

This contribution presents an experiment that was conducted to analyze the performance of the Enhanced aacPlus [67] codec used by 3GPP services when encoding First-Order Ambisonics audio representationsconverted from a 7.1.4 channel representation.

Specifically, this experiment uses the ITU-R BS.1534-3 [22] methodology to validate that dual stereo eAAC+ streams at reduced bit-rates is a viable transmission format for 4-channel B-format audio to be rendered over both a specific loudspeaker configuration (7.1.4) and binaural (headphone) endpoints.

The binaural testing methodology introduces distortions and localization errors associated with:

1) the limited speaker layout channel count in the renderer,

2) the differences between the single HRTF used and the assessor's individual HRTF and,

3) distortions in the headphone frequency response. Especially in the practically relevant case when using non-individualized HRTF and non-equalized headphones, spectral colouring distortions and more or less severe localization errors are frequently observed. Note that the testing methodology does not include any measure of absolute spatial accuracy. The test conditions were simply compared to the reference condition and all used the same non-individualised HRTF, with no indication of the intended spatial position.

The testing methodology over loudspeakers does not present the limitations listed in 2) and 3).

6.1.6.2 Test Method

In this experiment, a coding solution was tested at various bit-rates with 10 different test signals. The test signals were sourced from 7.1.4 channel-based content and converted to a First-Order Ambisonics representation, coded and then converted back to 7.1.4, through the process described in Clause 6.1.6.3. Given this process, the tests only assess the degradation caused by the encoding of the Ambisonics representation but not potential degradations caused by the format conversion itself.

The eAAC+ codec was used to encode stereo pairs at different configurations.

The test material used covers a range of audio content used to provide immersive experiences. The items were created from real recordings, as well as synthetic sounds and were authored in a post-production/mixing stage.

The test items were of the following type of content:

- amaze2: ambience, nature sounds
- audiosphere2: music, ambience
- bailando2: music, speech
- entity1: ambience
- leap1,2: ambience, nature sounds, speech
- santeria2: music, ambience, nature sounds
- silent6,7: music, ambience, nature sounds
- unfold1: ambience, synth-glitch hits

Testing was conducted in a critical, noise insulated (NC20) listening room with acoustic wall treatment.

Speaker-based listening and headphone-based listening tests were conducted.

For speaker presentation, all test material was rendered to a 7.1.4 speaker layout.

Loudspeaker delivery used Revel Salon 2 floor and Gem 2 ceiling speakers with Paradigm subwoofers. More specifically 5 inch single-cone full-range drivers were employed in a cabinet that is custom made for the listening room, to be as small as practicable possible. The speakers were placed at a distance of approximately 2.0m from the listener.

For headphone listening the material was subsequently binauralized. Headphones used were Sennheiser HD-600s with a Grace amplifier. No individualized HRTF or headphone equalization were used in this test, which may limit the generality of the results. No head-tracking was used in this test.

The participants were all members of Dolby Laboratories Inc. and experienced in audio quality evaluation. Post-screening of assessors was per ITU-R BS.1534-3 [22] section 4.1.2 and all assessors passed post-screening.

The listening panel contained 7 assessors for the loudspeaker listening test and 8 for headphone listening test. All listeners listened to all test content.

Participants were asked to consider all perceptual differences, including spatial characteristics, between the configurations under test and the reference signal when scoring the basic audio quality. They reported these measures of degradation on a 100 point ‘MUSHRA’ scale in accordance with the ITU-R BS.1534-3 [22] methodology, with standardized verbal anchors (100-80 is ‘excellent’, 60-80 ‘good’, 40-60 ‘fair’, 20-40 ‘poor’ and 0-20 ‘bad’).

6.1.6.3 Processing First-Order Ambisonics for Stereo eAAC+ Encoding

For the experiment, the original 7.1.4 signals were firstly down-mixed to First-Order Ambisonics representation. This was performed using a 4x12 matrix operation, with each speaker feed being panned to B format according to its direction of arrival, with no Near-Field-Compensation. In this operation, the LFE channel was ignored. The First-Order Ambisonics (FOA) signals were defined in terms of the standard Schmidt-Normalized ACN channel format (this is also often referred to as AmbiX format). As found based on theoretical considerations and confirmed by informal experiments with the content used in this test and further content, the four channels of the FOA signals were found less suitable for direct encoding/decoding via two stereo instances of eAAC+ at low bit-rates, due to the occurrence of out-of-phase components between each of the channel pairs.

Accordingly, the FOA signals were prepared for eAAC+ encoding by first remixing them into A-Format – in the form of 4 cardioid virtual-microphone signals. These 4 A-Format signals are defined in Table 6.7 as follows:

Table 6.7: Definition of A-Format signals

Signal Name	Orientation	Meaning
FL	Azimuth: +54.7°, Elevation 0°	Front left facing cardioid
FR	Azimuth: –54.7°, Elevation 0°	Front right facing cardioid
BU	Azimuth: 180°, Elevation +54.7°	Back upward facing cardioid
BD	Azimuth: 180°, Elevation –54.7°	Back downward facing cardioid

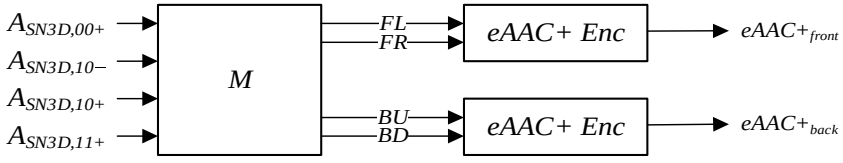
The four cardioid signals of A-Format are created from the FOA signals via a linear mixing matrix:

$$M = \begin{bmatrix} \frac{1}{2} & \frac{1}{\sqrt{6}} & 0 & \frac{1}{\sqrt{12}} \\ \frac{1}{2} & \frac{-1}{\sqrt{6}} & 0 & \frac{1}{\sqrt{12}} \\ \frac{1}{2} & 0 & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{12}} \\ \frac{1}{2} & 0 & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{12}} \end{bmatrix}$$

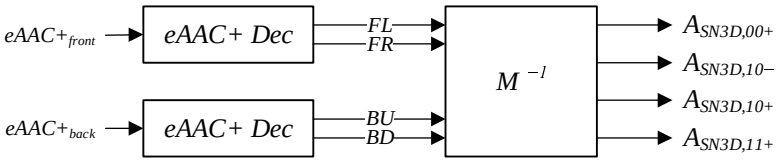
The mixing process is applied as follows:

$$\begin{bmatrix} FL \\ FR \\ BU \\ BD \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{\sqrt{6}} & 0 & \frac{1}{\sqrt{12}} \\ \frac{1}{2} & \frac{-1}{\sqrt{6}} & 0 & \frac{1}{\sqrt{12}} \\ \frac{1}{2} & 0 & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{12}} \\ \frac{1}{2} & 0 & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{12}} \end{bmatrix} \times \begin{bmatrix} A_{SN3D,00+} \\ A_{SN3D,11-} \\ A_{SN3D,10+} \\ A_{SN3D,11+} \end{bmatrix}$$

The overall process is shown in the following diagrams. Encoding is performed by mixing the FOA components into A-Format, splitting the A-Format signals into two stereo pairs, and encoding the stereo pairs with an eAAC+ encoder:



Decoding is performed by the reverse process:



wherein the M^{-1} matrix is defined as:

$$M^{-1} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{\sqrt{3}}{\sqrt{2}} & \frac{-\sqrt{3}}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & \frac{\sqrt{3}}{\sqrt{2}} & \frac{-\sqrt{3}}{\sqrt{2}} \\ \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} & \frac{-\sqrt{3}}{2} & \frac{-\sqrt{3}}{2} \end{bmatrix}$$

and hence, the FOA signals are recovered by the following matrix operation:

$$\begin{bmatrix} A_{SN3D,00+} \\ A_{SN3D,11-} \\ A_{SN3D,10+} \\ A_{SN3D,11+} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{\sqrt{3}}{\sqrt{2}} & \frac{-\sqrt{3}}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & \frac{\sqrt{3}}{\sqrt{2}} & \frac{-\sqrt{3}}{\sqrt{2}} \\ \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} & \frac{-\sqrt{3}}{2} & \frac{-\sqrt{3}}{2} \end{bmatrix} \times \begin{bmatrix} FL \\ FR \\ BU \\ BD \end{bmatrix}$$

Finally, the FOA signal went through an up-mix stage back to 7.1.4 representation, using a sub-band approach and direct panning to all of the 7.1.4 speaker channels.

Note: The $eAAC+_{front}$ signal may also be utilized as a stereo version of the original sound-field, being very similar to a standard “crossed cardioid” stereo recording (typical crossed-cardioid recordings make use of a pair of cardioid microphones at ±45°, whereas the FL and FR signals represent cardioids at ±54.7°). This stereo pair is suitable for listening on standard stereo playback devices.

Figure 6.10 visualizes the audio processing stages as described above.

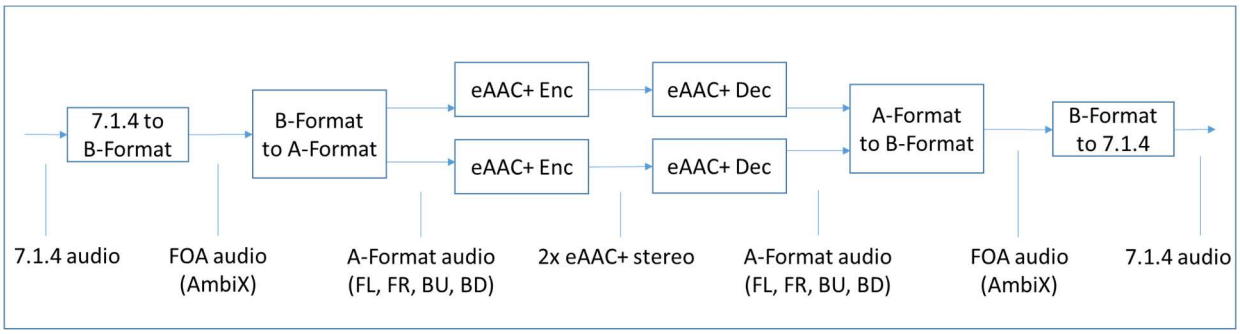


Figure 6.10: Audio processing stages

Table 6.8 summarizes the tested configurations.

Table 6.8: Configurations under test

Item	Description	Configuration
1	FOA reference	7.1.4->(downmix)->B-A-BFormat->(upmix)->7.1.4
2	7.0 LP Anchor	7.1.4->(downmix)->B-A-BFormat->(upmix)->7.1.4->(lowpass7.0k)->7.1.4
3	3.5 LP Anchor	7.1.4->(downmix)->B-A-BFormat->(upmix)->7.1.4->(lowpass3.5k)->7.1.4
4	eAAC+:2x96kbps	7.1.4->(downmix)->B-AFormat->(eAAC+:2x96kbps)->A-BFormat->(upmix)->7.1.4
5	eAAC+:2x64kbps	7.1.4->(downmix)->B-AFormat->(eAAC+:2x64kbps)->A-BFormat->(upmix)->7.1.4
6	eAAC+:2x48kbps	7.1.4->(downmix)->B-AFormat->(eAAC+:2x48kbps)->A-BFormat->(upmix)->7.1.4
7	eAAC+:2x32kbps	7.1.4->(downmix)->B-AFormat->(eAAC+:2x32kbps)->A-BFormat->(upmix)->7.1.4

6.1.6.4 Results

Figure 6.11 visualizes the absolute scores per test item and mean scores for all items. Table 6.9 below further summarizes the results across all test items. Figure 6.12 and Table 6.10 describe results for binaural delivery.

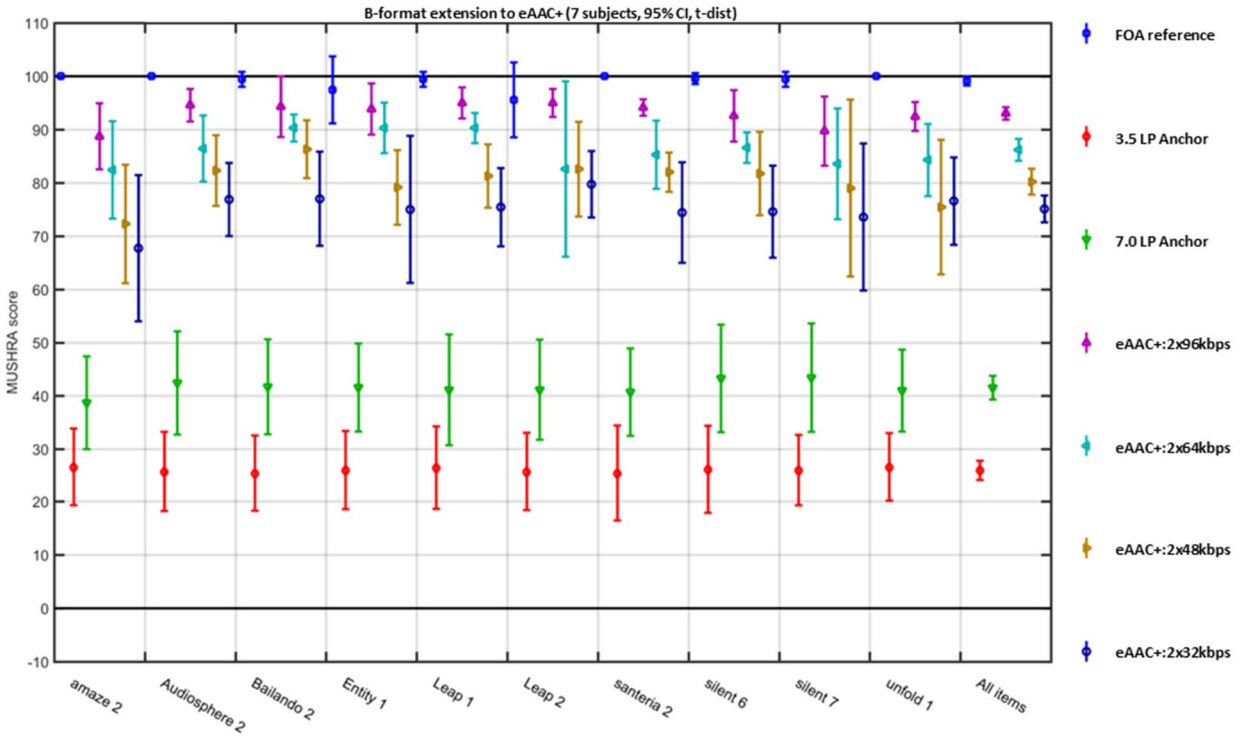


Figure 6.11: B-format extension using eAAC+, 7.1.4 loudspeaker configuration

Table 6.9: Loudspeaker scores

Condition	Amaze 2	Audiosphere 2	Bailando 2	Entity 1	leap 1	leap 2	santeria 2	silent 6	silent 7	unfold 1	all items
FOA reference	100	93	99	98	99	95	100	98	100	99	98
3.5 LP Anchor	28	26	24	27	28	27	27	25	26	28	27
7.0 LP Anchor	39	45	40	44	43	41	42	46	44	43	43
eAAC+:2x96kbps	88	93	93	93	95	93	92	93	86	95	92
eAAC+:2x64kbps	81	88	90	89	88	83	83	85	83	85	85
eAAC+:2x48kbps	72	80	83	74	82	79	82	81	79	77	79
eAAC+:2x32kbps	65	77	74	75	75	80	73	72	72	77	74

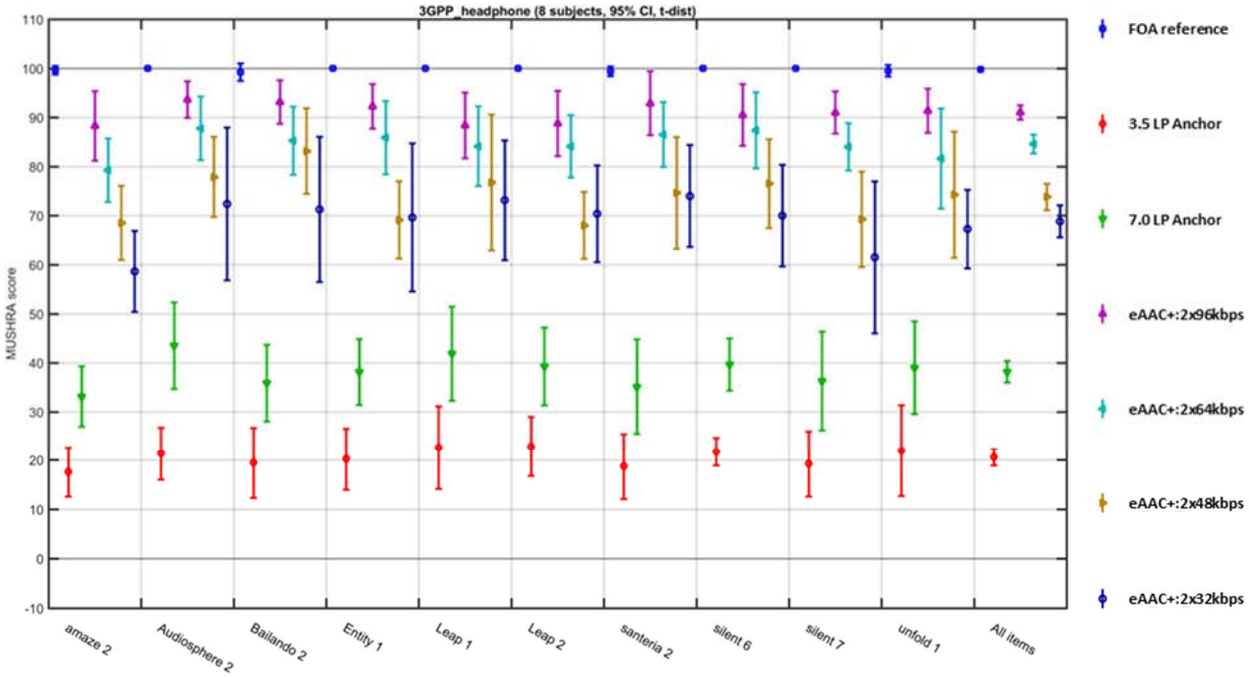


Figure 6.12: B-format extension to eAAC+, binaural rendering

Table 6.10: Binaural scores

Condition	Amaze2	Audiosphere 2	Bailando 2	Entity 1	leap 1	leap 2	santeria 2	silent 6	silent 7	unfold 1	all items
FOA reference	100	100	99	100	100	100	99	100	100	100	100
3.5 LP Anchor	18	21	20	20	23	23	19	22	19	22	21
7.0 LP Anchor	33	44	36	38	42	39	35	40	36	39	38
eAAC+:2x96kbps	88	94	93	92	88	89	93	91	91	91	91
eAAC+:2x64kbps	79	88	85	86	84	84	87	87	84	82	85
eAAC+:2x48kbps	69	78	83	69	77	68	75	77	69	74	74
eAAC+:2x32kbps	59	72	71	70	73	70	74	70	62	67	69

6.1.6.5 Conclusions

In this study, a workflow is tested in which a First-Order Ambisonics content stream is represented through 2 stereo eAAC+ streams. The workflow involves matrixing pre- and post-processing operations, from B to A-format and the reverse. Sending and receiving ends thus need to be aware of the chosen configuration with two eAAC+ instances and the specific conversions between A- and B-Formats. In the experiments, no precautions for potential overload after the format conversions were made.

The observed increase in quality over bitrate for the given test material is roughly linear for both endpoints, giving an impression of expected quality/bitrate trade-offs. Mean scores for the headphone endpoint are overall slightly lower than loudspeaker, but the shift is consistently very small (<5 MUSHRA points).

Together, these data suggest that dual stereo eAAC+ streams can reliably carry B-format audio derived from 7.1.4 channel audio, after conversion to A-format, with no unexpected encoding artifacts inherent to B-format, which would have affected the MUSHRA rating scores. However, a direct carriage of B-format was not formally evaluated and therefore no conclusions related to A-format vs. B-format can be drawn. As the study is limited to 3GPP eAAC+ no conclusions can be drawn on the suitability or necessary configurations of other 3GPP codecs for the coding of First-Order Ambisonics content.

6.1.7 Listening test for coding of First-Order Ambisonics using the EVS codec with loudspeaker rendering

6.1.7.1 Introduction

The experiment in clause 6.1.5 showed that the Ambisonics formats (1st to 3rd order) are performing better than what is achievable using existing 3GPP speech and audio services (mono or stereo) in terms of overall quality and spatial localization accuracy representing synthetic audio scenes.

In this experiment, the perceived overall and spatial quality of synthetic and recorded first-order Ambisonics (FOA) representations encoded by the EVS codec [68] were assessed. A comparison of encoding the A- or the B-format of the FOA representations with EVS was made as pre-screening for the main listening test. In addition, rendering of FOA was compared to stereo (two channel) rendering to relate the QoE to existing 3GPP speech and audio services for VR use cases.

6.1.7.2 Objectives

The main objective of the experiment was to assess the suitability of encoding FOA for VR use cases using existing 3GPP speech and audio codecs. The EVS codec was selected considering the need to cope well with speech and generic audio signals. As a speech and audio codec the EVS codec is suitable for conversational applications, requiring a low delay, as well as non-conversational use cases where higher quality is a key factor. The experiment covered several codec configurations to assess the quality obtained under certain bit rate constraints.

In addition, there was an objective to assess the suitability of FOA as a format for encoding a variety of audio material including recordings and synthetically generated audio scenes, e.g. by mapping a monophonic audio source to the B-format representation, especially in relation to what is achievable in existing 3GPP speech and audio services. The experiment aimed to cover a plurarity of auditory scenes to obtain more confidence in the generality of the results.

This experiment focused on the overall quality including additional evaluation of the spatial quality in terms of source localizability, width, height, depth, distance and spatial envelopment or immersiveness. Such spatial attributes are important factors to consider when assessing whether a particular Ambisonics representation is suitable for an immersive experience.

6.1.7.3 Test methodology

In this test, a method inspired by the ITU-R BS.1534-3 (MUSHRA) [22] test methodology was used. In contrast to an ordinary ITU-R BS.1534-3 test, there were two simultaneous ratings for each sound example, one for overall quality and one for spatial quality where the listeners were instructed to consider the source localizability, width, height, depth, distance and spatial envelopment or immersiveness. The use of more than one scale in the same trial have been tested successfully before in [65], and is also used in ITU-T P.806 [66] where 6 + 2 test questions are rated in the same trial.

6.1.7.4 Physical test setup

In order to avoid the effect of HRTF filtering the test was carried out with a 24-loudspeaker setup arranged in three circles, one at -30° elevation, one at 0° elevation and one at +30° elevation. The circles were offset 22.5° between the levels to distribute the speakers as evenly as possible. All speakers were positioned 2.0 m from the listening position as shown in Figure 6.13. The speakers chosen for this test, M-Audio AV-40, are compact 2-way speakers that, due to their size are limited in bass response but, on the other hand, provide a more defined acoustic center than bigger multi-driver speakers.

The room used for the test is a well sound proofed and acoustically treated audio lab, rated as NR10, with short reverberation time.

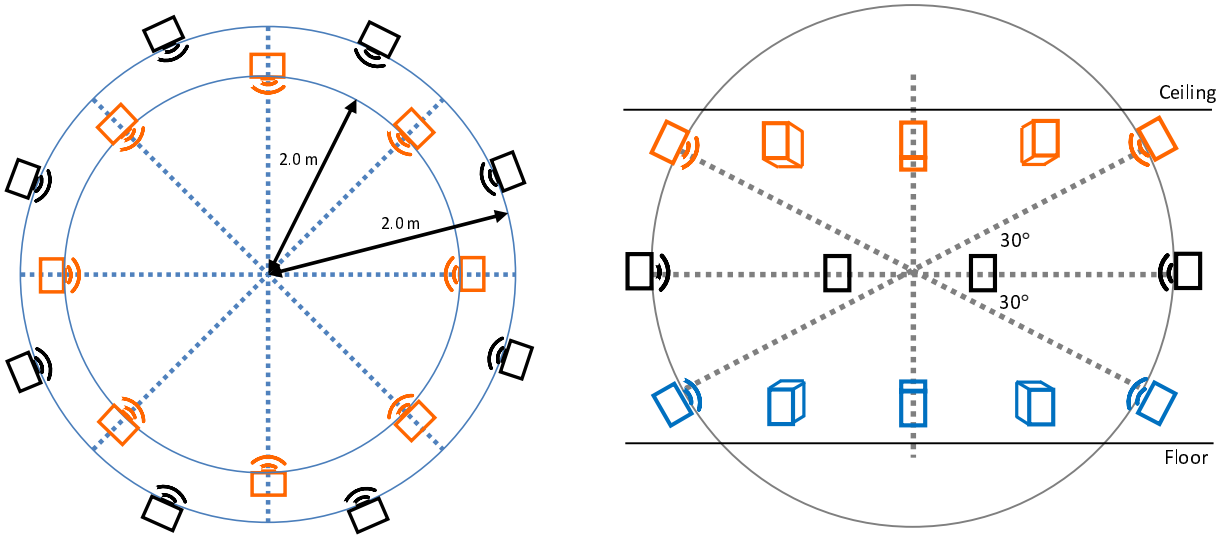


Figure 6.13: 24-loudspeaker setup. The blue speakers are placed at an elevation of -30 degrees, the black speakers at an elevation of 0 degrees and the orange speakers at an elevation of 30 degrees. All speakers are placed at a distance of 2.0 m from the center.

The test subjects were instructed to sit in the sweet spot of the loudspeaker array, while having the freedom to turn their head and move their upper torso slightly to better zoom in on a particular audio component of interest. The testing software provided a function to guide the user into the exact sweet spot. This was done by playing pink noise through all the speakers and finding the right position by minimizing the panning and the phasing effects that can be heard when moving out of the sweet spot.

6.1.7.5 Test material

The material included a variety of different audio scenes that are relevant for VR use cases. Ten different audio scenes were obtained and converted into a B-format representation. There were three types of scenes:

- Recorded scene
  - Recorded with a Sennheiser Ambeo microphone, i.e. a tetrahedral arrangement of four microphones (A-format), which were transformed to a B-format representation
- Synthetic scene
  - Virtual static and or moving audio sources encoded into a B-format representation
- Mixed scene
  - A mixture of a recorded scene and a synthetically encoded scene

Table 6.11 lists the scenes and gives a short decription of the content. Item11 was used in a training session and the results for that item are not presented as part of the test results.

Table 6.11: Audio scenes used in the test

Scene name	Scene description
Item1 (synthetic)	The sound of rattling screws in a glass, moving 360 degrees in azimuth, in an even counter clockwise circular movement, starting from straight ahead (0 degrees), at fixed elevation 30 degrees.
Item2 (recorded)	Two choirs singing in a church at the left and right sides of the microphone.
Item3 (recorded)	A choir singing and a chamber orchestra playing in a church at the front of the microphone.
Item4 (recorded)	A big band playing around the microphone. A trombone section in the front of the microphone, a solo saxophone to the left, and bass, guitar and drums at the back.
Item5 (recorded)	A subway train on a track above the microphone approaching the station from top left. Low level background with people talking at far distances and birds singing.
Item6 (recorded)	A shaken match box circulated around the microphone approximately fixed elevation of 0 degrees. Captured in a well sound proofed and acoustically treated audio lab, rated as NR10, with short reverberation time.
Item7 (recorded)	Two male talkers at fixed positions, approximately ± 45 degrees, conversating on sidewalk adjacent to a city street with moderate traffic. Woman with high heel shoes passing by on the left-hand side.
Item8 (recorded)	Two simultaneous conversations in a small conference room on the left respectively right hand-side of the microphone. The left-hand conversation between two male talkers at approximately azimuth 45 degrees (front left), and azimuth 120 degrees (back left). The righ-hand side conversation between one male talker at approximately azimuth -45 degrees (front right) and one female talker at approximately azimuth -120 degrees (right back).
Item9 (recorded)	Two male talkers conversating in a reverberant stair house. The first talker at a fixed position at approximately azimuth 45 degrees (front right). The second talker coming down the stairway from top left, stopping at a position of azimuth about -45 degrees (front left).
Item10 (mixed)	Two voices synthetically placed at fixed elevation of 0 degrees are heard from different azimuth angles. A female voice is heard from azimuth +30 degrees (front left) and a male voice from azimuth +150 (back left). Mixed at 15 dB SNR with recording of an outdoor bus station containing people talking at far distances and birds singing.
Item11 (recorded), only used for training session	A big band playing around the microphone. A trombone section in the front of the microphone, a saxophone section to the left, a trumpet section to the right and bass, guitar and drums at the back.

6.1.7.6 Test conditions

One important aspect of the assessment was to decide on which of the two Ambisonic representations of the audio scenes, the A- or the B-format, should be encoded. A short pre-test with the items 1, 2, 4, 8 and 9 and four listeners was performed with EVS encoded A- and B-format of the same audio scenes at 4x13.2 and 4x24.4 kbit/s. The input signals were sampled at 32 kHz and the codec was running in SWB mode. As for ordinary MUSHRA tests low-pass anchors at

3.5 and 7 kHz and a hidden reference were included. The conditions for the pre-test are presented in Table 6.12. The result showed that the audio scenes renderings from B-format encoded material obtained statistically significant better scores for overall and spatial quality, especially for the lowest bitrate, see figures 6.18 to 6.20 in Clause 6.1.7.9.

Table 6.12: Conditions evaluated in the pre-test

Short name	Description	Rendering details
ANCH-3k5	Low-pass filter anchor, cutoff = 3.5 kHz	Low-pass filter applied to each of the loudspeaker signals of the reference signal. Rendered via B-format to 24 loudspeaker signals.
ANCH-7k	Low-pass filter anchor, cutoff = 7 kHz	Low-pass filter applied to each of the loudspeaker signals of the reference signal. Rendered via B-format to 24 loudspeaker signals.
EVS-4x13k2-A	A-format FOA, EVS @ 4x13.2 kbit/s	Rendered via B-format to 24 loudspeaker signals.
EVS-4x13k2-B	B-format FOA, EVS @ 4x13.2 kbit/s	Rendered via B-format to 24 loudspeaker signals.
EVS-4x24k4-A	A-format FOA, EVS @ 4x24.4 kbit/s	Rendered via B-format to 24 loudspeaker signals.
EVS-4x24k4-B	B-format FOA, EVS @ 4x24.4 kbit/s	Rendered via B-format to 24 loudspeaker signals.
REF	Hidden reference, Uncoded FOA	Rendered via B-format to 24 loudspeaker signals.

For the main test the EVS codec was used to encode the four B-format components with an even bit distribution of 4x13.2, 4x24.4, 4x48 and 4x96 kbit/s. The input signals were sampled at 32 kHz and the codec was running in SWB mode. In addition, there was a stereo condition where the B-format was decoded onto two of the loudspeakers at elevation 0 degrees and azimuth angles ± 67.5 degrees. The test further included low-pass anchors at 3.5 and 7 kHz and a hidden reference. The conditions for the main test are presented in Table 6.13.

Table 6.13: Conditions evaluated in the main test

Short name	Description	Rendering details
ANCH-3k5	Low-pass filter anchor, cutoff = 3.5 kHz	Low-pass filter applied to each of the loudspeaker signals of the reference signal. Rendered via B-format to 24 loudspeaker signals.
ANCH-7k	Low-pass filter anchor, cutoff = 7 kHz	Low-pass filter applied to each of the loudspeaker signals of the reference signal. Rendered via B-format to 24 loudspeaker signals.
STEREO	Uncoded FOA rendered to stereo	Rendered via B-format to two loudspeakers at 0 degree elevation at ±67.5 degrees azimuth.
EVS-4x13k2-B	B-format FOA encoded by EVS @ 4x13.2 kbit/s	Rendered via B-format to 24 loudspeaker signals.
EVS-4x24k4-B	B-format FOA encoded by EVS @ 4x24.4 kbit/s	Rendered via B-format to 24 loudspeaker signals.
EVS-4x48k-B	B-format FOA encoded by EVS @ 4x48 kbit/s	Rendered via B-format to 24 loudspeaker signals.
EVS-4x96k-B	B-format FOA encoded by EVS @ 4x96 kbit/s	Rendered via B-format to 24 loudspeaker signals.
REF	Hidden reference, Uncoded FOA	Rendered via B-format to 24 loudspeaker signals.

The generation of the FOA B-format representation for the recorded audio scenes entailed two processing steps. First ambisonic decoding of the A-format microphone signals into the B-format, and then microphone equalization of the B-format signals as illustrated in Figure 6.14. The ambisonic decoding of the recorded A-format signal to a FOA signal was done by multiplying the A-format signal with the decoding matrix D, obtained as a pseudo-inverse from the spherical harmonic transform matrix Y that maps the microphone orientation angles to the FOA beams. The microphone equalization filters were obtained from the A-to-B-converter VST-plugin for the Sennheiser Ambeo microphone, which is available on Sennheiser's home page.

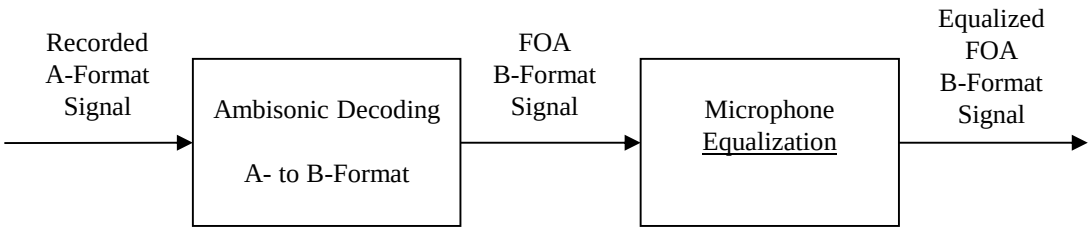


Figure 6.14: Generation of FOA B-format representation from a recorded A-format representation.

The generation of the FOA B-format representation for synthetic audio scenes simply entailed the ambisonic encoding of each virtual audio point source into a B-format signal as illustrated in Figure 6.15. The ambisonic encoding of the virtual sources to a FOA signal was done by multiplying each sample of a source signal with the spherical harmonic transform vector Y_e mapping the current position of the source to the FOA beams. For moving sources the Y_e matrix was updated at every sample.

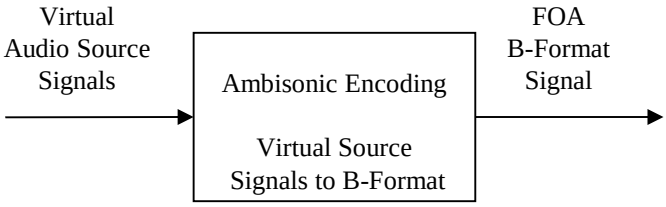


Figure 6.15: Generation of FOA B-format representation for a synthetic audio scene made up of different virtual audio sources.

The audio encoding of the B-format signals involved the consecutive steps of 50 Hz high-pass filtering, RMS level adjustment (-32 dBov) and 4xEVS audio encoding. The same high-pass filtering and level adjustment was applied to the non-coded conditions.

The mapping of the B-format signals to the loudspeaker signals entailed a standard ambisonic decoding of the B-format signals to the loudspeaker signals followed by a level adjustment as illustrated in Figure 6.16.

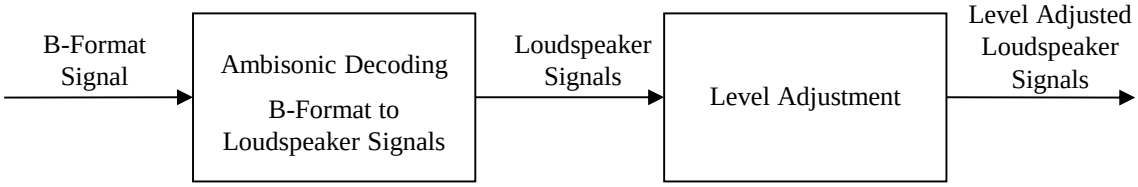


Figure 6.16: Mapping FOA B-format signals to loudspeaker signals.

The ambisonic decoding of the FOA signals to loudspeaker signals was done by first evaluating the spherical harmonic transform matrix Y_d that maps the loudspeaker positions to the FOA beams, then evaluating the decoding matix D as the pseudo inverse of Y_d , and then multiplying each FOA sample vector with decoding matrix D. The 24 loudspeaker signals were jointly normalized to an RMS level of -39.8 dBov, while the stereo channels were normalized to an RMS level of -29 dBov to be perceived similarily loud in average.

The ACN ambisonics channel order and the SN3D spherical harmonic normalization were used.

During the test the subjects were able to adjust the playback volume in a range of +/-4 dB, but were instructed not to change this setting while comparing the test samples.

6.1.7.7 Listening panel

The listening panel consisted of 8 experienced listeners of the Audio technology section at Ericsson Research. The main test duration was approximately 1.5 hours including the training session. Post-screening of the test subjects was made according to ITU-R BS.1534-3 [22] and none of the test subjects were rejected.

6.1.7.8 Software

The user interface used during the test was based on a typical MUSHRA test interface, with the addition of a second rating scale, see Figure 6.17.

Figure 6.17: Software GUI used in the test

6.1.7.9 Test results for A- and B-format pre-test

The test results for the overall quality of the pre-test comparing coding of A-format signals and B-format signals are shown in Figure 6.18-6.20. The difference overall scores of the B-format obtained a statistically significant better performance than the A-format at 4x13.2 and 4x24.4 kbit/s. Similar observations were made for the spatial quality.

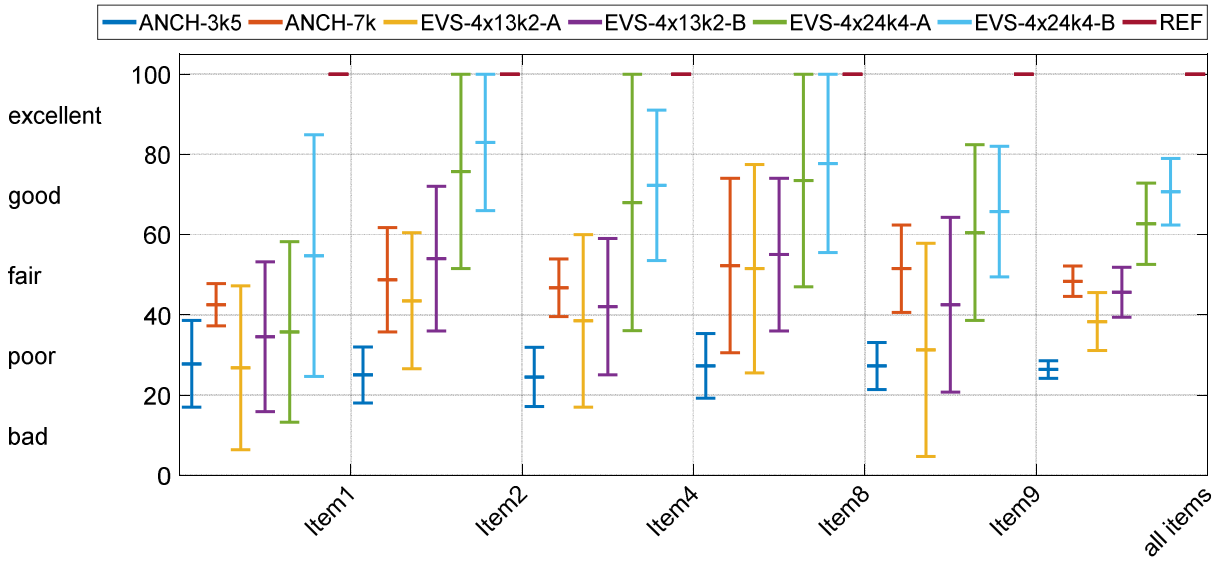


Figure 6.18: Absolute overall scores, with 95% CI.

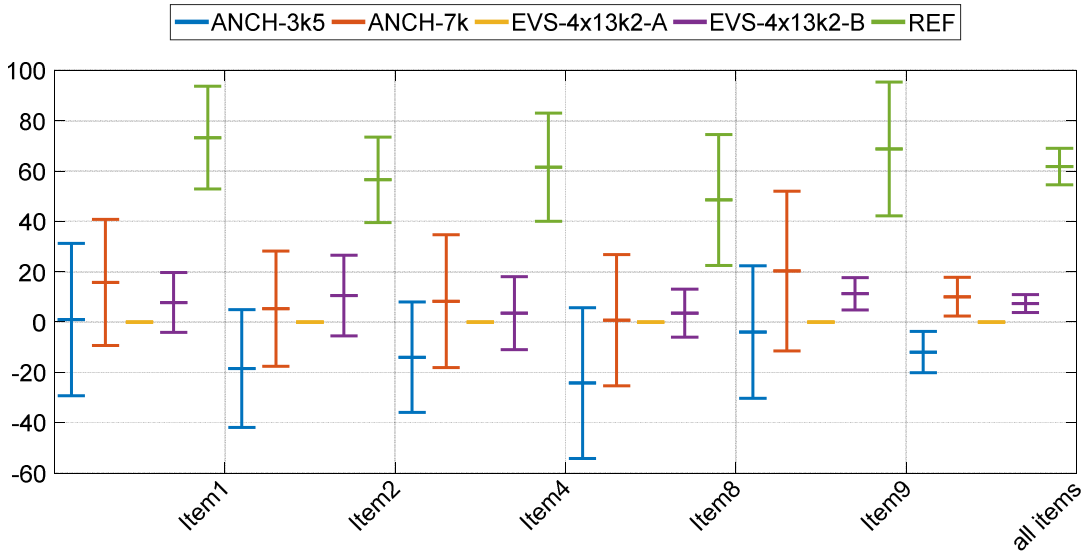


Figure 6.19: Difference overall scores for 4x13.2 kbit/s, relative to A-format 4x13.2 kbit/s, with 95% CI.

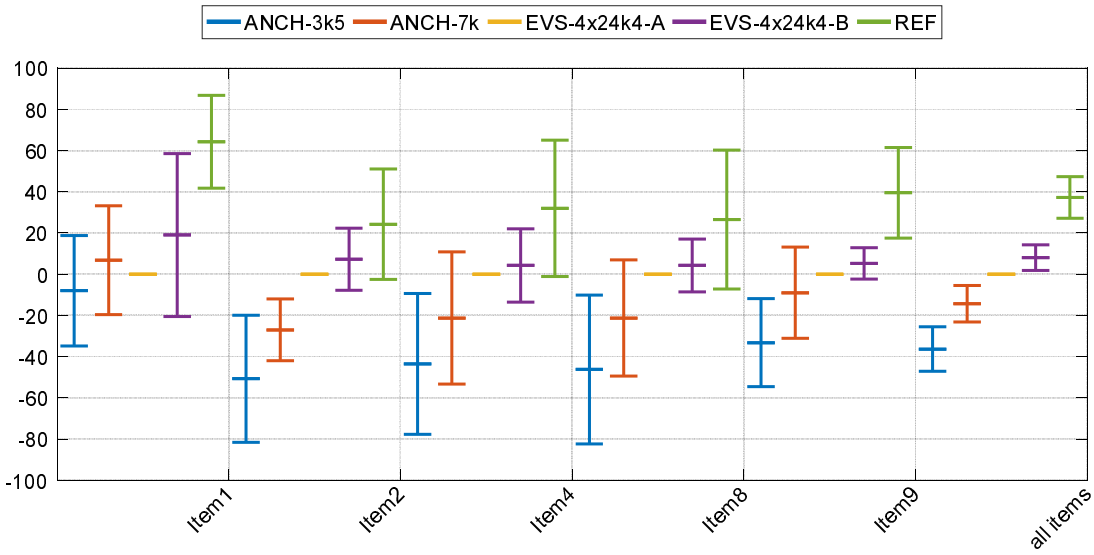


Figure 6.20: Difference overall scores for 4x24.4 kbit/s, relative to A-format 4x24.4 kbit/s, with 95% CI.

6.1.7.10 Test results for main test

The test results of the main test are shown in Figure 6.21-6.24.

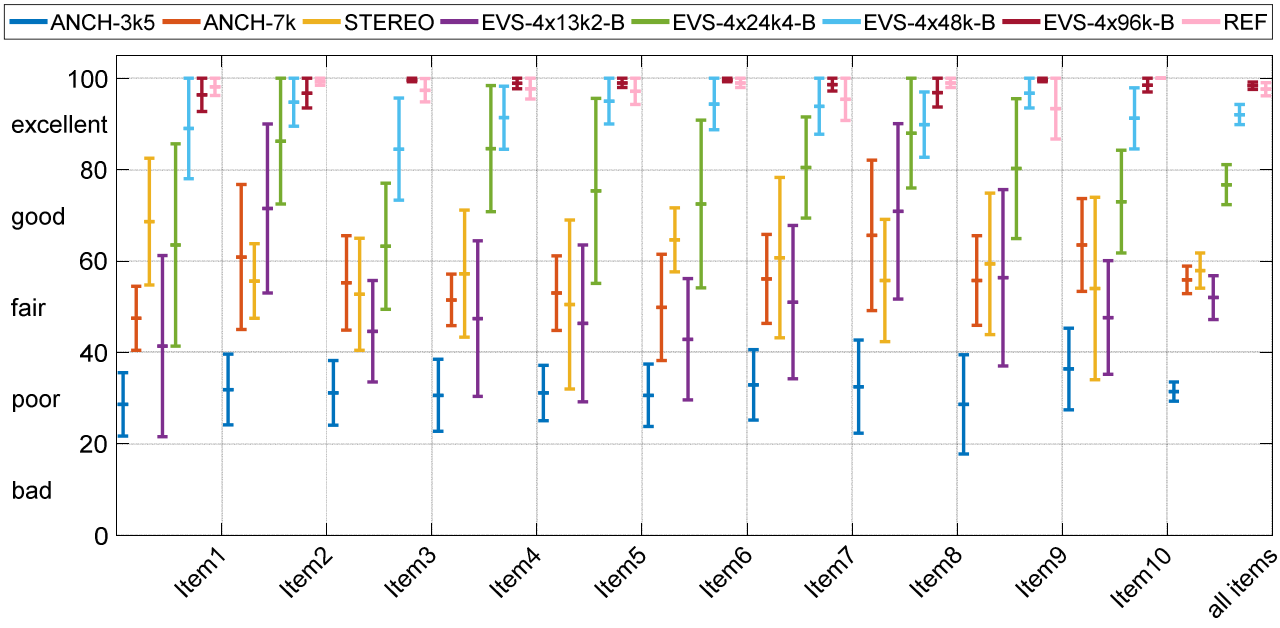


Figure 6.21: Absolute overall scores, with 95% CI.

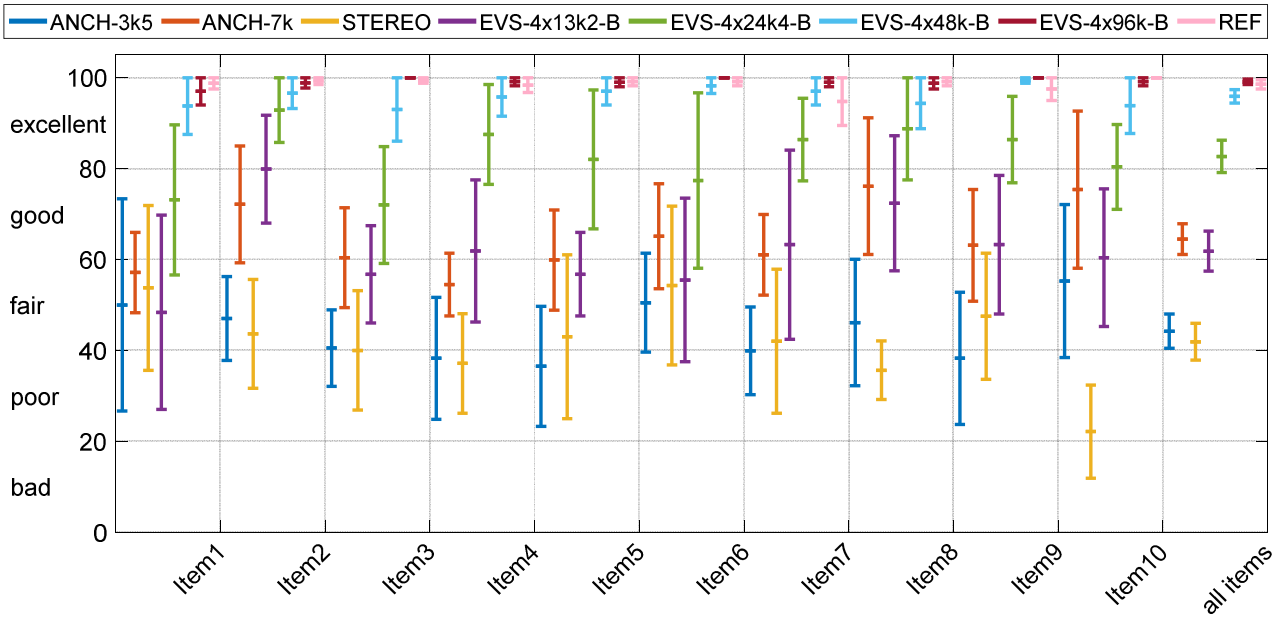


Figure 6.22: Absolute spatial scores, with 95% CI.

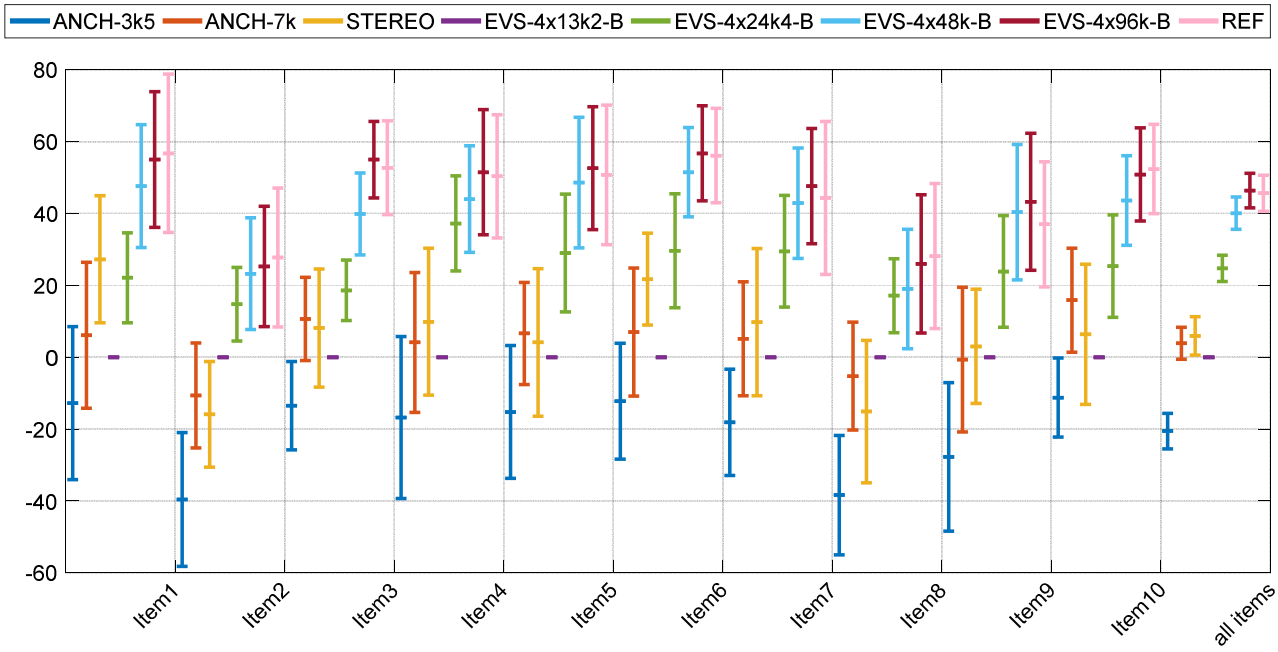


Figure 6.23: Differential overall scores, relative to 4x13.2 kbit/s, with 95% CI.

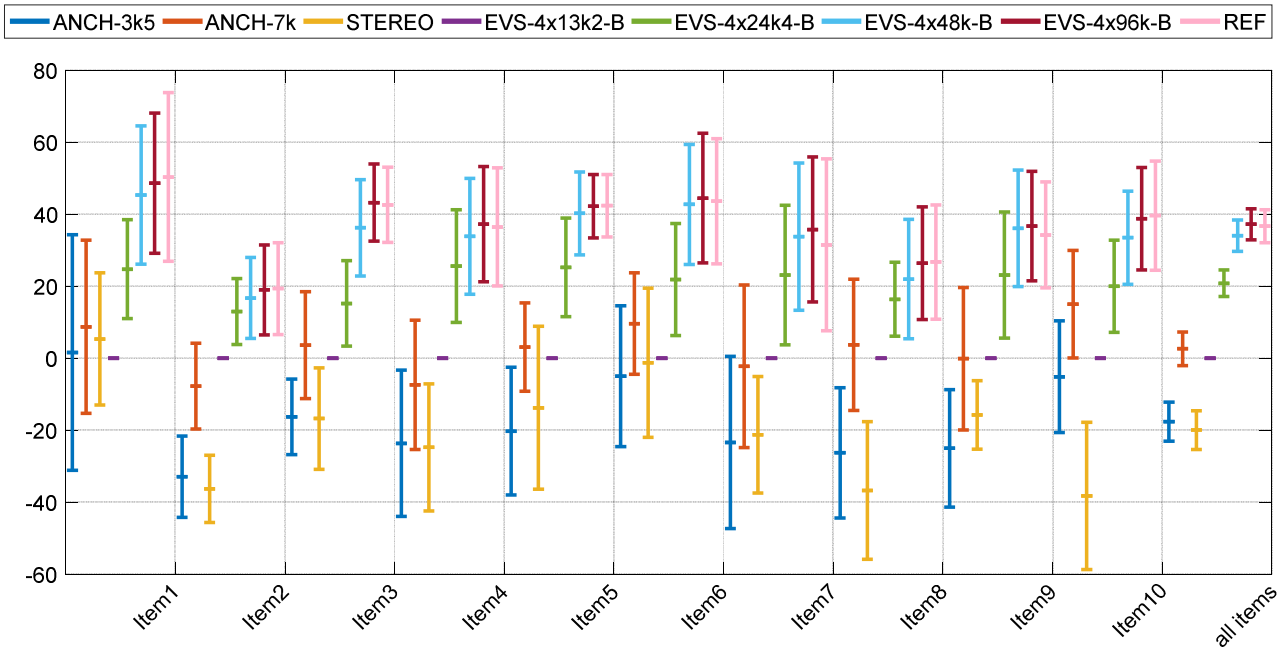


Figure 6.24: Differential spatial scores, relative to 4x13.2 kbit/s, with 95% CI.

6.1.7.11 Complexity

As reported in the performance characterization of the EVS codec [69], the combined encoder and decoder worst-case complexity is 87.97 WMOPS. Running four parallel EVS encoders and decoders would consequently in the worst-case consume 351.88 WMOPS. It should be noted that the in [69] reported operation modes consuming the worst-case complexity are not among the tested conditions which means this is a conservative complexity estimate. Additionally, considering decoder only, the worst-case complexity is reported to be 31.72 WMOPS, which means 126.88 WMOPS would be consumed for decoder-only use cases.

Similarly, the EVS codec (including both encoder and decoder) is reported to use 149 kW (16-bits) of RAM, 147 kW of ROM, and 114500 program instructions. When running four simultaneous instances of the EVS codecs it is assumed optimizations can be done such that ROM and program instructions can be shared, which means only the RAM consumption would increase by the number of codec instances, i.e. 596 kW RAM assuming no optimizations.

The EVS codec operates on 20 ms frames with an algorithmic delay of less than or equal to 32 ms [69].

6.1.7.12 Conclusions

The results from the listening test show that coding of super-wideband FOA signals without a statistically significant degradation is achievable using the EVS codec at 4x96 kbit/s. In addition, an overall and spatial quality within the ‘Excellent’ region was reached at 4x48 kbit/s. For 4x24.4 kbit/s an overall quality in the upper end of the ‘Good’ region and a spatial quality in the ‘Excellent’ region were observed. The overall quality dropped significantly to ‘Fair’ when using 4x13.2kbit/s.

Generally, a high spatial quality was observed for the EVS encoded conditions, which indicates that running four independent codec instances encoding the FOA B-format does not introduce severe spatial problems. The pre-test showed a better performance when encoding the B-format, but no further evaluation of encoding A-format signals with the EVS codec was done.

As the uncoded FOA condition was given as an explicit reference it was assumed that this representation was a preferable representation of the audio scenes. It was therefore also expected that the stereo condition, utilizing only two channels for rendering the audio scenes, would correspond to a quality less or equal to the FOA reference quality, which is also inline with the observations in Clause 6.1.5. It was of specific interest to find out whether the coded FOA representations would be perceived having better or worse quality than the uncoded stereo condition having a limited

capability in its spatial representation. It is clear that FOA provides a symmetric representation of the sound scene allowing for head rotations (3 DoF), which is of specific importance for VR use cases, that is not achievable by rendering stereo signals.

As seen from the results, all EVS FOA conditions except FOA at 4x13.2kbit/s provided a significantly better overall quality than the stereo condition and in terms of spatial quality the stereo condition performed significantly worse than the FOA conditions. It can be observed that the spatial quality for the stereo condition is performing relatively worse for items comprising complex audio scenes with surrounding sounds than e.g. for Item1 and Item6, comprising a synthetic and recorded scene of narrow (point-like) sound sources. This shows that even with uncoded stereo channels, the codec distortions introduced by the EVS codec for FOA at 4x24.4 kbit/s and above were perceived as less of degradation than what the stereo rendering was. However, as the FOA condition was given as the explicit reference, the stereo condition quality scores might have been negatively affected which affects the relative comparison.

The tests were performed using loudspeaker rendering to avoid effects of non-personalized HRTFs for binaural rendering. It is assumed that a good binauralization can be achieved by rendering of virtual loudspeakers, but it should be noted that the potential effects on the binaural rendering, e.g. the amount of externalization, coming from codec distortions were not assessed in this experiment.

The expected worst-case complexity for running four parallel EVS codecs (encoders and decoders) is 351.88 WMOPS based on the complexity reported in the performance characterization report for the EVS codec [69], while it would be 126.88 WMOPS for decoder-only use cases. Additionally, a combined encoder and decoder memory usage of 596 kW (16-bit) for RAM can be expected without specific optimizations, while ROM and program instructions should possibly be shared implying a ROM consumption of 147 kW and 114500 program instructions.

6.2 Audio quality evaluation of object-based formats

6.2.1 Introduction

Immersive audio experiences (3D audio) are an important element of next-generation audio entertainment systems. This section describes the evaluation of object-based audio formats.

6.2.2 Test of the AC-4 object-based coding scheme

A test has been conducted to analyse the coding efficiency of object-based audio coding with AC-4 as defined in ETSI TS 103190-2 [45]. A report describing a test system, the content under test, and the results of the experiment has been published by means of the Audio Engineering Society Convention Paper #9587, "Immersive Audio Delivery Using Joint Object Coding", presented at the 140th AES Convention 2016 June 4-7, Paris, France [64].

A 7.1.4 immersive speaker-based rendering system was chosen for evaluation of the object-based coding efficiency of the AC-4 system. Using this rendering scheme avoids the challenges with subjective or objective assessment of HRTF rendering.

This methodology has limitations related to the assessment of the rendering to VR use cases. For instance, the results do not include any potentially related quality degradation imposed by HRTF rendering. The choice for this rendering system enables the assessment of the coding efficiency of the object based coding scheme.

NOTE: AC-4 is designed for unidirectional media delivery and has not been evaluated in any other context, and is thus expected to be unsuitable for conversational applications.

6.3 Audio quality evaluation of channel-based formats

6.3.1 Introduction

Immersive audio experiences (3D audio) are an important element of next-generation audio entertainment systems. This section describes the evaluation of channel-based audio formats.

6.3.2 Test of the ISO/IEC 23008-3 MPEG-H 3D Audio channel-based coding scheme

The MPEG-H verification test report [36] provides details on four large-scale listening tests that were conducted to assess the performance of the Low Complexity (LC) Profile of MPEG-H 3D Audio. Test material was either channel-

based, channel plus objects, or scene-based, as Higher Order Ambisonics (HOA) of a designated order, possibly also including objects. Three tests were conducted over loudspeakers and one test over headphones binaurally rendered.

7 Video quality evaluation

7.1 Similarity ring metric

7.1.1 Challenges for Subjective Assessment

The main challenges for subjective assessment of omnidirectional video can be summarized as follows:

1) *Spatial correlation*. Ensure that the same parts of the omnidirectional content are watched when running several test sessions by one or more test subjects. Given the limited Field of View (FoV) of the human visual system and the HMDs, less than 360-degree content can be displayed and watched at any point in time. So, when moving through the omnidirectional space, a test subject may watch different parts of the omnidirectional video compared to parts watched for another test case. The watched parts in different tests may not overlap at all. Consequently, the scoring of the same video clip quality watched multiple times by a test subject or by multiple test subjects may be related to watching totally different parts of the video clip. This could make the test results invalid.

2) *Temporal correlation*. Ensure that 1. holds for each time instant of the assessed video clips. In fact, 1. is not sufficient to guarantee that test subjects are really watching the same thing. In fact, despite different people may be watching the same parts of a video clip, they might be watching it at different time instants T1 and T2. If video impairments are visible only at time T1, but not at time T2, the subject watching the same part of the content at time T2 will not perceive the impairments as the first subject does. Consequently, two subjects may be scoring the video quality differently, and the results might be inaccurate or even invalid.

3) *Relative Motion freedom*. Ensure that the test subjects have enough freedom in the motion when wearing the HMD. A 360-degree experience is such that a test subject should be free of exploring the content in all possible directions (possibly with a wireless HMD). However, a total motion freedom could make the subjective testing activity harder. A solution could be introducing a soft motion constraint for the test subjects. For example, to follow a specific motion pattern during the test sessions (see Figure 7.1). The given pattern should be content-dependent.

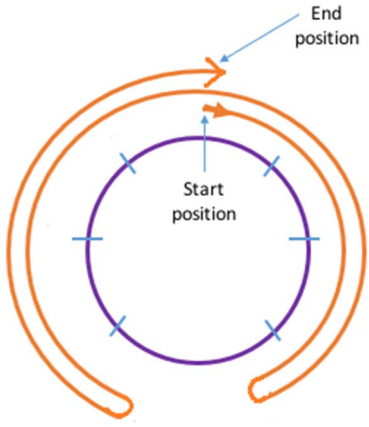


Figure 7.1: Test subjects motion pattern for 360 degree subjective assessment (aerial view)

Clauses 7.1.3 and 7.1.4 introduce a way to measure how far apart are two (or more) watching patterns (for the same test subject or different test subjects) for the same video clip for each time instant. With the possible aid of relative motion freedom (as described above) this metric can guarantee that, not only within a single subject, but also across multiple test subjects, the same viewing experience can be replicated, it follows a natural pattern, and can be objectively measured.

7.1.2 Void

7.1.3 HMD Tracking

A practical implementation of the metric in Clause 7.1.4 requires the recording of HMD motion information. For instance, the data to be stored during a subjective video quality test session, for each subject and for each test clip, is the following:

- Timestamp T.
- Yaw, Pitch and Roll denoting the head orientation at time T.
- Additionally, the viewport (or tile) watched at time T.

Data can be sampled at a given sufficient rate (e.g. 10 samples per second). The output is a data stream of timed HMD orientations and viewports/tiles watched by test subjects for each test clip.

7.1.4 Similarity Ring Metric (SRM)

7.1.4.1 Introduction

Once the HMD data is available as described in Clause 7.1.3, there needs to be a way to measure:

- Whether a subject has been watching the same parts (at the same time) of all the video clips belonging to the same test clip (*within-subject correlation*).
- Whether multiple subjects have been watching the same parts (at the same time) of the same video clip (*between-subjects correlation*).

For simplicity, the remainder of the discussion will focus on Yaw, which measures the horizontal FoV. The concepts can be extended to include a multidimensional assessment with Pitch and Roll and the processing of viewport/tile data.

A typical plot of Time vs. Yaw could look like illustrated in Figure 7.2. Here the curves represents the watching patterns of one (or more) subjects when watching the same test clip in different test sessions. It can be clearly seen that the curves are just partially overlapping, or close to each other by a certain distance. It is not possible that all curves overlap at the same time, since each test case carries some elements of variability even within the same subject. These elements are direction of motion and speed of motion.

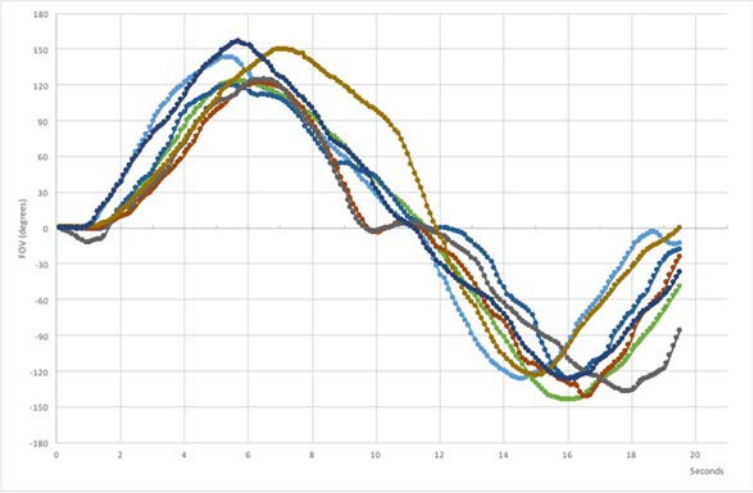


Figure 7.2: Head motion pattern example

However, it is possible to determine if the aggregate set of curves falls within a certain range. This range may be visualized as a "*ring*". The goal is then to ensure that the ring can travel through all curves from the beginning to the end of a test clip. If this happens, it can be stated that all clips (i.e. the curves) have been watched with high similarity.

If the curves are related to clips watched by the same test subject, high similarity means that the same subject has been watching the same content at the same time. If the curves are related to different test subjects, high similarity means then that the different test subjects have been watching the same content at the same time.

The ring size is determined by the HMD FoV, or otherwise can be the content FoV (e.g. primary viewports or tiles FoV). An example of a ring using HEVC tiles is shown in Figure 7.3. In the following the ring size is indicated simply as *FoV*.

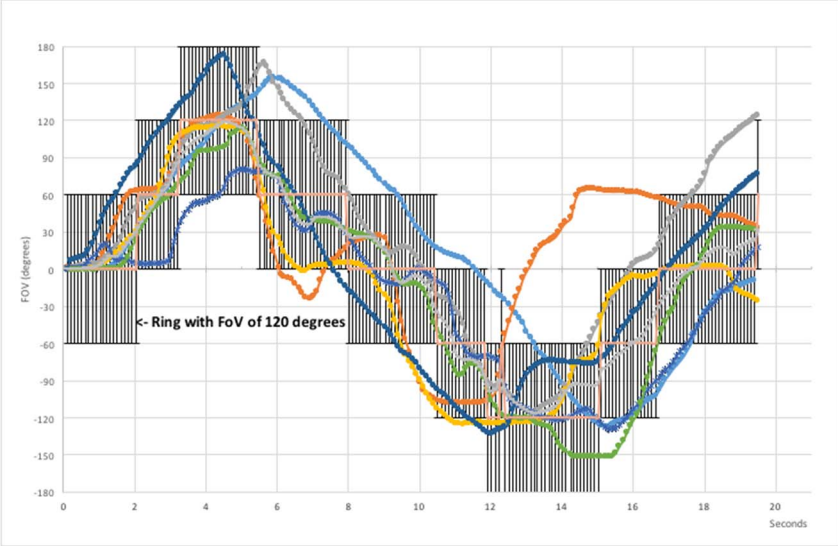


Figure 7.3: Example of ring with 120 degrees tiles

7.1.4.2 Calculating the SRM

Here follows a method to calculate the *Similarity Ring Metric (SRM)* of a given set of clips for one or more subjects. In other words, it is a metric to measure if different subjects have been watching the same parts of a test clip at the same time.

For each time instance, the center of the ring represents the ideal point where most of the people should watch. The ring has always a range (-FoV/2, FoV/2). So, if the FoV is 120 degrees, the ring has a range (-60, 60) degrees.

For the time instance $i=1,..,T$, when considering the set of Yaw values (Y_i^k) for the clip $k=1,\ldots,N$, it can be defined the point where the majority of the N clips have been watched as

$$\text{Mode}_i(Y_i^1, Y_i^2, \ldots, Y_i^N).$$

As alternative to the statistical mode, the median could be considered. Also, in case of multiple modes, the mode closest to the median could be chosen. The mode value defined above corresponds therefore to the center of the ring:

$$\text{Ring}_i = \text{Mode}_i(Y_i^1, Y_i^2, \ldots, Y_i^N).$$

$Ring_i$ has a range of (Ring_i - FoV/2, Ring_i + FoV/2). A Yaw value outside this range corresponds to a test subject not watching the same content as other test subjects, for the same time instant i.

It can be defined the *instantaneous similarity (IS)* at time i for clip k as:

$$IS_i^k = \begin{cases} 0 \; if \; Y_i^k \in \left(Ring_i - \frac{FoV}{2}, Ring_i + \frac{FoV}{2}\right), \\ 1 \; otherwise. \end{cases}$$

In other words, the value 0 is assigned to the clip k at time i if and only if it has been watched within the ring range (i.e. watching the same thing as most of the other test subjects). Otherwise the value 1 is assigned, meaning that what has been watched was not the same as the other test subjects.

The average of the instantaneous values at time i for N watched clips is given by the *Negative Similarity Ring Metric (NRSM)*:

$$NSRM_i = \sum_{k=1}^{N} \frac{IS_i^k}{N}$$

which is a number in the range [0, 1], where 1 indicates no watching similarity across the *N* watched clips, and 0 indicates perfect similarity across the watched clips.

The *Total NSRM (TNSRM)* can be calculated as the average for all T time instances (i.e. for the whole duration of the clip) as:

$$TNSRM = \sum_{i=1}^{T} \frac{NSRM_i}{T}$$

which yields again a number in the range [0, 1], with the same meaning as above.

Finally, the *Similarity Ring Metric (SRM)* can be calculated as follows:

SRM = (1-TNSRM) * 100,

where SRM is in the range [0, 100]%, and it expresses the *degree of watching similarity of all clips watched by either the same test subject or across several test subjects*, depending on the input data. A value of 0 % indicates no watching similarity, and a value of 100 % indicates perfect similarity.

7.1.4.3 Rejection criteria

As it is difficult to achieve a SRM=100 %, it is convenient to define a *Similarity Threshold ST*, e.g. 80 %. In this way, a rejection criteria could be established for the subjective tests: for example, the results of the tests could be rejected if SRM < ST.

7.1.4.4 Other usages of the SRM

The SRM may find applicability also for other related topics. For example:

- Categorization of different groups of test subjects, whenever they watch different parts of the omnidirectional video.
- Assess the motion-to-photon delay in case of head motion.

7.2 Subjective evaluation of Viewport-independent omnidirectional video streaming

7.2.1 Test description

7.2.1.1 Main objectives

In the context of many standardisation activities focusing on the short-term developments of 360-degrees video experience (3 degrees of freedom, aka 3DoF), the need has been raised to identify the impacts of video parameters on the perceived quality of experience as well as the associated bitrate needs.

Amongst such video parameters, the spatial resolution is identified as the most impacting one. Any change of resolution is likely to alter the rendering quality due to the lenses-based rendering system which basic principles are well defined in the 3GPP Technical report on Virtual Reality (see Clause 4.1).

> **First objective:** *Starting from the 8K (8192x4096) resolution of an equirectangular projected (ERP) video, assess the perceived video quality when decreasing the spatial resolution in a HMD environment.*

Once the video representation format is defined (resolution, frame rate…), the video compression is considered as the next challenging step for which the selected bitrate influences directly the quality of experience. The selection of the video format is dependent from two factors: Device decoding capabilities and test sequences availability in the chosen format.

> **Second objective:** *Starting from a 4K (4096x2048), 30 fps, ERP video, assess the perceived video quality when compressing the sequences with a state of the art video codec in a HMD environment.*

7.2.1.2 Test material

7.2.1.2.1 Head-Mounted Display (HMD)

A tethered HMD is used so as to be able to easily develop the software for playing and rating the test sequences. A tethered VR system connected to a PC has also the advantage to enable the playback of uncompressed sequences. The HMD in use has the following characteristics as described in Table 7.1.

Table 7.1: HMD features for video subjective tests

Display	OLED
Resolution	2160 x 1200
Refresh Rate	90Hz
Field of view	110 degrees
Sensors	Accelerometer, gyroscope, magnetometer, Constellation tracking camera

The system is used to render ERP (Equi-Rectangular Projection) contents in YUV/4:2:0/8bit up to 8kp30. The video sequences are rendered at the native system frame rate of 90Hz.

7.2.1.2.2 Software

The main challenge of conducting subjective tests with an HMD is to provide easy controls to the viewer for the playback and rating of the sequences and their respective versions. Two interfaces are defined accordingly:

- A graphic interface is developed for the selection of each sequences and the access to the rating, such as illustrated in Figure 7.4.

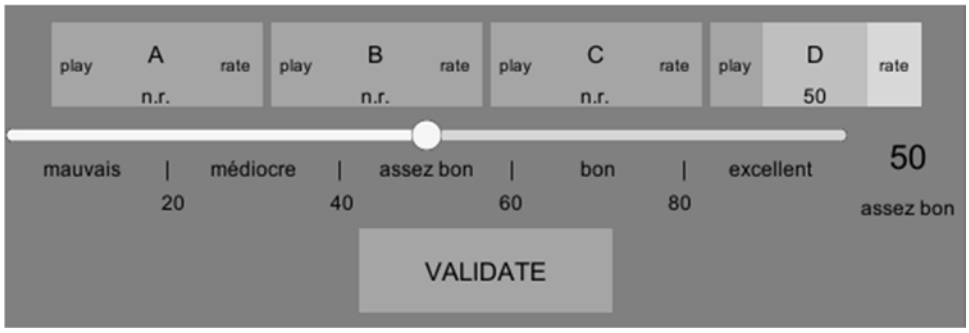


Figure 7.4: Graphic interface for sequence selection and rating

- A PC mouse is used as a controller because it is a simple (and well known) device to use that doesn't require the viewer to look at it. The mouse is configured, as illustrated in Figure 7.5, as follow:
  - Mouse wheel for navigation.
  - Mouse left simple click for selection/validation.

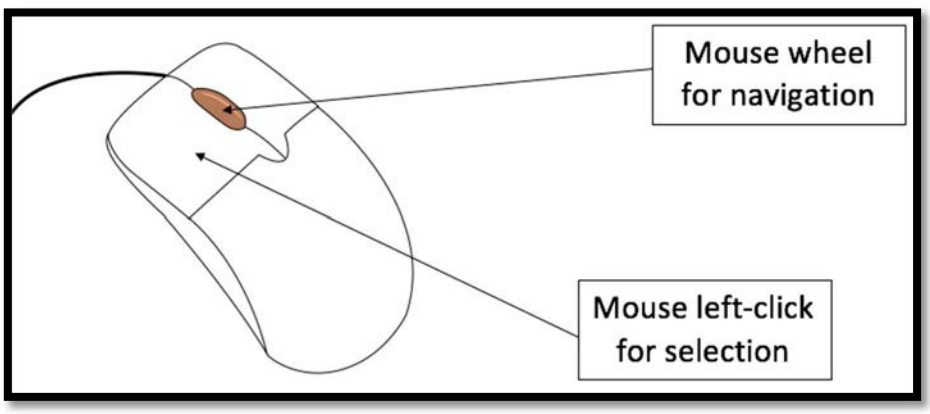


Figure 7.5: graphic interface for sequence selection and rating

7.2.1.3 Software implementation on content rendering

For 8K sub-resolutions, on the player side, the viewport projection being rectilinear, the viewport 3D coordinates are mapped on the sphere. These coordinates are then mapped to the ERP projection. Finally, the pixel values are interpolated using a Lanczos3 resizing filter.

On the frame rate aspects, the 30fps sequences are up-converted to the HMD native rendering frame rate 90ps by frame repetition.

7.2.1.4 Content preparation

For resolution test

Each test sequence is available in full equirectangular projection (ERP) representation, in 8K (8192x4096) resolution at 30 frames per second.

The following resolutions (all in 2:1 aspect ratio) are generated using a Lanczos3 resizing filter:

- 6K: 6 144 x 3 072
- 4K: 4 096 x 2 048
- 3K: 3 077 x 1 536
- 2K: 2 048 x 1 024
- 1K: 1 024 x 512

For compression test

Each test sequence is available in full ERP representation, in 4K (4096x2048) resolution at 30 frames per second. Sequences originally available were downsized using a Lanczos3 filter.

NOTE: The 30 fps frame rate is chosen based on the availability of test contents, due to the lack of relevant 60 fps sequences that can be used for the purpose of the present test (reproducibility requirement).

The HEVC video codec is selected as the state-of-the-art compression solution. The HM reference encoder in version **16.15** is configured with the following parameters for the generation of the Bitstreams:

- Profile: Main
- Bit depth: 8bit
- RAP period: 1s (32 images at 30fps)

The detailed configuration parameters are described in Annex A.

7.2.1.5 Test Sequences

The test on spatial resolution contains 5 sequences presented below in Figure 7.6.

ChairLiftRide (Courtesy of GoPro)

Gaslamp (Courtesy of InterDigital)

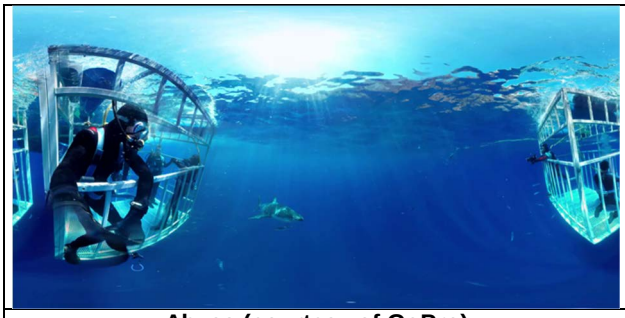
Harbor (Courtesy of InterDigital)

KiteFlite (Courtesy of InterDigital)

Trolley (Courtesy of InterDigital)

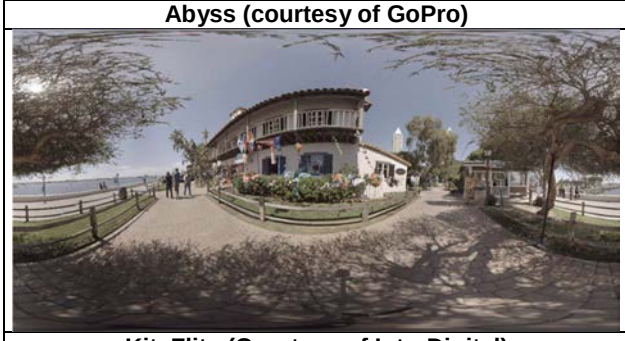
Figure 7.6: Source sequences for the resolution test

The test on compression also contains 5 sequences presented in Figure 7.7.



Abyss (courtesy of GoPro)

Harbour (Courtesy of InterDigital)



KiteFlite (Courtesy of InterDigital)

Skateboard (Courtesy of GoPro)

Trolley (Courtesy of InterDigital)

Figure 7.7: Source sequences for the compression test

7.2.1.6 Test methodology

7.2.1.6.1 Introduction

Without any existing standardized approach for subjective quality assessment in immersive environments, the decision is taken to start from the SAMVIQ (Subjective Assessment Methodology for Video Quality).

7.2.1.6.2 SAMVIQ based approach

In the SAMVIQ methodology (Recommendation ITU-R BT.1788 [60]), each viewer has access to all sequences and the explicit reference (if it exists) at any time and in any order the viewer wants. The viewer can watch each sequence as many times as he/she wants. This methodology improves the discrimination between the different cases that have to be scored. The scoring is done using a continuous quality scale graded from 0 to 100 and annotated with 5 quality labels: Excellent, Good, Fair, Poor and Bad (see Figure 7.4 in French language). The requirements for the completion of a test for a given sequence are: A sequence cannot be rated

7.2.1.6.3 Specific adaptations

In order to limit the differences of viewing conditions over the participants, it is decided to put some additional constraints on the viewing conditions:

- Some sequences are rotated so as to ensure that the viewer has a critical part of the 360-degrees scene to assess. Many 360-degrees sequences have a limited area of interest (in terms of objects movements and texture). The scene rotation then ensure that this specific part is displayed into the default viewport.

- For the same reason, only 3/4th of the 360-degrees is rendered. The remaining 1/4th centered on the back of the viewer is replaced by a black area.
- In this context, each viewer is seated on a fixed chair so as to be presented the same default viewport while seated in front of the table. The viewing environment is illustrated in Figure 7.9.

7.2.1.7 Test organization

7.2.1.7.1 Introduction

The test is split into 3 steps:

1) Each viewer passes a visual acuity test checking their vision (used as a rejection criteria when 10/10 vision at least on both eyes is not achieved).
2) A training test is conducted with assistance on explaining how the system works.
3) The main test is run by the viewer in full autonomy.

7.2.1.7.2 Visual acuity check

A general test on visual acuity is conducted as a first step. The long-distance vision is used as a rejection criteria because even if the actual distance from the screen to eyes is very close, the optical distance is infinite in the case of 2D videos. A similar vision-test machine as below in Figure 7.8 is used.

Figure 7.8: Visual acuity test machine

Based on this test all the candidates with less than 10/10 long distance visual acuity on any eye are excluded during the test results analysis.

7.2.1.7.3 Training

A training session is used prior to the test with the following objectives:

- Help the tester to adapt the HMD in a comfortable position using the straps and adjust the HMD for the best vision (inter-eye distance, vertical orientation).
- Take notice of the test instructions (that are documented in Annex B).
- Familiarize the tester with the mouse controls for content playback and rating.

The training session consists of one sequence presented in 3 versions, namely the reference and two versions of the same clip lettered A and B. The test sequence is of course not reused for any of the main tests.

7.2.1.7.4 Main session

During the main session, the tester is in full autonomy. The tester first selects one of the five sequences before putting on the HMD and run the playbacks/ratings for each version. When a sequence has all its versions rated, the test can be validated for the sequence. Then, the user needs to remove the HMD for selecting another sequence… and so on until the 5 sequences are all rated.

Figure 7.9: Testing environment

7.2.2 Test results

7.2.2.1 Introduction

Due to the uncertainty around the reliability of the test method used for quality assessment of virtual reality contents using HMDs, it was decided to have a large set of testers (43 on definition, 48 on compression) and have a threshold of 10/10 on visual acuity for both eyes for long-distance vision. Based on these criteria, 14 persons were rejected in the resolution test and 11 people in the compression test.

Following the SAMVIQ ITU-R BT.1788 Recommendation [60], a statistical analysis was performed using both Spearman and Pearson algorithms that removed 4 people in each test.

The test results are based on 28 viewers in the definition test and 32 in compression test.

So as to get consistent results, due to lack of maturity on test methodology in the field of VR, it is advised to increase the number of testers compared to regular 2D visual assessments.

In the figures illustrating the test results, the implicit and explicit references are placed on different abscissa for clarity and avoid overlaps of confidence intervals.

7.2.2.2 Test results on spatial resolution

It can be concluded that based on the selected HMD and the available set of sequences, the optimal resolution is around 4K (4096x2048). These results are consistent over all sequences.

Even if confidence intervals are quite important, it is commonly admitted that given the number of testers retained, the shape of curves will not change. Increasing the number of testers will only improve the confidence intervals.

These confidence intervals also reflect the fact that the dynamic of scores is limited. Indeed, the quality level is quite low for this test. Maximum often reach "Good" quality.

However, the test was difficult as 4K, 6K and reference are really close to each other. This can also explain the size of confidence intervals. These limitations are mainly due to the HMD which has limited resolution.

To conclude on the optimal resolution regardless of the visualization device technology, it would be interesting to perform the same test on a regular 4K TV set.

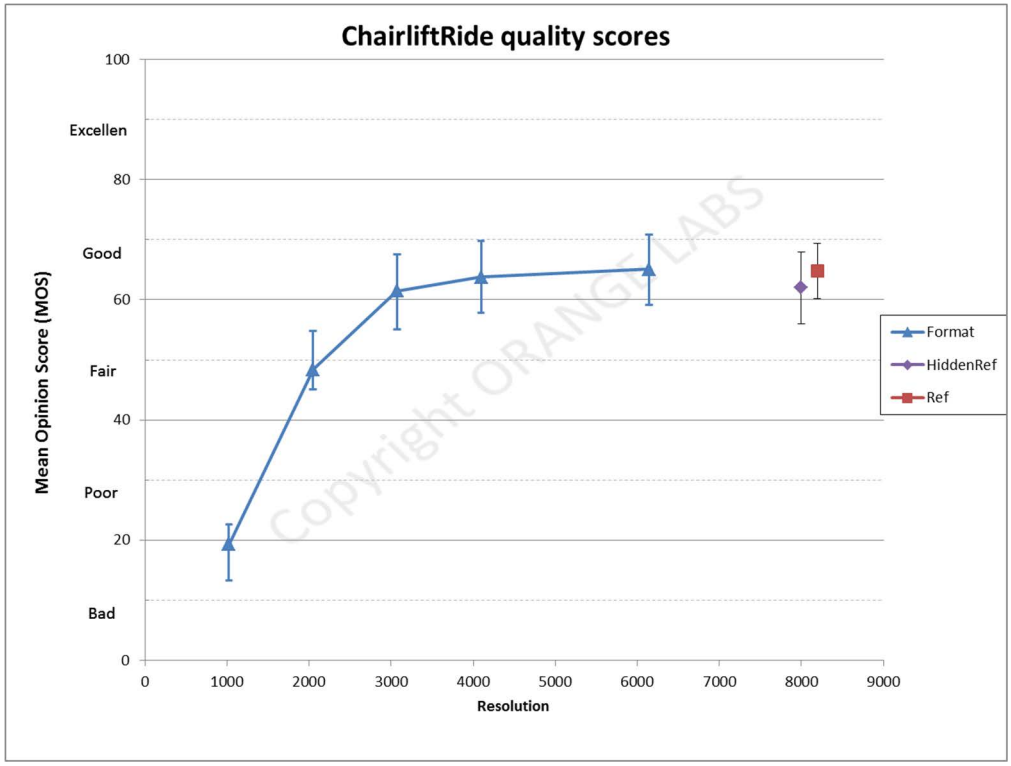


Figure 7.10: Spatial resolution quality scores for ChairliftRide sequence

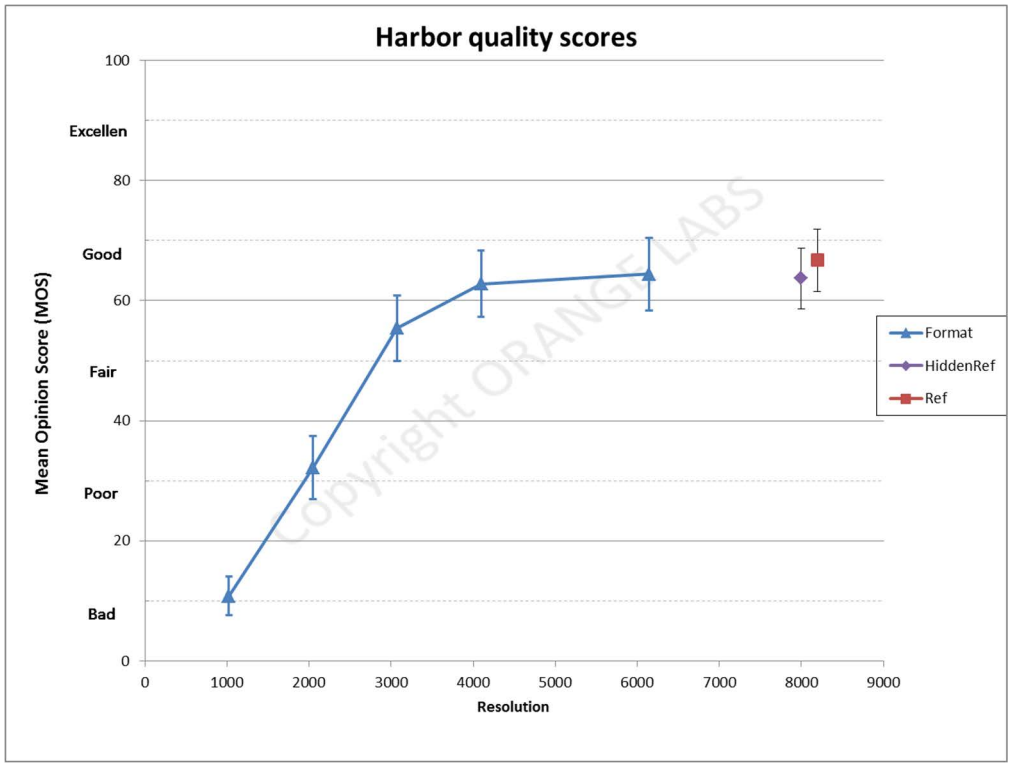


Figure 7.11: Spatial resolution quality scores for Harbor sequence

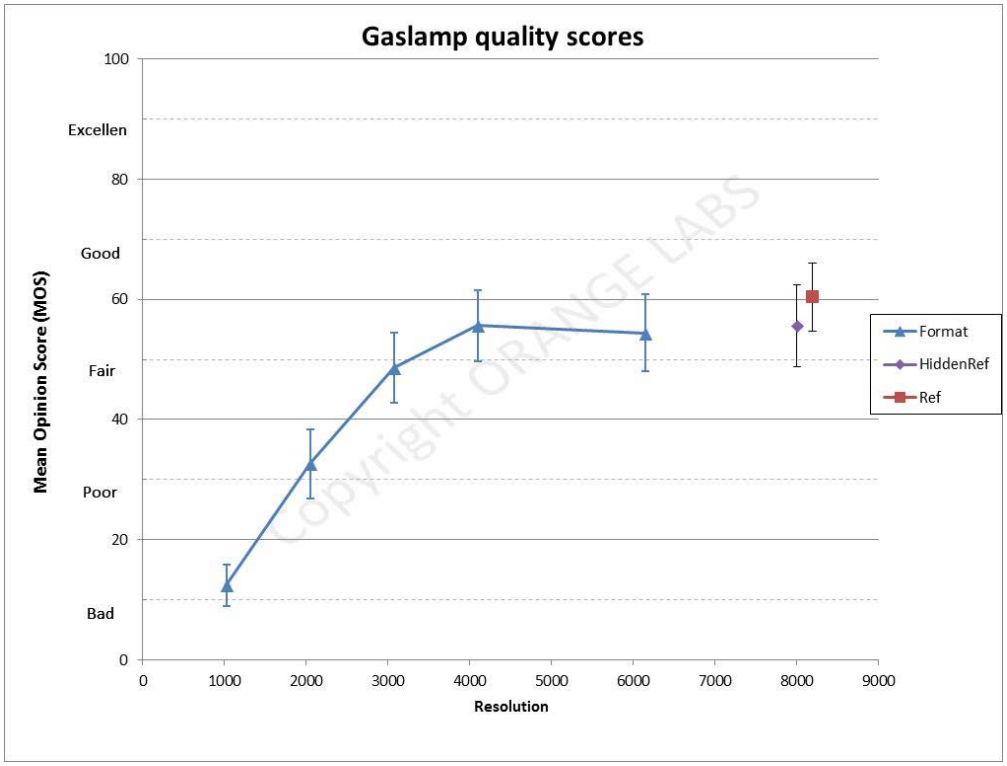


Figure 7.12: Spatial resolution quality scores for Gaslamp sequence

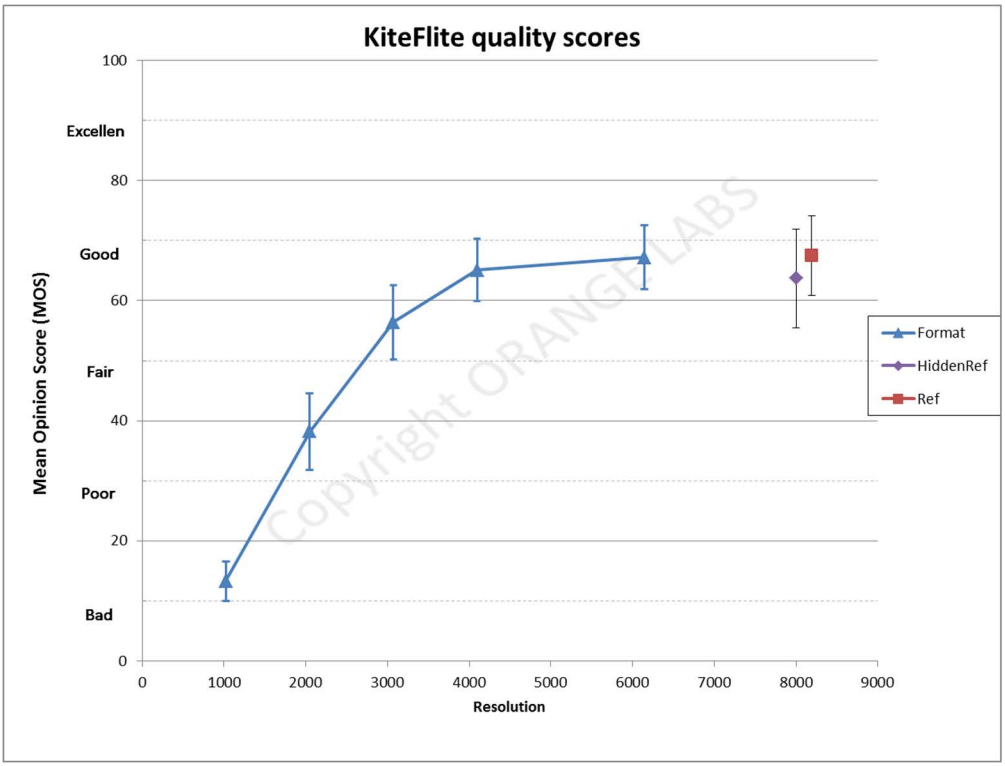


Figure 7.13: Spatial resolution quality scores for Kiteflite sequence

Figure 7.14: Spatial resolution quality scores for Trolley sequence

7.2.2.3 Test results on video compression

It can be concluded that based on the selected HMD and the available set of sequences, the optimal bitrate ranges from 5 to 12 Mbps depending on the sequence complexity.

As for the resolution test, the confidence intervals are quite important. But the quality level for the maximum bitrate is equivalent to the reference (without compression) across all test sequences.

It can be noticed that the quality level of the reference in the Skateboard sequence is very low. One of the main reasons is the jerkiness induced by the 30Hz capture rate combined to a short shutter speed.

It would be interesting to assess visual quality based on 60Hz sequences shot with optimal shutter speed.

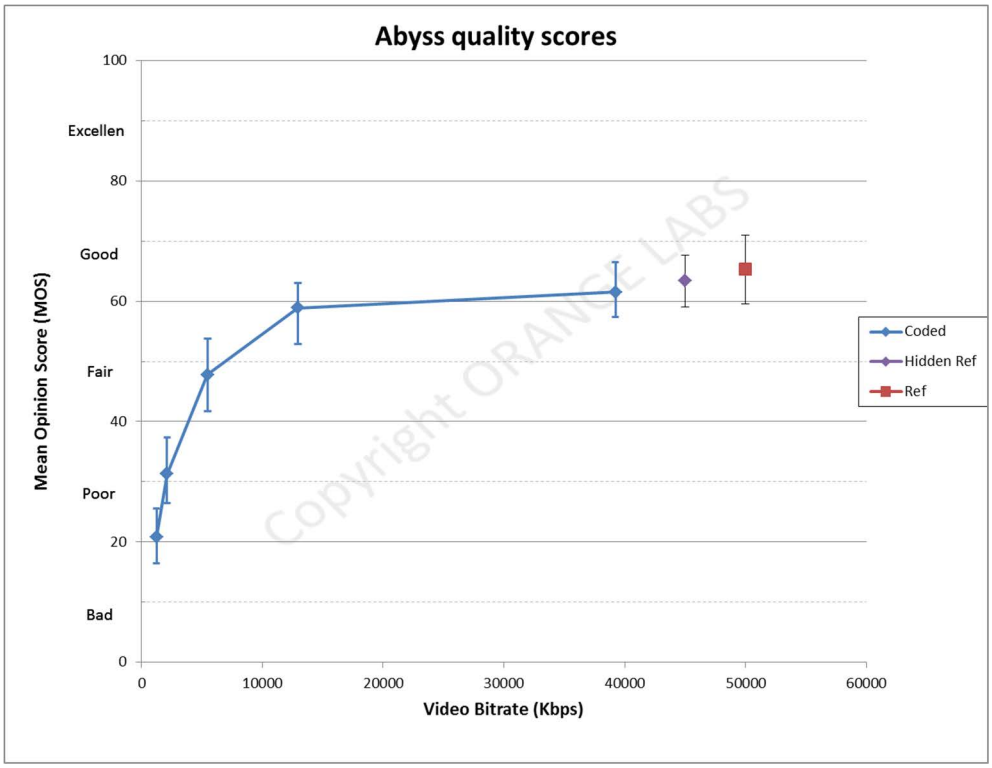


Figure 7.15: Compression quality scores for Abyss sequence

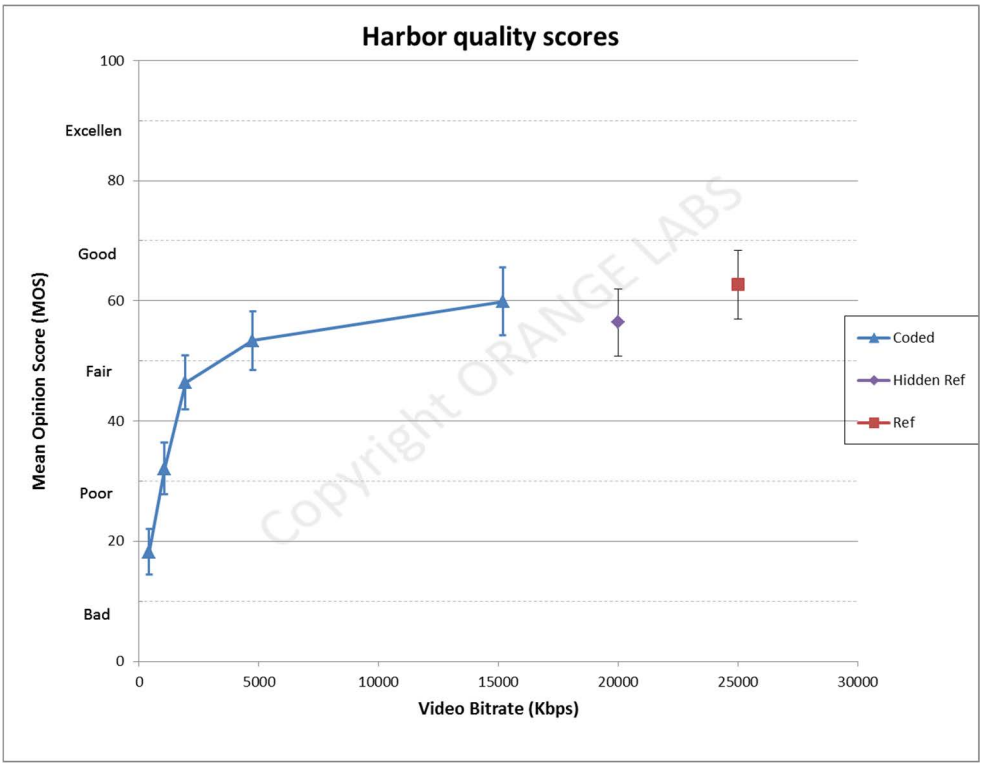


Figure 7.16: Compression quality scores for Harbor sequence

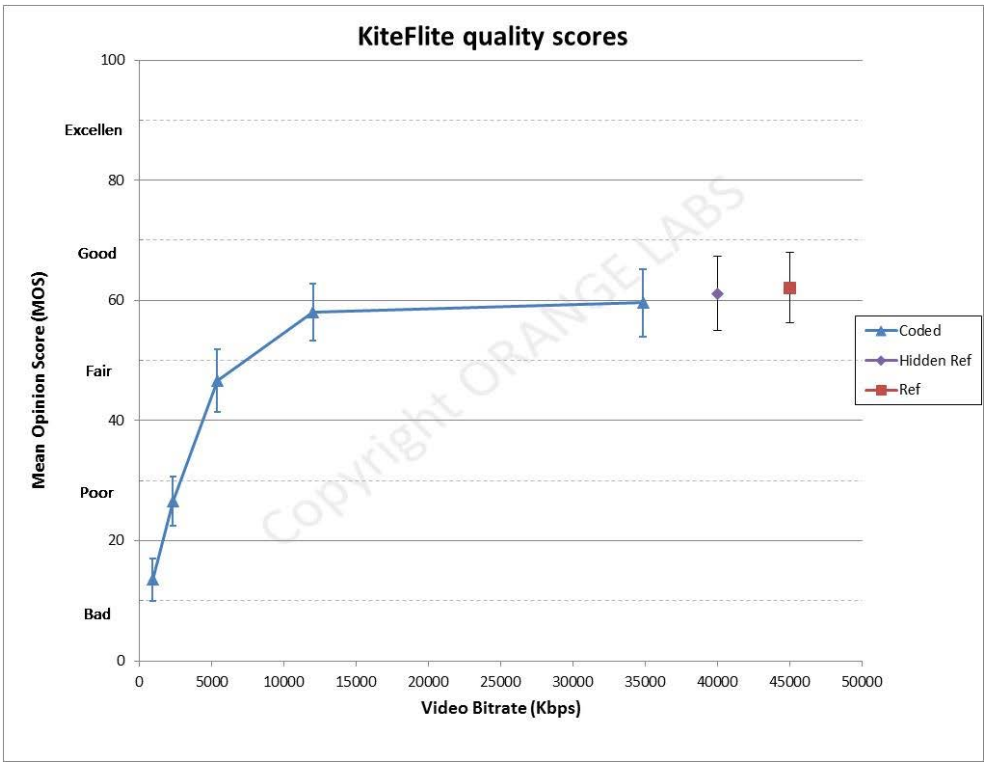


Figure 7.17: Compression quality scores for KiteFlite sequence

Figure 7.18: Compression quality scores for Skateboard sequence

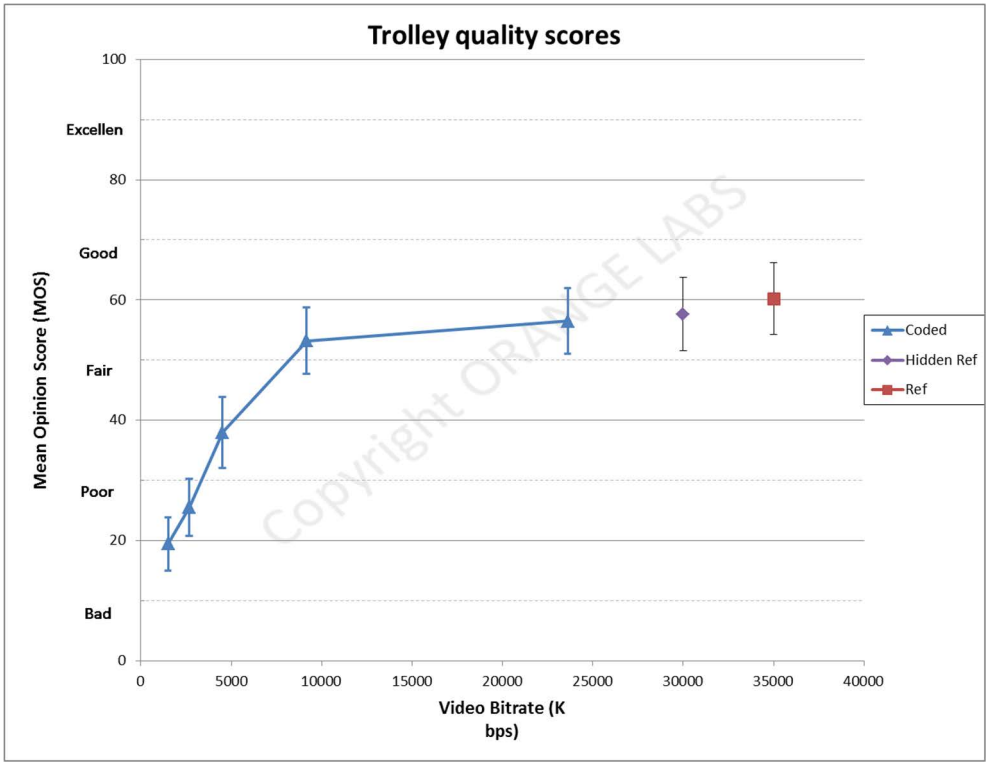


Figure 7.19: Compression quality scores for Trolley sequence

7.3 Subjective evaluation of Viewport-dependent omnidirectional video streaming

7.3.1 Introduction

This Clause introduces elements of a test methodology used to perform subjective assessment experiments of 360-degree video streaming. Also, results for an experiment are reported. The objective was to use the subjective evaluation results for gathering data about optimal operation points with the goal of reducing the required bandwidth for viewport-dependent omnidirectional video streaming.

To achieve this, an experiment was designed and divided in two parts with the idea of maximizing visual quality while at the same time minimizing streaming bandwidth. The questions under investigation were:

- How it is perceived the temporary quality difference when moving away from the foreground tile towards the background tile, given that the former is streamed at high quality, and the latter adaptively streamed at lower quality?

- Is it better to stream foreground tile data at high SNR quality, and background tile data at lower SNR quality both at the same (full) spatial resolution, or stream foreground tile data at high SNR quality, and background tile data at higher SNR quality but at lower spatial resolution?

7.3.2 Test Methodology

7.3.2.1 Introduction

This section contains few contributing elements of test methodology that could be useful for running subjective tests of omnidirectional video.

7.3.2.2 Assessment Method

The selected method for assessing omnidirectional video quality is the single-stimulus *Absolute Category Rating with Hidden Reference (ACR-HR)* [61], [62]. This is also recently considered in comparative literature [63]. A 5-grade scale [54] (with adjective category judgment) was used, with fractional values.

7.3.2.3 Instructions for the Assessment

To the test subjects it was given a description of the task in each session. The subjects are requested to move around in the 360-degree space and explore it following also a motion pattern as depicted in Figure 7.1, avoiding staying in a still position.

7.3.2.4 Data Analysis

A Differential MOS (DMOS) was computed between each test sequence and its corresponding reference sequence [55].

One important data set used for analysis was the test subjects' head orientation data (yaw, pitch and roll) in the space. This data can be used to verify that there is fair assessment for a single subject or between multiple test subjects, i.e. that multiple subjects watch the same thing at the same time for different test conditions as described in Clauses 7.1.3 and 7.1.4.

7.3.3 Subjective Test Experiments

7.3.3.1 Test Environment

The test system used in our experiment was a complete end-to-end streaming system for omnidirectional video. The server was streaming video sequences via the HTTP protocol, and the video files were MPEG DASH segmented prior to it. The transmission occurred via a private high-bandwidth WLAN 802.11ac tri-band access point. The streaming client was MPEG DASH compliant (with the needed extensions), and was implemented on a Samsung S7 phone, which was plugged to a Samsung Gear VR 2016 HMD.

7.3.3.2 Test Contents

The stimulus displayed was panoramic video (equi-rectangular projection). The test content was chosen to accommodate different genres: sport, documentary and entertainment. In particular:

- The sport sequence was a high-motion pole vault competition (Pole).
- The documentary sequence contained a bear approaching the camera location at medium motion (Bear).
- The entertainment sequence was a music clip with medium motion (Kids).

All sequences were 21 seconds long and have been shot with the Nokia OZO camera.

All contents were captured at 30 fps and encoded using a HEVC encoder with tiles. Each DASH segment size was set to be 1 second long. The vertical FoV was split into three areas: top and bottom tiles (with 360x34.5 degrees horizontal and vertical FoV each), and equatorial tiles (with 111 degrees vertical FoV). The 360-degree equatorial area of panorama frames was split into a foreground tile of 120 degrees horizontal FoV, and a background (non-visible) tile of 240 degrees horizontal FoV. For different orientations, the foreground tiles were overlapping by 60 degrees, as shown in Figure 7.20.

The content was stereoscopic, so there were two separate streams for left and right views. The maximum aggregate streaming bit rate was fixed to 23 Mbps. The full resolution panorama frames were 4K (3840x1920 pixels) with 2:1 aspect ratio. The idea was to encode the foreground tiles at full resolution (and always at high quality), and the background (plus top and bottom) tiles at either:

a) full resolution and decreasing SNR quality; or at

b) half spatial resolutions in both horizontal and vertical directions and with increasing SNR quality.

The decoding complexity is not identical for the two cases, as the first case is generally more complex than the second case.

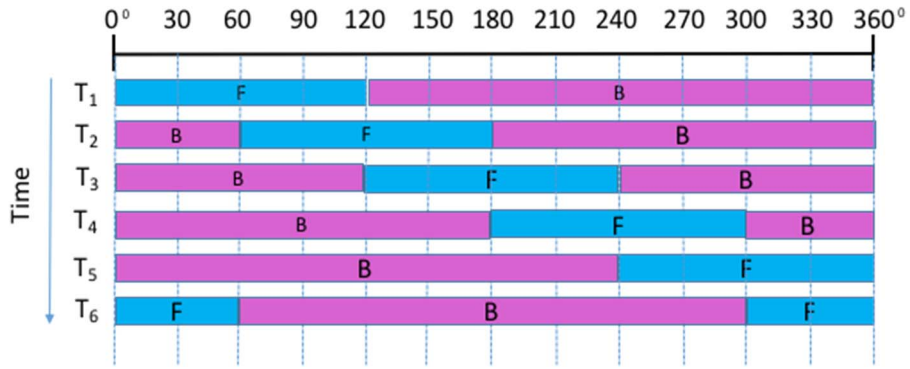


Figure 7.20: Overlapping tiling scheme (panorama view)

The video content was created at 4 different quality levels for both the full resolution and low resolution cases. To vary the sequences SNR quality, the video sequences were encoded using different Quantization Parameters (QPs) for foreground (FG) and non-foreground (i.e. background) tiles as indicated in Table 7.2. The table shows also the maximum streaming bit rates for the different genre clips. These bit rates have been calculated by taking into account all possible head orientations in the omnidirectional space.

Table 7.2: QPs and Maximum Streaming Bit Rates

		QP for FG tiles	QP for non-FG tiles	Max 360 degrees bit rate (Mbps) Bear	Max 360 degrees bit rate (Mbps) Pole	Max 360 degrees bit rate (Mbps) Kids
Full Res	A	27	27	22,5	18,9	12,7
	B	27	30	11,9	11,6	6,8
	C	27	34	8,7	9,2	4,8
	D	27	40	7,4	8,2	4,0
Low Res	E	27	28	9,6	10,2	5,9
	F	27	27	11,0	11,5	7,0
	G	27	26	13,4	13,3	8,6
	H	27	25	17,9	16,2	11,1

7.3.3.3 Test Implementation

A group of 12 expert subjects participated to the test; all had normal visual capabilities. The subjects were of age between 30 and 50 years; all were male except one. All the subjects are working with video coding or VR related video research. The hidden reference sequence for all test conditions (one for each of the genres) was encoded using QP=27 for all tiles.

In the experiment setup there were three test sessions, and the total experiment time was about 65 minutes, approximately half of which was made of test sessions wearing the HMD. One repetition of each test clip was also included. The total number of video sequences watched per test subject was 61 (including training and stabilizing clips).

The test subjects were instructed to follow a motion pattern as described in Figure 7.1 following a natural speed. This motion was matching the actual video content motion and the action of the clips. The head orientations were sampled and stored at intervals of 100 ms.

7.3.4 Results

The first research question was aiming at understanding how test subjects perceive the temporary quality difference when moving away from the foreground tile towards the background tile (or top/bottom tiles), given that the foreground tile is streamed at high quality, and the other tiles adaptively streamed at lower quality.

When moving outside of the foreground tile, the background tile with lower quality was visible for a certain time (about 1 s) before the new foreground tile was streamed at high quality (tile switch). When moving horizontally uni-directionally through the 360-degree space there were several chances of tile switches, i.e. in total several seconds were watched at a lower quality compared to the quality of the foreground tile, which was exactly the purpose of this experiment. So, the subjects' votes expressed exactly the judgment of this watching experience.

The second research question under investigation was aiming at verifying whether background tiles at full resolution and lower quality were preferred against background tiles at low resolution but higher quality in the same adaptive streaming scenario.

For both research questions, the overall goal was to find the best subjective video quality experience for the lowest streaming bit rates.

Figure 7.21 shows DMOS values related to the first research question for all three genres for decreasing video quality. Results in Table 7.3 are calculated from Table 7.2. They show that a DMOS of 3-fair for the usage of adaptive background tiles (case D) can yield a considerable streaming bit rate reduction (64,1 % on average) for different genres over the uniform high quality tiles configuration. Similarly, a DMOS of 4.5 (between 'good' and 'excellent') (case B) can yield an average streaming bit rate reduction of 44 %.

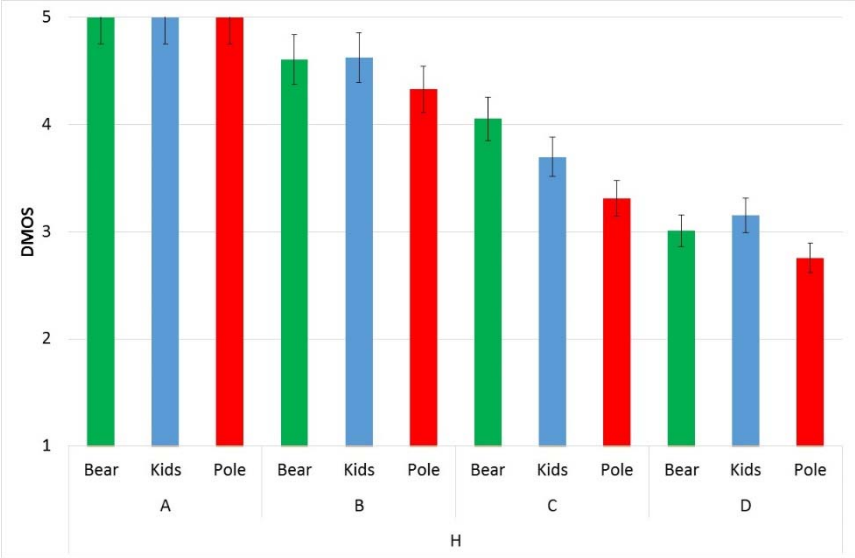


Figure 7.21: DMOS for different genres (full resolution) with 95 % confidence interval.

Table 7.3: Summary of Streaming Bit Rate Reductions for Different DMOS Values

	Avg. DMOS=3,0 (case D) Full Res	Avg. DMOS=4,5 (case B) Full Res	Avg. DMOS=3,8 (case E) Low Res	Avg. DMOS=4,2 (case G) Low Res
Bear	67,1 %	47,1 %	57,3 %	40,4 %
Pole	56,6 %	38,6 %	46,0 %	29,6 %
Kids	68,5 %	46,4 %	53,5 %	32,3 %
Average	**64,1 %**	**44,0 %**	**52,3 %**	**34,1 %**

Figure 7.22 shows data for the second research question in increasing QP quality order. Here the trend of increasing DMOS is confirmed, as expected. Results show that a DMOS of 3,8 (almost 'good') for the usage of adaptive background tiles (case E) can yield a considerable streaming bit rate reduction (52,3 % on average) for different genres over the uniform high quality tiles configuration. Similarly, a DMOS of 4.2 (more than 'good') (case G) can yield an average streaming bit rate reduction of 34,1 %. From the summary data in Table 7.3, results show that using background tiles encoded and streamed at lower resolution (but higher quality) brings about 10 % less bit rate reduction compared to background tiles encoded and streamed at full resolution (but lower quality), for comparable DMOS values (cases B and G).

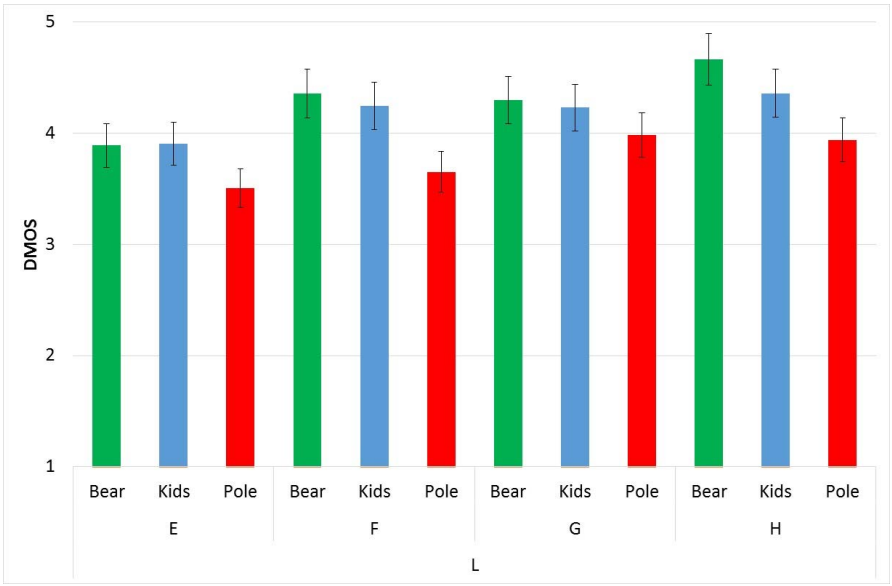


Figure 7.22: DMOS for different genres (low resolution) with 95 % confidence interval.

Finally, to show the head motion similarity within subjects, Figures 7.23 and 7.24 show the instantaneous plots of the horizontal orientations for two of the test subjects for the Bear video sequence. A similar plot is shown in Figure 7.25 to show the head motion similarity between all 12 subjects involved in the experiment. The similarities in the patterns show a quite good correlation (despite not always perfect) when considering an instantaneous moving FoV of 120 degrees. So, the subjects have been watching approximately similar parts of the same video clips at the same time instances.

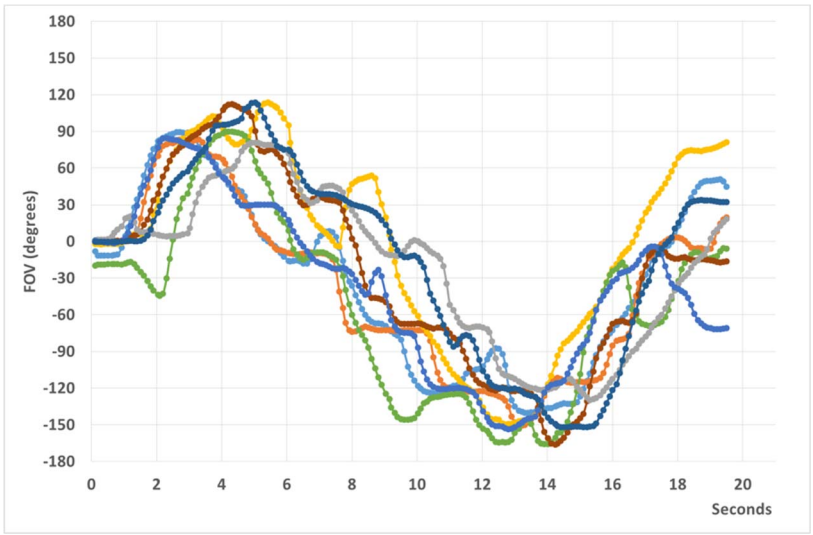


Figure 7.23: Head motion pattern within Subject 1 (Bear).

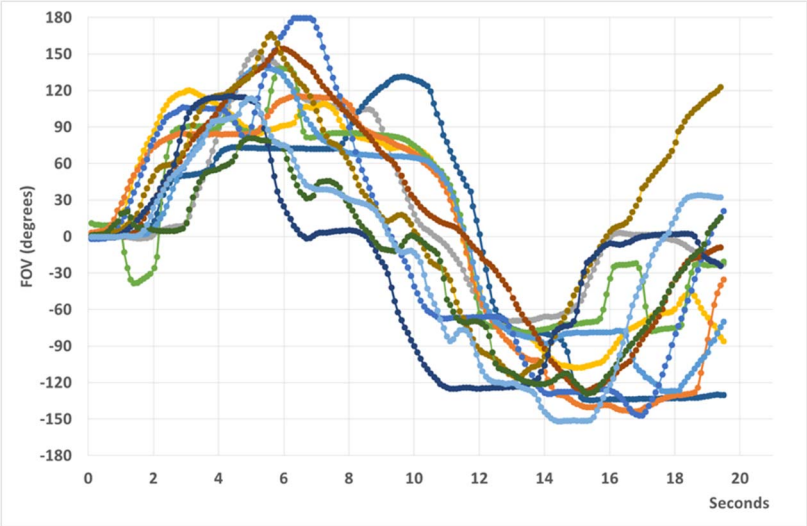


Figure 7.24: Head motion pattern within Subject 2 (Bear).

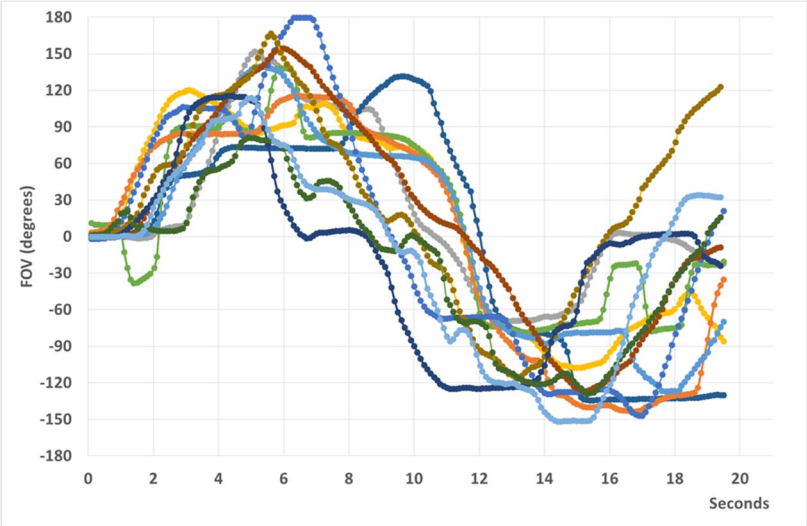


Figure 7.25: Head motion pattern between all 12 subjects (Bear).

8 Latency and synchronization aspects

8.1 Interaction latency

8.1.1 Introduction

Interaction latency is the time delay between the user interacting with a VR system and the system responding to that user interaction.

Although other interactions in VR systems may be envisaged, the main feature of virtual reality, as stated in Clause 4, is that the user is able to move and for the output sensory stimuli of the simulation to change in a manner which is consistent with those movements. It is well known that conflicts between the movement of the user and their senses, usually the visual and the vestibular senses (sensory conflict theory) may lead to nausea or motion sickness which in this case is known as virtual reality sickness. Therefore, from a performance point of view, the accuracy with which the movements of the user are reflected in the visual and audio cues and the latency before these cues respond to the user's movements are key parameters for any VR system.

8.1.2 Video interaction (Motion-to-photon) latency

The main driver on performance of the video interaction latency, often referred to as the motion-to-photon latency, comes from the angular or rotational vestibulo-ocular reflex (where the gaze is shifted in direct response to head orientation changes detected by the vestibular system by an equal and opposite reaction). Although research has shown that adaptation to sensory conflicts and other shortcomings in VR systems is possible [24], [25], [26] at real world (1.0x) magnifications, it has also been shown that sensitivity to virtual reality sickness is sometimes worse depending upon gender, general health and other factors and hence it seems reasonable that VR systems should strive to mimic the real world experience as closely as possible.

The latency of action of the angular or rotational vestibulo-ocular reflex is known to be of the order of 10 ms [27] or in a range from 7-15 milliseconds [28] and it seems reasonable that this should represent a performance goal for VR systems. The frame rate from the renderer to the viewer for VR video is usually at least 60 frames per second but more recently systems have been reporting frame rates up to 90 frames per second (~11 ms) or higher, which are more consistent with the latency requirements of the angular or rotational vestibulo-ocular reflex, albeit without any allowance for the detection of user movement and image processing times. When such detection times and image processing delays are taken into account it would seem appropriate to set a requirement of 20 ms, although it is clear that some acute users will be able to discern much lower interaction latency times [29]. It would therefore seem useful to set an objective for the interaction latency time around 10 ms.

8.1.3 Audio interaction (Motion-to-sound) latency

The response time of the human auditory system in the presence of head movement is less well characterized than the visual system but it is known to be dependent upon the nature of the sounds being heard and their direction in relation to the user.

From studies of human perception of the effect of motion-to-sound latency, there is evidence that the requirements for VR and AR are quite different [30]. In the case of VR, the user is immersed in a wholly artificial situation where there is no real world zero latency "reference" to highlight the non-zero latency of the audio rendering system.

The conclusions of [30] state that the most sensitive listeners in the test were able to detect latencies of 60 ms (with 70 % probability) for isolated auditory stimuli (VR) and of 38 ms (with 70 % probability) when a low-latency reference tone was also present in the stimulus (AR). AR is out of the scope of the present document.

These studies [30], [14] do not include a simultaneously rendered video component which may influence the perception of these latencies and the user experience. It is unclear whether such simultaneous rendering of video and audio will result in a relaxation or tightening of the motion-to-sound latency requirements. More extensive studies are highly desirable however; considering greater numbers of subjects, personalized Head Related Transfer Functions (HRTFs), and using VR/AR equipment more closely representative of the current state-of-the-art and assessing the impacts on the user experience.

The audio component interaction latency requirements of a VR system are for further study.

8.2 Audio/Video synchronization

Due to the relatively slower speed of sound compared to light it is natural that users are more accustomed to, and therefore tolerant of, sound being relatively delayed with respect to the video component than sound being relatively in advance of the video component. This effect is seen in Figure 2 of [32] depicting the detectability thresholds obtained through subjective viewing experiments. These results show that in a range from 125ms (audio delayed) to 45ms (audio advanced) it is difficult for viewers to detect the lack of synchronization. In recent years though the results of [32] have received significant scrutiny mainly because they were obtained with interlaced video of 25 or 30 Hz.

More recent studies have led to tighter recommendations e.g. [33] recommending an accuracy of between 15 ms (audio delayed) and 5 ms (audio advanced) for the synchronization, with recommended absolute limits of 60 ms (audio delayed) and 40 ms (audio advanced) for broadcast video. These figures therefore lead to an indication of an appropriate range.

In applying absolute limits for the audio/video synchronization thresholds to the VR application there needs to be appropriate account taken of the apparent distance in virtual space from the user to the source of the sound under evaluation. The limits they should be computed relative to the delays expected due to the speed of sound over the free-space path length in the virtual environment.

8.3 Report of one listening experiment for derivation of Motion-to-Sound Latency Detection Thresholds

8.3.1 Introduction

8.3.1.1 Motion to Sound Latency definition and impacts to VR QoE.

The M2S latency is the time elapsed from the transduction of a listener action (such as a head movement) until the consequences of that action are made available to the listener [48]. The M2S latency includes e.g. latencies introduced by the finite motion sensor update rate, the computation of source positions relative to the listener, and the binauralization processing. A quantity of interest in the design of virtual reality systems for 3GPP is the maximum M2S latency that can be tolerated without introducing audible artifacts.

A non-exhaustive list of audible artifacts that have been previously reported include:

1) Spatio-temporal displacement of the audio and video images, i.e. when, during and immediately after motion, the sound scene lags the visual image displayed in the field of view;

2) A sensation of spatial "slewing", i.e. where an otherwise fixed sound source appears to be floating in space due to the position error (difference between the actual head angle and the source relative to the head) introduced by the M2S. This is described e.g. [53].

3) A decreased ability to localize sources. For example, in [49] participants were asked to conduct a sound localization task. That study found that the azimuth error for localization was statistically significantly increased when the M2S latency increased from 29ms to 96ms. Similarly, localization errors with increased latency were observed in [13].

8.3.1.2 Motion to Sound Latency Detection Thresholds and its meaning for 3GPP purposes

To estimate the M2S latency detection threshold for human beings, psychophysical experiments have been conducted in the past. The detection threshold in 2AFC experiments is typically adopted as the value of the stimuli that results in 75 % of correct responses. In a 2AFC experiment, the 75 % correct response rate represents the level where the user correctly detects (i.e. excluding detections through guessing) the presence of latency 50 % of the time. Ignoring a further correction for lapse errors, the psychometric function for percentage of correct responses, with a high threshold assumption and correction for guesses can be written as [59]:

$$P(x) = p(x) + \gamma[1 - p(x)]$$

The high threshold assumption is that the listener is in one of two states: a detect state, with probability $p(x)$, and a non-detected state. If the stimulus is not detected, the listener guesses with a guess rate γ. In forced choice tasks, the guess rate is dependent on the number of alternatives given, with:

$$\gamma = \frac{1}{number\ of\ alternatives}.$$

For a 2AFC experiment, $\gamma = 0.5$, resulting that $P(x)$ =0.75 for a $p(x)$=0.5. I.e. the listener correctly detecting the presence of the impairment half of the time. In practice, the actual threshold reported in experiments vary depending on the choice of experiment chosen, since N-alternative adaptive forced choice experiments converge at different levels [57].

In addition, the level of latency that produces 75 % correct responses in 2AFC experiments is not necessarily the M2S level that should be targeted for 3GPP immersive audio systems. A condition where the user is noticing artifacts half of the time may not be the appropriate target for 3GPP immersive audio services. For example, Lindau and Weinzierl [21] have assumed a $P(x)$ =0,55 to be the detection level where "plausibility" is reached, with "plausibility" defined as *a simulation in agreement with the listener's expectation towards an equivalent real acoustic event*.

8.3.1.3 Results of previous experiments to determine M2S Latency Detection Thresholds

Table 8.1 summarizes some of the previous experiments conducted to find M2S latency detection thresholds, including values for the most discerning participant and values for the average of all participants in the experiment. Further details

can be found in the references provided. Because every test system has a mTSL associated with it, the experiments do not exactly measure the M2S latency detection threshold but rather estimate it by comparing the participant's response to the added latency over the mTSL. The mTSL for the various systems is also indicated in Table 8.1.

Table 8.1: Summary of previous experiments conducted to determine motion to sound latency detection thresholds

Study	Subj.	Stimuli	Task	mTSL[1] [ms]	DT[2] [ms]	t[3] [s]	Feedback[4]	Motion strategy[5]
Stitt et al [51]	10	Maracas	spatial stability preference	12.5	N/A** / 63.5* @ Pc=70 %	5	?	Predefined*
Stitt et al [51]	10	Complex scene	spatial stability preference	12.5	N/A** / 73.5* @ Pc=70 %	5	?	Predefined*
Yairi, S. et al [52]	10	multitone	Yes/No	9.93	42** / 94* @ Pc=75 %****	4	?	Predefined**
Mackensen, P. [58]	17	Castanets	2 AFC	50	N/A** / ~89* @ Pc=75 %	9(?)	?	Predefined**
Lindau, A. anechoic [14]	22	Male Speech	3 AFC	43	64** / 100* @ Pc=75 %	4.5	Train.	Individual
Lindau, A. reverb. [14]	22	Male Speech	3 AFC	43	52** / 100* @ Pc=75 %	4.5	Train.	Individual
Brungart, D. et al (exp1) [56]	9	Noise (contin.)	2 AFC	11.7	N/A** / 71.7* @Pc=70 %	20	Yes	Individual
This study	7	Speech (contin.)	2 AFC	2.3	63.5** / 89.6* @Pc=79.4 %***	10	Yes	Individual

NOTE 1: The averaged "minimum total system latency" of the test system used for the experiments. Note that some systems had variable mTSL during the experiments.
NOTE 2: The estimated M2S latency detection threshold reported for:
(*) the average of subjects @ threshold; or
(**) the most discerning subject;
(***) refers to levels derived after a preliminary test to detect the threshold;
(****) converting to what the percentage of correct responses would have been if this was an alternative forced choice test.
NOTE 3: The time in seconds during which the stimulus was presented.
NOTE 4: Whether feedback was given to participants after their responses. Some experiments only had feedback during the training session.
NOTE 5: Whether the participant was free to choose her motion strategy or followed a predefined motion:
(*) Listener was asked to move his/her head only one cycle (front-right-left-front or front-left-right-front).
(**) Listener was allowed to make small rotations around the vertical axis.

8.3.1.4 Analysis of previous experiments results and methodology

A first observation is that the previous experiments vary in methodology and the thresholds derived are dependent on the methodology chosen so the results cannot be directly compared between experiments. The choices of stimuli, task, feedback to participants, and whether participants were limited in their choice of head motion, all affect the psychophysical response of the participants to the stimuli. Finally, the percentage correct used for detection threshold differs between the experiments as explained above. Nevertheless, the range of M2S latency detection thresholds found in all studies is reasonably confined, varying from 63,5 ms to 100 ms for the threshold averaged across all participants, and from 42 ms to 64 ms for the most discerning participant threshold.

The previous experiments have a non-negligible mTSL because of the head-tracking and the binauralization systems employed. In addition, there are possible influences to the results due to the choice of renderer used. Another important aspect is that the studies were somewhat limited in the number of participants and there was significant variation in the detection thresholds observed between participants.

It is of interest to 3GPP to be able to further assess the impact of M2S latency to the user experience. To establish a system that can be used for further experimentation, minimizing the mTSL and rendering aspects mentioned, a new test setup was devised and a preliminary experiment conducted.

8.3.2 Test methodology

8.3.2.1 Apparatus

A new test apparatus was developed to determine M2S latency detection thresholds. The test apparatus was designed to reduce the mTSL from previous experiments and to minimize possible impacts of the rendering system. The general idea with the test apparatus was to:

1) capture the sound around the participant through the means of two microphones located close to the participant's ear; and

2) reproduce the sound around the participant through a delayed acoustic path with controlled amounts of latency. A block diagram of the test apparatus is shown in Figure 8.1.

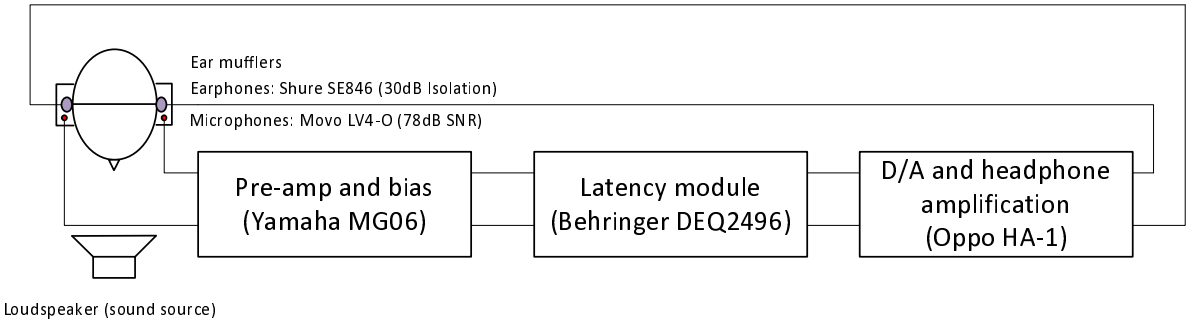


Figure 8.1: Block diagram of apparatus for Motion to Sound Latency tests

Controlled amounts of latency were introduced through a DSP module (Behringer DEQ2496) that delays the sound captured by the two microphones located close to the participant's ear. The DSP module allows the latency to be adjusted in small steps (0.3ms) varying from 0ms to 300ms. For the experiments, the amount of latency was varied through a script (implemented in Python) that could send MIDI commands to the DSP module.

To avoid possible comb-filtering effects due to the interaction between direct and delayed acoustic paths, it was necessary to isolate the participants from the direct acoustic path. This was accomplished by reproducing the delayed acoustic path to the participant through high isolation earphones (Shure SE846) and adding an additional layer of acoustic isolation through ear-mufflers. In addition, for the experiments, the stimuli around the participants were adjusted to a level low enough that participants were only able to listen the delayed acoustic path when wearing both earphones and ear-mufflers.

To avoid noticeability of microphone self-noise when reproducing the captured sound through the delayed acoustic path, two small high SNR microphones (MOVO LV4-O, 78dBSNR) were employed, capturing the sound around the participant's head. These two microphones were placed on the outside of the ear-mufflers and positioned as close as possible to the participant's left and right ears. This positioning attempted to preserve Inter-aural level and time differences as much as possible.

A loudspeaker was setup at the front centre location serving both as a sound source as well as a visual reference to the participant. The loudspeaker image can be interpreted as a simultaneous visual presentation with zero latency. This aspect was thought to be important as, in practice, most immersive audio applications are accompanied by visual stimuli. By providing a visual reference that could be used to compare the visual and auditory stimuli, effects related to the lack of synchronism between the two stimuli could also be tested.

An image of a participant wearing the test apparatus and facing the sound source is shown in Figure 8.2.

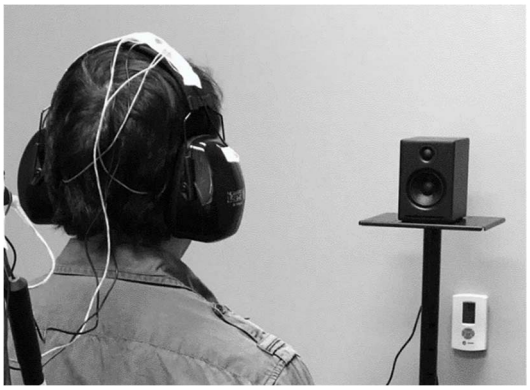


Figure 8.2: Participant wearing the test apparatus

The mTSL of the test apparatus was determined through cross-correlation and found to be only 2,3 ms (corresponding to the A/D, D/A conversion processes).

8.3.2.2 Subjects

7 subjects participated in a preliminary experiment meant to determine the general range where M2S latency detection thresholds could be observed. All participants were male with ages ranging from 31yr to 39yr. The participants were experienced in audio algorithm development but none had taken part in M2S latency detection threshold derivation tests before.

8.3.2.3 Task

In the preliminary experiment, a modified 2AFC test with an adaptive 3-down 1-up rule was conducted. The participants were presented with two samples in random order (one with added latency and one without), and were asked to determine which of the presented samples had the added latency.

For each trial, participants were given feedback as to whether their responses were correct or not. The starting latency for the test was set to a high level of 200 ms. If the participant correctly determined the sample containing latency, the latency was reduced by one step. Otherwise, the latency was increased by three steps. The initial step size was set to 200 ms and was decreased by half after three reversals. 4 step sizes were used. The three reversals for the last step size were averaged to determine a value close to the M2S latency detection threshold. Due to the 3-down, 1-up nature of the adaptive procedure, the procedure converges to a percentage of correct detections = 79,4 %.

8.3.2.4 Participant's motion instructions

Participants were seated in a chair and were free to make the exploratory head movements that they felt was necessary to detect the M2S latency. Eventually, as trials progressed and feedback was being given, the participants adapted their head movements to the motion that better facilitated their discrimination. However, it is possible that a participant did not find the best motion strategy during the experiment.

8.3.2.5 Stimuli

The stimulus chosen was a continuous speech signal, taken from podcasts. While previous studies have chosen specific test signals for a high repeatability of psychophysical stimulation, the idea in this preliminary experiment was to stimulate the participant with content that is more approximate to the actual use case of interest and hence, a long podcast was played throughout the experiment. For each trial, the actual stimulus presentation length was limited to 10 s. This duration was thought to be a good compromise between the time required to make exploratory head motions and the need to limit the duration of the experiment.

8.3.2.6 Test environment

The test environment used to conduct the tests was an office-like space with an RT60 ~ 0,5 s. The choice of a space with some amount of reverberation was made following the observation that testing in purely anechoic environments resulted in higher latency detection thresholds [14].

8.3.3 Results and listeners feedback

The preliminary experiment results on the M2S latency detection thresholds for each of the subjects are listed in Table 8.2. These results represent only an estimate of the threshold and further repeats and a different adaptive approach are required to derive a more precise threshold for each of the subjects:

Table 8.2: Results of preliminary experiment to determine range of motion to sound latency detection thresholds P_c=79,4 %

Listener	1	2	3	4	5	6	7	Average
DT[ms]	78	69	76	148	90	63	103	89,6

Each session took about 45min to 1 hour depending on the adaptation tracks for each subject.

In a post-experiment interview, some subjects reported a feeling of motion sickness. Although latency existed only for the audio stimuli (there was no latency on the visual stimuli), it is possible that the head movements themselves caused the motion sickness. Further investigation on motion sickness aspects require further evaluation. Some subjects felt that, when latency was present, the sound image was not stable, and used that artifact as a cue to detect the M2S latency. This is consistent with the spatial "slewing" effect observed in [53] and the actual task specified in [51]. Other subjects focused on the spatio-temporal displacement between the audio and visual images as a cue to detect the M2S latency thresholds.

8.3.4 Conclusions

The results obtained with this preliminary experiment seem in line with results previously reported in academic studies, although the experiment only aimed to find a range of latencies to be further tested with the test apparatus. The mTSL of the test apparatus (2,3 ms) is significantly lower than the values obtained for the most discriminating subjects. Therefore, the test apparatus/methodology can be used to derive actual psychometric curves and further estimate M2S latency at other levels of detectability.

9 Gap Analysis, Recommended Objectives and Candidate solutions for VR Use Cases

9.1 Introduction

The relevant use cases described in the present document can be broadly grouped into the following categories:

- UE consumption of managed and third-party VR content:
  - 5.2 Event broadcast/multicast use cases
  - 5.3 VR streaming
  - 5.4 Distributing 360 A/V content library in 3GPP
  - 5.5 Live services consumed on HMD
  - 5.7 Cinematic VR
  - 5.13 Use cases for highlight region(s) in VR video
- VR services including UE-generated content:
  - 5.6 Social TV and VR (note that for this use case, a detailed architectural decomposition still has to be done)
  - 5.8 Learning application use cases
  - 5.9 VR calls use cases
  - 5.10 User generated VR use cases

- 5.11 Communications in Virtual Worlds

9.2 UE consumption of managed and third-party VR content

9.2.1 Gap Analysis

These use cases, primarily the one documented in Clause 5.4, may be implemented using a service architecture as introduced in Clause 4.4.2. In this case, the delivery may use progressive download, DASH-based streaming or DASH-over-MBMS for encapsulation and delivery.

Based on the architecture in Figure 4.24, a system as above requires to define the following components:

- Original content formats on interface B:
  - For audio that can be used by a 3D audio encoding engine.
  - For video that can be used by pre-processing and image/video encoding.
- Mapping formats from a 3-dimensional representation to a 2D representation in order to use regular video encoding engines.
- Encapsulation of the media format tracks to ISO file format together, adding sufficient information on to decode and render the VR content. the information may be on codec level, file format level, or both.
- Delivery of the formats through regular download, DASH delivery and DASH over MBMS delivery.
- Static and dynamic capabilities and environment data that is collected from VR application and the VR platform. This includes decoding and rendering capabilities, as well as sensoring data.
- Media decoders that support the decoding of the formats delivered to the receiver.
- In case decoding and rendering are not performed in an integrated block (decoder/renderer), information for audio and video rendering present the information on the VR display and rendering environments.

Based on the considerations above, to support the use case, the following functions are missing in the present document:

- Consistent content contribution formats for audio and video for 360/3D AV applications including their metadata. This aspect may be informative and may be considered outside the scope of 3GPP, but there should at the minimum an assumption on these formats.
- Efficient encoding of 360 video content. In the initial versions, this encoding is split in two steps, namely a projection mapping from 360 video to 2D (projection mapping) and a regular video encoding. In order to address the latter, high-end video decoding platforms should be targeted. The TV video profile codecs in TS 26.116 [34] may fulfill the requirements. In an extension to basic encoding, viewport specific encoding may be considered. This may for example be supported by the use of specific projection maps or tile-based encoding. This aspect is considered as an optimization rather than essential feature. In addition, the appropriate encoding of metadata to use used by the display/rendering, is necessary. Typically, SEI messages are defined to support the rendering.
- Efficient encoding of 3D audio.
- Encapsulation of VR media into a file format for download delivery. This requires extensions to the 3GP file format.
- Providing the relevant enablers for DASH delivery of VR experiences based on the encoding and encapsulation.
- Providing the necessary capabilities for static and dynamic consumption of the encoded and delivered experiences in the Internet media type and the DASH MPD.
- A reference client architecture that provides the signalling and processing steps for download delivery as well as DASH delivery as well as the interfaces between the VR service platform, the VR application (e.g. sensors), and the VR rendering system (displays, GPU, loudspeakers).
- Decoding requirements for the defined 360 video formats.
- Decoding requirements for the defined 3D audio formats.

- Possibly rendering requirements or recommendations for the above formats, for both separate and integrated decoding/rendering.

9.2.2 Recommended Objectives

The following terms are used in the following:

- Media Profile: file format track, including elementary stream constraints for a specific media type enabling a VR component.
- Presentation Profile: Combination of different tools, including audio and video media profile, to provide a full VR Experience.
- ISO BMFF Profile: The inclusion of a presentation profile into to an ISO BMFF file to provide a full VR Experience.
- DASH Profile: Mapping of media to a DASH Media Presentation.

Based on the discussion and the use case above, the following requirements are derived for a solution addressing the use case in Clause 5.4.

General

1) The solution is expected to provide for interoperable exchange of VR360 content.
2) The solution is expected to avoid multiple tools for the same functionality to reduce implementation burden and improve interoperability.
3) The solution is expected to enable good quality and performance.
4) The solution is expected to enable interoperable and independent implementations, following common specification rules and practices in 3GPP SA4, e.g. conformance and test tools.
5) The solution is expected to enable full interoperability between services/content and UEs/clients:
    5.1) The solution is expected a very low number of fully specified interoperability points that include what is traditionally known as Profile and Level information.
    5.2) Interoperability points addressing a Media Profile file format constraints, elementary stream constraints and rendering information.
    5.3) Interoperability points are expected to address a Presentation Profile for a full VR experience including different media, enabling their temporal synchronization and spatial alignment.
    5.4) The solution is expected to define at least one media profile for audio.
    5.5) The solution is expected to define at least one media profile for video.
    5.6) The solution is expected to define at least one presentation profile that includes one audio and one video media profile.
6) The solution is recommended to take into account the capabilities of high quality devices such as HMDs that are on the market today or that are on the market by the time the specification is published.
7) The solution is expected to support the representation, storage, delivery and rendering of:
    7.1) Omnidirectional (up to 360° spherical) coded image/video (monoscopic and stereoscopic) with 3 DoF.
    7.2) 3D audio.
8) The solution is expected to work with existing 3GPP PSS and MBMS storage and delivery formats.
9) The solution is expected to support temporal synchronization and spatial alignment between different media types, in particular between audio and video.
10) The solution is expected to enable applications to use hardware-supported or pre-installed independently manufactured decoders and renderers through defined conformance points.

11) The solution is expected to support viewport-dependent processing (this may include delivery, decoding and rendering).

12) The solution is expected to support at least one Presentation Profile that requires support for neither viewport-dependent delivery nor viewport-dependent decoding.

NOTE: It is obvious that there will be viewport-dependent **rendering**, both for visual and audio components.

Delivery

13) The Specification is expected to support the following methods of distribution:

13.1) Download and Progressive Download as defined in PSS based on HTTP and the 3GP/ISO BMFF file format.

13.2) Download Delivery as defined in MBMS using the 3GP/ISO BMFF file format.

13.3) DASH-based streaming as defined in PSS.

13.4) DASH-based distribution over MBMS.

Visual

14) The solution is expected to enable content exchange with high visual perceptual quality.

15) The solution is expected to support distribution of full panorama resolutions up to 4K to decoders capable of decoding only up to 4K@60fps.

16) The solution may support distribution of full panorama resolutions beyond 4K (e.g. 8K, 12K), to decoders capable of decoding only up to 4K@60fps, if sufficient interoperability can be achieved.

17) The solution is expected to support metadata for the rendering of spherical video on a 2D screen.

18) The solution is expected to support encoding of equirectangular projection (ERP) maps for monoscopic and stereoscopic video, in an efficient manner.

Audio

19) An audio media profile is expected to:

19.1) support sound quality adequate for entertainment/broadcast (assessed by subjective testing, for example a scale of *Excellent* with ITU-R BS.1534)

19.2) support binauralization and immersive rendering with sufficiently low motion-to-sound latency

19.3) support 3D Audio distribution, decoding & rendering.

19.4) support immersive content, e.g. higher order Ambisonics,

19.5) support a combination of diegetic and non-diegetic content sources.

19.6) be capable to ingest and carry all content types:

19.6.1) audio channels;

19.6.2) audio objects;

19.6.3) scene-based audio;

19.6.4) and combinations of the above.

19.7) be able to carry dynamic meta-data for combining, presenting and rendering all content types.

Security

20) The solution is expected to not preclude:

20.1) Solution and rendering to support secure media pipelines.

20.2) Efficient distribution for multiple DRM systems (e.g. using common encryption).

9.2.3 Candidate Solutions

9.2.3.1 Summary

MPEG initiated work for media and presentation profiles called OMAF (Omnidirectional MediA Format) [35].

A candidate solution for the use case, addressing the recommended objective above is the "OMAF Baseline Viewport-Independent Presentation Profile". This profile includes two media profiles:

- OMAF 3D Audio Baseline Media Profile.
- Viewport-Independent baseline media profile.

Spatial alignment and temporal synchronization is provided by the integration into the ISO BMFF file format and/or a DASH Media Presentation.

To address extended use cases, also the "HEVC viewport dependent baseline media profile" may be considered.

In the following, a summary of each of those media profiles referenced above is provided. For details, please refer to the expected LS from MPEG.

9.2.3.2 OMAF 3D Audio Baseline Profile

This media profile fulfils requirements to support 3D audio. Channels, objects and Higher-Order Ambisonics is supported, as well as combinations of those. The profile is based on MPEG-H 3D Audio [46].

Note that MPEG-H 3D audio is designed for unidirectional media delivery and has not been evaluated in any other context, and is thus expected to be unsuitable for conversational applications.

MPEG-H 3D Audio [46] specifies coding of immersive audio material and the storage of the coded representation in a ISO Base Media File Format (ISOBMFF) track. The MPEG-H 3D Audio decoder has a constant latency, see "MPEG-H 3DA functional blocks, internal processing domain and delay numbers" of [3DA]. With this information, content authors can synchronize audio and video portions of a media presentation, e.g. ensuring lip-synch. When orientation sensor inputs (i.e. pitch, yaw, roll) of an MPEG-H 3D Audio decoder change, there will be some algorithmic and implementation latency (perhaps tens of ms) between user head movement and the desired sound field orientation. This latency will not impact audio/visual synchronization (i.e. lip synch), but only represents the lag of the rendered sound field with respect to the user head orientation.

MPEG-H 3D Audio specifies methods for binauralizing the presentation of immersive content for playback via headphones, as is needed for 360 Media VR presentations. MPEG-H 3D Audio specifies a normative interface for the user's orientation, as Pitch, Yaw, Roll, and 3D Audio technology permits low-complexity, low-latency rendering of the audio scene to any user orientation.

9.2.3.3 Viewport-Independent baseline media profile

This media profile fulfils basic requirements to support omnidirectional video. Both monoscopic and stereoscopic spherical video up to 360 degrees is supported. The profile does neither require viewport dependent decoding nor viewpoint dependent delivery. Regular DASH clients, file format parsers and HEVC decoder engines can be used for distribution and decoding. The profile also minimizes the options for basic interoperability. The profile requires clients to support:

- HEVC Main 10 profile, Main tier, Level 5.1 with some restrictions and SEI messages to support signalling of:
  - Equirectangular projection maps.
  - Frame-packing using either SbS or TaB to support stereoscopic video.
- Simple extensions to the ISO file format (based on OMAF) to signal projection maps and frame-packing.
- Mapping to DASH by using CMAF media profile constraints for HEVC and a restricted amount of signalling.

9.2.3.4 Viewport-Dependent baseline media profile

This media profile enables viewport-dependent delivery and decoding based on HEVC Main 10 profile, Main tier, Level 5.1.

The profile requires clients to support:

- HEVC Main 10 profile, Main tier, Level 5.1 with some restrictions and SEI messages to support signalling of:
  - Equirectangular projection maps.
  - Frame-packing using either SbS or TaB to support stereoscopic video.
- Advanced extensions to the ISO file format (based on OMAF) to signal projection maps, frame-packing, region-wise packing, tiling, extractors and viewport-adaptation.
- Mapping to DASH using Preselection to signal different viewport.

9.2.3.5 Metadata for highlight region description

The highlight regions in the VR video, for all use cases in Section 5.13, may for example be represented by indicating regions or points on the sphere.

In this case:

1) The region on the sphere can be described as a central point (*yaw_center, pitch_center, roll_center*), with horizontal and vertical range of the sphere region, and region shape. The sphere region description can indicate the recommended viewport that is intended to be displayed when the user does not have control of the viewing orientation or has released control of the viewing orientation. The recommend viewport metadata may be used to describe highlight regions of content consumed on either VR or non-VR devices (such as a TV), for example in use case 5.13.3 and 5.13.4.

2) The point on the sphere can be described as a spherical point *(yaw, pitch, roll).* The point description can describe initial viewpoint metadata to indicate initial viewport orientations intended to be used when playing 360 video. Specifically, the initial viewpoint metadata may be used to describe initial viewing orientation of content consumed on VR devices (such as HMD), for example in use cases 5.13.1 - Initial view point for on demand content, and 5.13.2 - View point for random tuning in.

9.3 VR services including UE generated content

9.3.1 Gap Analysis

For the use cases of VR services including UE generated content the following gaps are identified in the present document:

General:

- A reference end-to-end architecture that provides signalling and processing steps for real-time delivery as well as the interfaces between the UEs and network nodes.
- Delivery of the formats through appropriate protocols and mechanisms.
- Client architecture includes the interfaces to sensors, the VR rendering system (e.g. displays, headphones), and audio and video capturing.
- Performance and quality requirements and design constraints for the coding and rendering of 360° video and immersive audio.
- Static and dynamic capabilities and environment data that is collected from VR application and the VR platform. This includes decoding and rendering capabilities, as well as sensor data.
- In addition, the appropriate provisioning of metadata to describe the spatial content that can be used by the display/rendering, is necessary.
- Decoding of the formats and metadata delivered to the receiver.

- Methods to inform the media sender about the rendering and output capabilities of the receiving UE.
- Encapsulation of the media formats for real-time transport, live upload or storage, adding sufficient information to describe, decode and/or render the VR content.
- Media encoders and decoders that support the encoding and decoding of the formats.
- A limited subset of consistent user generated VR content formats for audio and video including their metadata for different use cases.

For Video

- Video formats for 360° video.
- Mapping formats from a 3-dimensional representation to a 2D representation in order to use regular video encoding engines.
- In an extension to basic encoding, viewport specific encoding may be considered for certain use cases (in particular in one-to-one communication), whereas for other use cases highest quality of a non-viewport dependent quality is necessary (one-to-many).
- Efficient encoding and decoding of 360˚ video content, adapted to the use case.

For Audio

- Input audio formats for VR services including UE generated content.
- Efficient coding of relevant immersive audio input signals that are captured by multiple microphones.
- Immersive audio encoder and decoder specifications for VR services including UE generated content.
- Immersive rendering of audio signals for different playback scenarios considering sensor data.

9.3.2 Recommended Objectives

Based on the above identified gaps, the solution for services including UE generated content should meet the following objectives:

General

- The solution is expected to provide interoperability between different types of 3GPP VR UEs (mobile terminals, high-end IOT devices, surveillance cameras), network nodes for new VR services and existing services.
- The solution is expected to enable interoperable and independent implementations, following common specification rules and practices in 3GPP SA4, e.g. conformance and test tools.
- The solution is expected to enable temporal synchronization and spatial alignment of audio and video.
- The solution is expected to enable sufficiently low algorithmic delay for conversational VR use cases.
- The solution is expected to support a wide bitrate range to allow for operating under different and varying channel and operating conditions.
- The solution is expected to handle asymmetrical conversations, e.g. with "VR media" in one direction.

Video

- The solution is expected to enable high perceptual visual quality for 360° video.
- The solution is expected to support distribution of full panorama resolutions
- The solution is expected to support metadata for the rendering of spherical video on a 2D screen.
- The solution is expected to support encoding of equi-rectangular projection (ERP) maps for monoscopic and stereoscopic video, in an efficient manner.

- Solutions are expected to support UE with different capturing capabilities, ranging from low tier to high end devices, e.g. 2D cameras up to multi-lens stereoscopic cameras.
- The solution is expected to support UE with different presentation capabilities ranging low tier to high end devices, e.g. device with 2D screens up to stereoscopic HMDs.
- The solution is expected to be interoperable with MTSI.

Audio

- The solution is expected to handle encoding/decoding/rendering of speech, music and generic sound.
- The solution is expected to support encoding of channel-based audio (e.g. mono, stereo or 5.1) and scene-based audio (e.g. higher-order ambisonics) inputs including geometric information about the sound field and sound sources. This includes support for diegetic and non-diegetic input.
- The solution is expected to provide a decoder for the encoded format that is expected to include a renderer with an interface for listener positional information enabling immersive user experience with sufficiently-low motion to sound latency.
- The solution is expected to support low latency that would enable both conversational and live-streaming services over 4G/5G.
- The solution is expected to support advanced error robustness under realistic transmission conditions from clean channels to channels with packet loss and delay jitter and to be optimized for 4G/5G.
- The solution is expected to be suitable for a wide range of potential UEs, ranging from low tier to high end.
- The solution is expected to be interoperable with MTSI.

9.3.3 Candidate Solutions

A candidate solution for the real-time VR use cases, addressing the objectives in Clause 9.3.2.

For Video the extensions to existing MTSI codecs (or existing tele-presence codecs) are potential solutions. The elementary stream constraints of the Viewport-Independent baseline media profile may be used. Multiple video streams may be generated with MTSI codecs, for example to support multi-camera use cases. Encapsulation into RTP as well as SDP signalling is FFS.

For Audio, the MTSI codecs are potential solutions for channel-based audio based on multi mono operation and obviously provide MTSI interoperability. Encapsulation into RTP and ISO file formats for MMS, as well as SDP signalling is ffs.

9.4 Quality of experience of VR

9.4.1 Introduction

Generally speaking, the technical factors that will impact quality of experience may be attributed to: the transmission network, the content type and the device. This section analyses the VR QoE technology gap based on these three aspects.

9.4.2 Network impact on quality of experience (QoE)

Network performance is important for the VR service that is streamed to the UE in real-time when it is needed. For example, in case of FOV downloading, when a FOV video (rather than 360° video) is downloaded in real-time, and user interaction occurs, e.g. head movement, then the relevant FOV video segment needs to be delivered, within certain latency limits to enable a believable experience without producing VR sickness effects. In this case, the information about how long the user waits before the high-quality version appears would be useful for the network to better adapt streaming configuration in order to enable better user experience.

Similarly, in case of adaptive ROI streaming, when the user chooses a ROI from the thumbnail video, the server will respond with the corresponding overlapping high resolution tiles within a certain delay to ensure a seamless experience.

Based on the above analysis, it is clear that events involved in VR services are more complex than for traditional streaming video. In order to better understand what happens to the user end, the network operators need to collect QoE metrics that are able to represent these features and events. Based on these metrics, the server or the operator is able to analysis which part of delay contributes the most, and according solution is applied.

As a result, further study is required to determine what to report and how to describe these events in a simple and clear way for better understanding. A baseline reference for this work could be the Play List defined in [39] for DASH and Progressive downloading.

9.4.3 Content impact on quality of experience

In [39], the MPD information sent by the client for content quality evaluation in DASH based streaming video applications includes: @bandwidth, @qualityRanking, @width, @height, @mimeType, and @codecs. While many of these will apply for DASH based VR video, a VR service needs to create a virtual environment that enables users to feel immersed with sound, image or other stimuli. This places more requirements on the creation and delivery of content. For example, when pyramid projection is used, the user's current viewing area will be represented with higher resolution while the area outside the viewport will be encoded with lower resolution. This will inevitably lead to degraded video quality when a user moves his head to look at objects within the lower resolution area. In this case, the information about how the content is projected and mapped is useful to evaluate the content quality from the user's point of view. For example, the audio bandwidth is also likely to affect the feeling of immersion experienced by the user.

Compared with the MPD information defined in [39], more information for a VR service is therefore very likely to be needed by the network operator to better understand and manage the delivered media characteristics. As a result, further study will be necessary to determine what kind of additional information is needed for quality evaluation of VR experience.

9.4.4 Device impact on quality of experience

The device plays an important role in the end-to-end user experience. For a traditional mobile phone, relevant QoE metrics are defined in [39] to represent its impact on end user experience, including displayed video resolution as well as the physical screen characteristics, but a VR device will have more features. A typical VR device usually has such attributes as: screen size, resolution, pixel size, field of view, refresh rate, head-tracking/eye-tracking latency, degree of freedom, weight, etc. and for example the audio bandwidth reproduced by the device is also likely to have an impact on the VR experience. As a result, all of the potential device information that could be related to the user experience of VR services should be investigated and requires further study.

10 Conclusion

10.1 Conclusion on Ambisonics audio aspects

While technology trends in the VR area seem manifold, there are strong indications that scene-based VR audio representations based on Ambisonics are highly relevant. Although the audio experiments presented in this technical report are limited, e.g. in the number of audio items and renderers utilized, there is an indication that FOA-based VR services utilizing some 3GPP codecs are possible. It has not been studied whether it is feasible to support HOA using existing 3GPP codecs.

One study, presented in Clause 6.1.5, indicates that FOA can provide a statistically significantly better spatial localization quality than what can be achieved in mono or stereo (as supported in 3GPP speech and audio services). An important aspect of the Ambisonics representations is the ability to present immersive sound allowing time-variant rotations (yaw, pitch, roll) which results in a more natural listening experience than is achievable with static mono or stereo channels.

Two studies, presented in Clauses 6.1.3 and 6.1.5, have also shown that the localization quality of FOA representations is statistically significantly lower than for HOA representations. The findings suggest that solutions providing HOA representations can provide a higher quality than what can be obtained by FOA solutions.

Two other studies, presented in Clauses 6.1.6 and 6.1.7, assessed coding of FOA B-Format signal representations with 3GPP eAAC+ and 3GPP EVS codecs, respectively. These were comparative studies assessing the quality of coded FOA representations compared to uncoded FOA.

The study presented in Clause 6.1.6 suggests that two 3GPP eAAC+ stereo streams can appropriately carry B-format audio that was derived from 7.1.4 channel audio, after a conversion to A-format. Renderered to a 7.1.4 loudspeaker

array as well as binauralized to headphones with generic HRTFs and without head-tracking MUSHRA listening test results indicate that the “Good” to “Excellent” quality range is achievable at bitrates from 64kbps (2x32kbps) to 192kbps (2x96kbps). In that range a quality increase is observed that is commensurate with increased bitrate.

Clause 6.1.7 shows that the 3GPP EVS codec can be used to encode super-wideband FOA B-format representations obtaining MUSHRA scores in the “Good” to “Excellent” quality range with increasing quality from 4x24.4 kbit/s to 4x96 kbit/s compared to uncoded FOA. This shows the potential of using the 3GPP EVS codec for FOA-based VR audio services for conversational and streaming use cases.

An EVS-based FOA solution would without optimizations result in a worst-case computational complexity and RAM usage of four times the EVS codec. This would be 351.88 WMOPS and 596 kW (16-bit) RAM for the combination of encoders and decoders, while it would be 126.88 WMOPS for decoder-only use cases. It can however be assumed that program instructions and ROM can be shared between the codec instances, i.e. resulting in the same requirements as for the EVS codec for mono signals.

By using ACN/SN3D formatted FOA signals, it was shown in Clause 6.1.7 that a generic FOA renderer (designed to process ACN/SN3D formatted FOA signals) can be used to deliver coded audio signals to the listeners over loudspeakers. Aspects of binauralization were not particularily assessed in the study.

The results presented in this report indicate that: (1) FOA enables a user experience for VR exceeding the experience with mono or stereo audio in a statistically significant manner, and (2) there is additionally a statistically significant quality increase with HOA over FOA. The results show further that some existing 3GPP speech/audio codecs are capable of encoding FOA audio with high quality. This suggests that normative work specifying VR services could rely on existing 3GPP codecs at least as short-term solution for enabling the carriage of the VR audio component. By specifying a generic Ambisonics format such as ACN/SN3D and a fixed bit-distribution between the channels the need for additional metadata and SDP parameters would be limited. It could e.g. be that the existing EVS SDP parameters for transport of 4 EVS channels are used and a new SDP parameter implying the ambisonics format is defined. Noting that there are audio rendering and binauralization technologies that already exist outside of 3GPP, end-to-end solutions for the audio component of 3GPP VR services could be enabled in a brief time frame.

10.2 Subjective tests for VR audio systems evaluations

This technical report has collected a range of potentially applicable tests, quality attributes and experimental methodologies for VR QoE. Test methodologies previously used for the evaluation of 3D audio systems can be considered for assessing the quality of coding and rendering schemes with some limitations, e.g. lack of head-tracking during the rendering. However, no standardized subjective test methodology has been specified yet for VR audio. It is recommended to start the definition and specification of test methodologies for the evaluation of perceived listening quality in immersive audio systems.

10.3 Conclusion on QoE aspects

User experience based network management is important for operators so that they may provide the best quality of experience for VR. As described in clause 9.4, it is necessary to take into account the following three aspects: content quality, the network constraints and the device limitations. Defining VR-service specific QoE metrics will allow operators to understand and manage how end users are experiencing specific VR services. Based on these QoE metrics, operators may also perform problem analysis and trouble shooting in a similar manner to the way in which services such as DASH, progressive downloading and MTSI are provided with QoE metrics and tools at the present time. It is proposed that a study item on QoE metrics relevant to VR user experience be initiated.

10.4 Conclusion on VR streaming

For UE consumption of managed and third-party VR content, the following conclusions, to adequately support these use cases, a number of gaps in 3GPP specifications have been identified in Clause 9.2.1. The report documents recommended objectives in Clause 9.2.2 and some candidate solutions to address the recommended objectives in Clause 9.2.3. It is recommended to provide enablers in 3GPP services to support the use cases based on the recommended objectives and taking into account the candidate solutions.

Annex A: Encoding configuration parameters for viewport-independent video quality evaluation

This annex describes the HM configuration file used for the Bitstreams generation (here an example for QP=20).

```
#======== File I/O =====================
BitstreamFile : hm16.15_ra@Sequence_30p_3840x1920@qp=20.bin # Modified param
ReconFile : encoder_hm16.15_ra@Sequence_vr_30p_3840x1920@qp=20.yuv # Modified param

#======== Profile ================
Profile                       : main

#======== Unit definition ================
MaxCUWidth                    : 64          # Maximum coding unit width in pixel
MaxCUHeight                   : 64          # Maximum coding unit height in pixel
MaxPartitionDepth             : 4           # Maximum coding unit depth
QuadtreeTULog2MaxSize         : 5           # Log2 of maximum transform size for
                                            # quadtree-based TU coding (2...6)
QuadtreeTULog2MinSize         : 2           # Log2 of minimum transform size for
                                            # quadtree-based TU coding (2...6)
QuadtreeTUMaxDepthInter       : 3
QuadtreeTUMaxDepthIntra       : 3

#======== Coding Structure =============
IntraPeriod : 32 # Modified param
DecodingRefreshType           : 1           # Random Access 0:none, 1:CRA, 2:IDR, 3:Recovery Point SEI
GOPSize                       : 16           # GOP Size (number of B slice = GOPSize-1)
ReWriteParamSetsFlag          : 1           # Write parameter sets with every IRAP

IntraQPOffset                 : -3
LambdaFromQpEnable            : 1           # see JCTVC-X0038 for suitable parameters for IntraQPOffset, QPoffset, QPOffsetModelOff, QPOffsetModelScale when enabled
#        Type POC QPoffset QPOffsetModelOff QPOffsetModelScale CbQPoffset CrQPoffset QPfactor tcOffsetDiv2 betaOffsetDiv2 temporal_id #ref_pics_active #ref_pics reference pictures     predict deltaRPS
#ref_idcs reference idcs
Frame1:  B   16   1     0.0         0.0      0   0   1.0   0   0   0   2   3   -16 -24 -32       0
Frame2:  B    8   1    -4.8848      0.2061   0   0   1.0   0   0   1   2   3   -8  -16   8       1    8
4        1 1 0 1
Frame3:  B    4   4    -5.7476      0.2286   0   0   1.0   0   0   2   2   4   -4  -12   4  12    1    4
4        1 1 1 1
Frame4:  B    2   5    -5.90        0.2333   0   0   1.0   0   0   3   2   5   -2  -10   2   6  14   1    2
5        1 1 1 1 1
Frame5:  B    1   6    -7.1444      0.3      0   0   1.0   0   0   4   2   5   -1    1   3   7  15   1    1
6        1 0 1 1 1 1
Frame6:  B    3   6    -7.1444      0.3      0   0   1.0   0   0   4   2   5   -1   -3   1   5  13   1   -2
6        1 1 1 1 1 0
Frame7:  B    6   5    -5.90        0.2333   0   0   1.0   0   0   3   2   4   -2   -6   2  10      1   -3
6        0 1 1 1 1 0
Frame8:  B    5   6    -7.1444      0.3      0   0   1.0   0   0   4   2   5   -1   -5   1   3  11   1    1
5        1 1 1 1 1
Frame9:  B    7   6    -7.1444      0.3      0   0   1.0   0   0   4   2   5   -1   -3  -7   1   9   1   -2
6        1 1 1 1 1 0
Frame10: B   12   4    -5.7476      0.2286   0   0   1.0   0   0   2   2   3   -4  -12   4          1   -5
6        0 0 1 1 1 0
Frame11: B   10   5    -5.90        0.2333   0   0   1.0   0   0   3   2   4   -2  -10   2   6      1    2
4        1 1 1 1
Frame12: B    9   6    -7.1444      0.3      0   0   1.0   0   0   4   2   5   -1   -9   1   3   7   1    1
5        1 1 1 1 1
Frame13: B   11   6    -7.1444      0.3      0   0   1.0   0   0   4   2   5   -1   -3 -11   1   5   1   -2
6        1 1 1 1 1 0
Frame14: B   14   5    -5.90        0.2333   0   0   1.0   0   0   3   2   4   -2   -6 -14   2      1   -3
6        0 1 1 1 1 0
Frame15: B   13   6    -7.1444      0.3      0   0   1.0   0   0   4   2   5   -1   -5 -13   1   3   1    1
5        1 1 1 1 1
Frame16: B   15   6    -7.1444      0.3      0   0   1.0   0   0   4   2   5   -1   -3  -7 -15   1   1   -2
6        1 1 1 1 1 0

#=========== Motion Search =============
FastSearch                    : 1           # 0:Full search  1:TZ search
SearchRange                   : 256         # (0: Search range is a Full frame)
ASR                           : 1           # Adaptive motion search range
MinSearchWindow               : 64          # Minimum motion search window size for the adaptive window ME
BipredSearchRange             : 4           # Search range for bi-prediction refinement
HadamardME                    : 1           # Use of Hadamard measure for fractional ME
FEN                           : 1           # Fast encoder decision
FDM                           : 1           # Fast Decision for Merge RD cost

#======== Quantization =============
QP : 20 # Modified param
MaxDeltaQP                    : 0           # CU-based multi-QP optimization
MaxCuDQPDepth                 : 0           # Max depth of a minimum CuDQP for sub-LCU-level delta QP
DeltaQpRD                     : 0           # Slice-based multi-QP optimization
RDOQ                          : 1           # RDOQ
RDOQTS                        : 1           # RDOQ for transform skip
SliceChromaQPOffsetPeriodicity: 0           # Used in conjunction with Slice Cb/Cr QpOffsetIntraOrPeriodic. Use 0 (default) to disable periodic nature.
SliceCbQpOffsetIntraOrPeriodic: 0           # Chroma Cb QP Offset at slice level for I slice or for periodic inter slices as defined by SliceChromaQPOffsetPeriodicity. Replaces offset in the GOP table.
SliceCrQpOffsetIntraOrPeriodic: 0           # Chroma Cr QP Offset at slice level for I slice or for periodic inter slices as defined by SliceChromaQPOffsetPeriodicity. Replaces offset in the GOP table.

#=========== Deblock Filter ============
LoopFilterOffsetInPPS         : 1           # Dbl params: 0=varying params in SliceHeader, param = base_param + GOP_offset_param; 1 (default) =constant params in PPS, param = base_param)
LoopFilterDisable             : 0           # Disable deblocking filter (0=Filter, 1=No Filter)
LoopFilterBetaOffset_div2     : 0           # base_param: -6 ~ 6
LoopFilterTcOffset_div2       : 0           # base_param: -6 ~ 6
DeblockingFilterMetric        : 0           # blockiness metric (automatically configures deblocking parameters in bitstream). Applies slice-level loop filter offsets (LoopFilterOffsetInPPS and
LoopFilterDisable must be 0)

#=========== Misc. ============
InternalBitDepth : 8 # Modified param

#=========== Coding Tools =================
SAO                           : 1           # Sample adaptive offset  (0: OFF, 1: ON)
AMP                           : 1           # Asymmetric motion partitions (0: OFF, 1: ON)
TransformSkip                 : 1           # Transform skipping (0: OFF, 1: ON)
TransformSkipFast             : 1           # Fast Transform skipping (0: OFF, 1: ON)
SAOLcuBoundary                : 0           # SAOLcuBoundary using non-deblocked pixels (0: OFF, 1: ON)

#============ Slices ================
SliceMode                : 0                # 0: Disable all slice options.
                                            # 1: Enforce maximum number of LCU in an slice,
                                            # 2: Enforce maximum number of bytes in an 'slice'
                                            # 3: Enforce maximum number of tiles in a slice
SliceArgument            : 1500             # Argument for 'SliceMode'.
                                            # If SliceMode==1 it represents max. SliceGranularity-sized blocks per slice.
                                            # If SliceMode==2 it represents max. bytes per slice.
                                            # If SliceMode==3 it represents max. tiles per slice.

LFCrossSliceBoundaryFlag : 1                # In-loop filtering, including ALF and DB, is across or not across slice boundary.
                                            # 0:not across, 1: across

#============ PCM ================
PCMEnabledFlag                      : 0              # 0: No PCM mode
PCMLog2MaxSize                      : 5              # Log2 of maximum PCM block size.
PCMLog2MinSize                      : 3              # Log2 of minimum PCM block size.
PCMInputBitDepthFlag                : 1              # 0: PCM bit-depth is internal bit-depth. 1: PCM bit-depth is input bit-depth.
PCMFilterDisableFlag                : 0              # 0: Enable loop filtering on I_PCM samples. 1: Disable loop filtering on I_PCM samples.

#============ Tiles ================
TileUniformSpacing                  : 0              # 0: the column boundaries are indicated by TileColumnWidth array, the row boundaries are indicated by TileRowHeight array
                                                     # 1: the column and row boundaries are distributed uniformly
NumTileColumnsMinus1                : 0              # Number of tile columns in a picture minus 1
TileColumnWidthArray                : 2 3            # Array containing tile column width values in units of CTU (from left to right in picture)
NumTileRowsMinus1                   : 0              # Number of tile rows in a picture minus 1
TileRowHeightArray                  : 2              # Array containing tile row height values in units of CTU (from top to bottom in picture)

LFCrossTileBoundaryFlag             : 1              # In-loop filtering is across or not across tile boundary.
                                                     # 0: not across, 1: across

#============ WaveFront ================
WaveFrontSynchro                    : 0              # 0:  No WaveFront synchronisation (WaveFrontSubstreams must be 1 in this case).
                                                     # >0: WaveFront synchronises with the LCU above and to the right by this many LCUs.

#=========== Quantization Matrix =================
ScalingList                   : 0                    # ScalingList 0 : off, 1 : default, 2 : file read
ScalingListFile               : scaling_list.txt     # Scaling List file name. If file is not exist, use Default Matrix.

#============ Lossless ================
TransquantBypassEnableFlag : 0                       # Value of PPS flag.
CUTransquantBypassFlagForce: 0                       # Force transquant bypass mode, when transquant_bypass_enable_flag is enabled

#============ Rate Control ======================
RateControl : 0 # Modified param
TargetBitrate : 0 # Modified param
KeepHierarchicalBit                 : 2              # Rate control: 0: equal bit allocation; 1: fixed ratio bit allocation; 2: adaptive ratio bit allocation
LCULevelRateControl                 : 1              # Rate control: 1: LCU level RC; 0: picture level RC
RCLCUSeparateModel                  : 1              # Rate control: use LCU level separate R-lambda model
InitialQP : 20 # Modified param
RCForceIntraQP                      : 0              # Rate control: force intra QP to be equal to initial QP


#============ Added params ======================
FramesToBeEncoded : 900
SourceHeight : 1920
InputFile : /Sequences/360VR/Sequence_30p_3840x1920.yuv
InputBitDepth : 8
SEIDecodedPictureHash : 1
Level : 5
SourceWidth : 3840
FrameRate : 30
```

Annex B:
Test instructions for viewport-independent video quality evaluation

NOTE: These instructions were made available in both French and English (tests run in France).

Welcome to Orange Labs,

You are about to take part in an evaluation of the quality of video sequences (video only, no sound) in Virtual Reality environment. For each sequence, you have to assess the overall video quality for the entire duration. In addition, the visibility level of visual degradations can be used to assess the video quality.

Five various clips of about 10-15 seconds long have been selected. Each of them has been treated with different processes indicated by the letters A, B, C, …. The reference clip ("R" button) has not been processed.

You may view each sequence in any order and repeat it as many time as you want (at least one time entire duration) using the "Play" button. After the visualization of each sequence, you can report your opinion moving the slider on the quality scale (numbered from 0 to 100) according to quality labels "Bad", "Poor", "Fair", "Good", "Excellent".

The scoring can be modified or refine at any time. You have to score the sequences of one clip before to assess the next clip pressing the "VALIDATE" button.

At the end of the last sequence of the last clip, the "END" button becomes active. Press it to complete the test session.

In order for you to get used to the equipment (adjustment of the head-mounted-display, getting started with navigation and rating controls) you will be assisted during a preliminary training session.

Thank for your participation.

Annex C:
Change history

Change history							
Date	**Meeting**	**TDoc**	**CR**	**Rev**	**Cat**	**Subject/Comment**	**New version**
06-2017	SA#76	SP-170331				Presented to TSG SA#76 for information	1.0.0
06-2017	SA4#94	S4-170651				Completion of the TR with the inclusion of scope, media formats description, audio and video quality subjective tests, motion to sound analysis, use cases completion, QoE metrics analysis and a set of conclusions	1.1.0
06-2017	SA4#94	S4-170752				Editorial modifications	1.2.0
09-2017	SA#77	SP-170619				Presented to TSG SA#77 for approval	2.0.0
09-2017						Version 15.0.0	15.0.0
12-2017	SA#78	SP-170832	0003	-	B	Findings and Conclusions from study on 3GPP codecs for VR audio	15.1.0
03-2018	SA#79	SP-180030	0004	1	F	Corrections to subjective listening tests	15.2.0

History

Document history		
V15.2.0	July 2018	Publication