ETSI TS 126 118 V15.3.0 (2021-01)

5G;
Virtual Reality (VR) profiles for streaming applications
(3GPP TS 26.118 version 15.3.0 Release 15)

Reference
RTS/TSGS-0426118vf30

Keywords
5G

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

Important notice

The present document can be downloaded from:
http://www.etsi.org/standards-search

Copyright Notification

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.

Legal Notice

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

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

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

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**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 Specification 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

The present document provides technologies for interoperable Virtual Reality services with focus on streaming and consumption.

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.

Suitable media formats for providing immersive experiences are specified to enable Virtual Reality Services in the context of 3GPP bearer and user services.

1 Scope

The present document defines interoperable formats for Virtual Reality for streaming services. Specifically, the present document defines operation points, media profiles and presentation profiles for Virtual Reality. The present document builds on the findings and conclusions in TR 26.918 [2].

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] 3GPP TR 26.918: "Virtual Reality (VR) media services over 3GPP".

[3] Recommendation ITU-R BT.709-6 (06/2015): "Parameter values for the HDTV standards for production and international programme exchange".

[4] Recommendation ITU-R BT.2020-2 (10/2015): "Parameter values for ultra-high definition television systems for production and international programme exchange".

[5] Recommendation ITU-T H.264 (04/2017): "Advanced video coding for generic audiovisual services" | ISO/IEC 14496-10:2014: "Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding".

[6] Recommendation ITU-T H.265 (02/2018): "High efficiency video coding" | ISO/IEC 23008-2:2018: "High Efficiency Coding and Media Delivery in Heterogeneous Environments – Part 2: High Efficiency Video Coding".

[7] void.

[8] 3GPP TS 26.247: "Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH)".

[9] ISO/IEC 14496-15: "Information technology - Coding of audio-visual objects - Part 15: Carriage of network abstraction layer (NAL) unit structured video in ISO base media file format".

[10] ISO/IEC 23001-8: "Information technology -- MPEG systems technologies -- Part 8: Coding-independent code points".

[11] Recommendation ITU-R BT.2100-1: "Image parameter values for high dynamic range television for use in production and international programme exchange".

[12] 3GPP TS 26.116: "Television (TV) over 3GPP services; Video profiles".

[13] ISO/IEC 23090-2: "Coded representation of immersive media -- Part 2: Omnidirectional media format".

[14] ISO/IEC DIS 23091-2: "Information technology -- Coding-independent code points -- Part 2: Video".

[15] 3GPP TS 26.260: "Objective test methodologies for the evaluation of immersive audio systems".

[16] 3GPP TS 26.259: "Subjective test methodologies for the evaluation of immersive audio systems".

[17] ISO/IEC 14496-12: "Information technology -- Coding of audio-visual objects -- Part 12: ISO base media file format".

[18] ISO/IEC 23009-1: "Information technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats".

[19] ISO/IEC 23008-3:2015: "Information technology -- High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio", ISO/IEC 23008-3:2015/Amd2:2016: "MPEG-H 3D Audio File Format Support ", ISO/IEC 23008-3:2015/Amd 3:2017: "MPEG-H 3D Audio Phase 2", ISO/IEC 23008-3:2015/Amd 5: "Audio metadata enhancements".

[20] IETF RFC 6381: "The 'Codecs' and 'Profiles' Parameters for "Bucket" Media Types", R. Gellens, D. Singer, P. Frojdh, August 2011.

[21] AES69-2015: "AES standard for file exchange - Spatial acoustic data file format", 2015.

3 Definitions, symbols and abbreviations

3.1 Definitions

For the purposes of the present document, the terms and definitions given in 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 TR 21.905 [1].

bitstream: a bitstream that conforms to a video encoding format and certain Operation Point.

field of view: the extent of visible area expressed with vertical and horizontal angles, in degrees in the 3GPP 3DOF reference system.

operation point: a collection of discrete combinations of different content formats including spatial and temporal resolutions, colour mapping, transfer functions, rendering metadata and the encoding format.

pose: position derived by the head tracking sensor expressed by (azimuth, elevation, tilt angle).

receiver: a receiver that can decode and render any bitstream that is conforming to a certain Operation Point.

viewport: the part of the 3DOF content to render based on the pose and the field of view.

3.2 Symbols

For the purposes of the present document, the following symbols apply:

α	yaw of the 3GPP 3DOF coordinate system
β	pitch of the 3GPP 3DOF coordinate system
γ	roll of the 3GPP 3DOF coordinate system
ϕ	azimuth of the 3GPP 3DOF coordinate system
θ	elevation of the 3GPP 3DOF coordinate system

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in 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 TR 21.905 [1].

3DOF	3 Degrees of freedom
ACN	Ambisonics Channel Number
API	Application Programming Interface
AVC	Advanced Video Coding
BMFF	Base Media File Format
BRIR	Binaural Room Impulse Response
CMP	Cube-Map Projection

CIBR	Common Informative Binaural Renderer
DASH	Dynamic Adaptive Streaming over HTTP
DRC	Dynamic Range Control
EOTF	Electro-Optical Transfer Function
ERP	EquiRectangular Projection
ESD	Equivalent Spatial Domain
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FOA	First Order Ambisonics
FOV	Field Of View
GPU	Graphics Processing Unit
HDR	High Dynamic Range
HDTV	High Definition TeleVision
HEVC	High Efficiency Video Coding
HMD	Head Mounted Display
HOA	High Order Ambisonics
HRD	Hypothetical Reference Decoder
HRIR	Head-Related Impulse Responses
HRTF	Head-Related Transfer Function
HTTP	HyperText Transfer Protocol
IFFT	Inverse FFT
IRFFT	Inverse RFFT
MAE	MPEG-H Audio Metadata information
MHAS	MPEG-H Audio Stream
MIME	Multipurpose Internet Mail Extensions
MPD	Media Presentation Description
MPEG	Moving Pictures Experts Group
NAL	Network Abstraction Layer
OMAF	Omnidirectional MediA Format
PCM	Pulse Code Modulation
RAP	Random Access Point
RFFT	Real FFT
RWP	Region-Wise Packing
SDR	Standard Dynamic Range
SEI	Supplemental Enhancement Information
SN3D	Schmidt semi-normalisation
SOFA	Spatially Oriented Format for Acoustics
SPS	Sequence Parameter Set
SRQR	Spherical Region-wise Quality Ranking
VCL	Video Coding Layer
VST	Virtual Studio Technology
VUI	Video Usability Information
VR	Virtual Reality

4 Architectures and Interfaces for Virtual Reality

4.1 Definitions and Reference Systems

4.1.1 Overview

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, assumes 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 as shown in Figure 4.1-1.

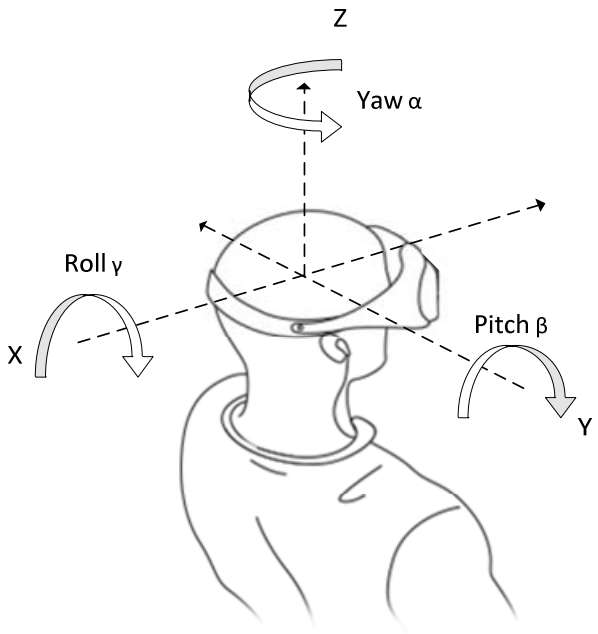

Figure 4.1-1: Reference System

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. Sensors typically are able to track the user's pose in the reference system. Additional means to interact with the virtual reality simulation may be provided but are not strictly necessary.

VR users are expected to be able to look around from a single observation point in 3D space defined by either a producer or the position of one or multiple capturing devices. When VR media including video and audio is consumed with a head-mounted display or a smartphone, only the area of the spherical video that corresponds to the user's viewport is rendered, as if the user were in the spot where the video and audio were captured.

This ability to look around and listen from a *centre point* in 3D space is defined as 3 degrees of freedom (3DOF). According to the figure 4.1-1:

- tilting side to side on the X-axis is referred to as *Roll*ing, also expressed as γ
- tilting forward and backward on the Y-axis is referred to as *Pitch*ing, also expressed as β
- turning left and right on the Z-axis is referred to as *Yaw*ing, also expressed as α

It is worth noting that this *centre point* 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 3 degrees of freedom. For a full 3DOF VR experience, such video content may be combined with simultaneously captured audio, binaurally rendered with an appropriate Binaural Room Impulse Response (BRIR). The third relevant aspect is the interactivity: Only if the content is presented to the user in such a way that the movements are instantaneously reflected in the rendering , then the user will perceive a full immersive experience. For details on immersive rendering latencies, refer to TR 26.918 [2].

4.1.2 3GPP 3DOF Coordinate System

The coordinate system is specified for defining the sphere coordinates azimuth (ϕ) and elevation (θ) for identifying a location of a point on the unit sphere, as well as the rotation angles yaw (α), pitch (β), and roll (γ). The origin of the coordinate system is usually the same as the centre point of a device or rig used for audio or video acquisition as well as the position of the user's head in the 3D space in which the audio or video are rendered. Figure 4.1-2 specifies principal axes for the coordinate system. The X axis is equal to back-to-front axis, Y axis is equal to side-to-side (or lateral) axis, and Z axis is equal to vertical (or up) axis. These axis map to the reference system in Figure 4.1-1.

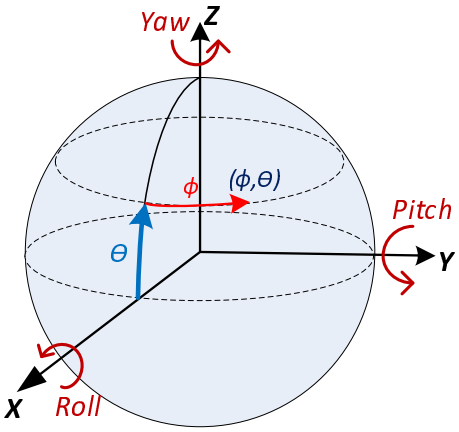

Figure 4.1-2: Coordinate system

Signals defined in the present document are represented in a spherical coordinate space in angular coordinates (ϕ, Θ) for use in omnidirectional video and 3D audio. The viewing and listing perspective are from the origin sensing/looking/hearing outward toward the inside of the sphere. Even though a spherical coordinate is generally represented by using radius, elevation, and azimuth, it assumes that a unit sphere is used for capturing and rendering of VR media. Thus, a location of a point on the unit sphere is identified by using the sphere coordinates azimuth (ϕ) and elevation (θ). The spherical coordinates are defined so that ϕ is the azimuth and Θ is the elevation. As depicted in Figure 4.1-2, the coordinate axes are also used for defining the rotation angles yaw ($\alpha\alpha$), pitch ($\beta\beta$), and roll ($\gamma\gamma$). The angles increase clockwise when looking from the origin towards the positive end of an axis. The value ranges of azimuth, yaw, and roll are all −180.0, inclusive, to 180.0, exclusive, degrees. The value range of elevation and pitch are both −90.0 to 90.0, inclusive, degrees.

Depending on the applications or implementations, not all angles may be necessary or available in the signal. The 360 video may have a restricted *coverage* as shown in Figure 4.1-3. When the video signal does not cover the full sphere, the coverage information is described by using following parameters:

- *centre azimuth*: specifies the azimuth value of the centre point of sphere region covered by the signal.
- *centre elevation*: specifies the elevation value of the centre of sphere region.
- *azimuth range*: specifies the azimuth range through the centre point of the sphere region.
- *elevation range*: specifies the elevation range through the centre point of the sphere region.
- *tilt angle*: indicates the amount of tilt of a sphere region, measured as the amount of rotation of the sphere region along the axis originating from the origin passing through the centre point of the sphere region, where the angle value increases clockwise when looking from the origin towards the positive end of the axis.

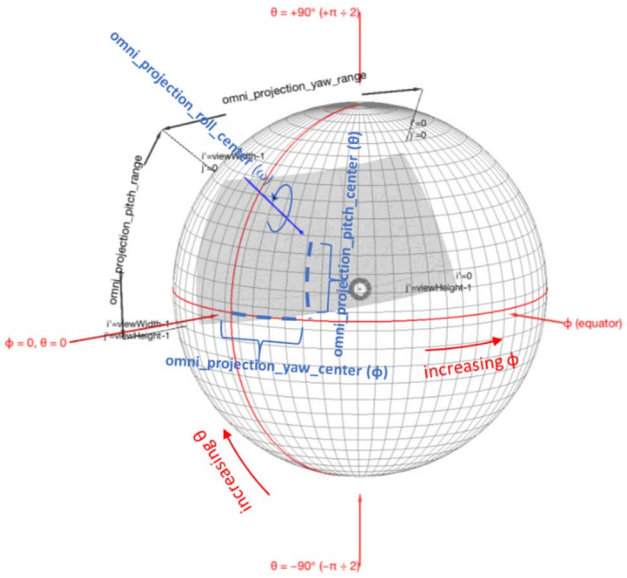

Figure 4.1-3: Restricted coverage of the sphere region covered by the cropped output picture with omni_projection_{yaw | pitch | roll}_center the center of the coverage region.

For video, such a centre point may exist for each eye, referred to as *stereo* signal, and the video consists of three color components, typically expressed by the luminance (Y) and two chrominance components (U and V).

The coordinate systems for all media types are assumed to be aligned in 3GPP 3DOF coordinate system. Within this coordinate system, the *pose* is expressed by a triple of azimuth, elevation, and tilt angle characterizing the head position of a user consuming the audio-visual content. The pose is generally dynamic, and the information may be provided through sensors in a frequently sampled version.

The *field of view (FoV)* of a rendering device is static and defined in two dimensions, the horizontal and vertical FoV, each in units of degrees in the angular coordinates (ϕ, θ). The pose together with the field of view of the device enables the system to generate the user viewport, i.e., the presented part of the content at a specific point in time.

4.1.3 Video Signal Representation

Commonly used video encoders cannot directly encode spherical videos, but only 2D textures. However, there is a significant benefit to reuse conventional 2D video encoders. Based on this, Figure 4.1-4 provides the basic video signal representation in the context of omnidirectional video in the context of the present document. By pre-processing, the spherical video is mapped to a 2D texture. The 2D texture is encoded with a regular 2D video encoder and the VR rendering metadata (i.e. the data describing the mapping from the spherical coordinate to the 2D texture) is encoded and provided along with the video bitstream, such that at the receiving end the inverse process can be applied to reconstruct the spherical video.

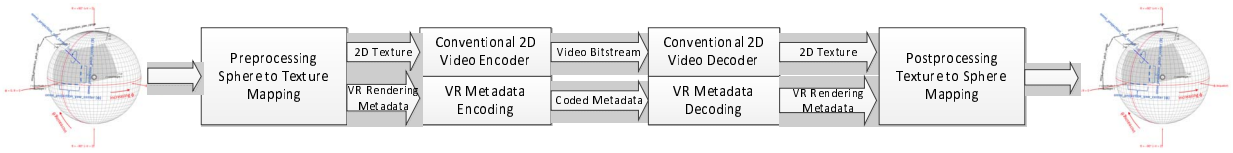

Figure 4.1-4: Video Signal Representation

Mapping of a spherical picture to a 2D texture signal is illustrated in Figure 4.1-5. The most commonly used mapping from spherical to 2D is the equirectangular projection (ERP) mapping. The mapping is bijective, i.e. it may be expressed in both directions.

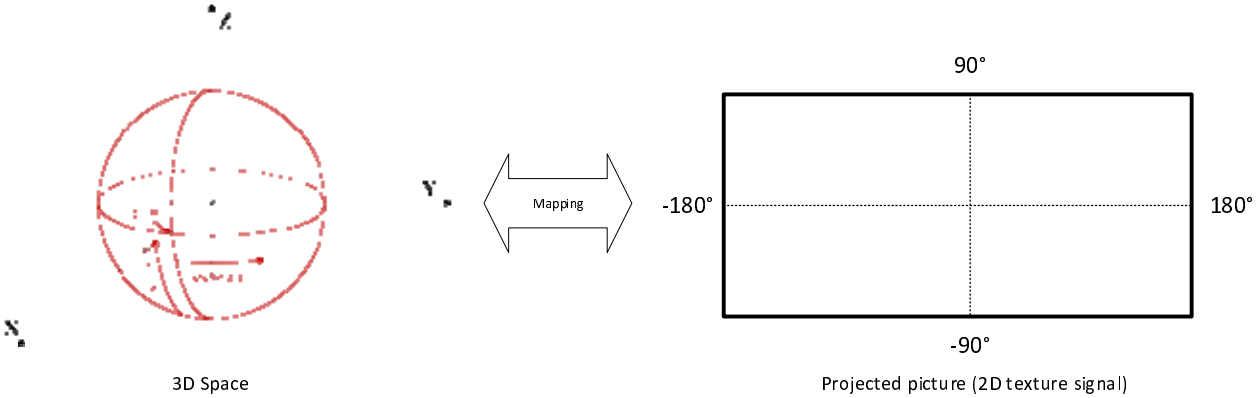

Figure 4.1-5: Examples of Spherical to 2D mappings

Following the definitions in clause 4.1.2, the mapping of the color samples of 2D texture images onto a spherical coordinate space in angular coordinates (ϕ, θ) for use in omnidirectional video applications for which the viewing perspective is from the origin looking outward toward the inside of the sphere. The spherical coordinates are defined so that ϕ is the azimuth and θ is the elevation.

Assume a 2D texture with `pictureWidth` and `pictureHeight`, being the width and height, respectively, of a monoscopic projected luma picture, in luma samples and the center point of a sample location `(i,j)` along the horizontal and vertical axes, respectively, then for the *equirectangular* projection the sphere coordinates (ϕ, θ) for the luma sample location, in degrees, are given by the following equations:

$$\phi = (0.5 - i \div pictureWidth) * 360$$
$$\theta = (0.5 - j \div pictureHeight) * 180$$

Whereas ERP is commonly used for production formats, other mappings may be applied, especially for distribution. The present document also introduces cubemap projection (CMP) for distribution in clause 5. In addition to regular projection, other pre-processing may be applied to the spherical video when mapped into 2D textures. Examples include region-wise packing, stereo frame packing or rotation. The present document defines different pre- and post-processing schemes in the context of video rendering schemes.

4.1.4 Audio Signal Representation

Audio for VR can be produced using three different formats. These are broadly known as channels-, objects- and scene-based audio formats. Audio for VR can use any one of these formats or a hybrid of these (where all three formats are used to represent the spherical soundfield). The audio signal representation model is shown in Figure 4.1-6.

The present document expects that an audio encoding system is capable to produce suitable audio bitstreams that represent a well-defined audio signal in the reference system as defined in clause 4.1.1. The coding and carriage of the VR Audio Rendering Metadata is expected to be defined by the VR Audio Encoding system. The VR Audio Receiving system is expected to be able to use the VR Audio Bitstream to recover audio signals and VR Audio Rendering metadata. Both signals, audio signals and metadata, are well-defined by the media profile, such that different audio rendering systems may be used to render the audio based on the decoder audio signals, VR audio rendering metadata and the user position.

In the present document, all media profiles are defined such that for each media profile at least one Audio Rendering System is defined as a reference renderer and additional Audio Rendering systems may be defined. The audio rendering system is described based on well-defined output of the VR Audio decoding system.

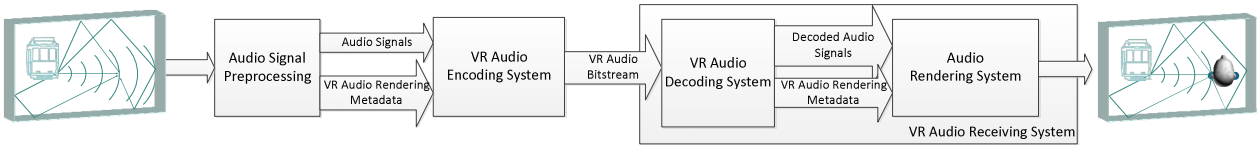

Figure 4.1-6: Audio Signal Representation