# ETSI TS 126 250 V18.0.0 (2024-07)


**LTE;**
**5G;**
**Codec for Immersive Voice and Audio Services**
**- General overview**
**(3GPP TS 26.250 version 18.0.0 Release 18)**

Reference
RTS/TSGS-0426250vi00

Keywords
5G,LTE

***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 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

# Intellectual Property Rights

## Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are 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 Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, 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.

**DECT™**, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

# 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 https://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

Intellectual Property Rights ..... 2
Legal Notice ..... 2
Modal verbs terminology ..... 2
Foreword ..... 4
1 Scope ..... 6
2 References ..... 6
3 Definitions of terms, symbols and abbreviations ..... 6
3.1 Terms ..... 6
3.2 Symbols ..... 7
3.3 Abbreviations ..... 7
4 General ..... 7
5 Transcoding functions ..... 9
6 Rendering ..... 9
7 C-code ..... 10
8 Test sequences ..... 10
9 Discontinuous transmission (DTX) ..... 10
10 Error concealment of lost frames ..... 10
11 Frame structure ..... 11
12 RTP Payload Format ..... 11
13 Jitter Buffer Management ..... 11
14 Performance characterization ..... 11
**Annex A (informative): Change history ..... 12**
History ..... 13

# 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.

In the present document, modal verbs have the following meanings:

**shall** indicates a mandatory requirement to do something

**shall not** indicates an interdiction (prohibition) to do something

The constructions "shall" and "shall not" are confined to the context of normative provisions, and do not appear in Technical Reports.

The constructions "must" and "must not" are not used as substitutes for "shall" and "shall not". Their use is avoided insofar as possible, and they are not used in a normative context except in a direct citation from an external, referenced, non-3GPP document, or so as to maintain continuity of style when extending or modifying the provisions of such a referenced document.

**should** indicates a recommendation to do something

**should not** indicates a recommendation not to do something

**may** indicates permission to do something

**need not** indicates permission not to do something

The construction "may not" is ambiguous and is not used in normative elements. The unambiguous constructions "might not" or "shall not" are used instead, depending upon the meaning intended.

**can** indicates that something is possible

**cannot** indicates that something is impossible

The constructions "can" and "cannot" are not substitutes for "may" and "need not".

**will** indicates that something is certain or expected to happen as a result of action taken by an agency the behaviour of which is outside the scope of the present document

**will not** indicates that something is certain or expected not to happen as a result of action taken by an agency the behaviour of which is outside the scope of the present document

**might** indicates a likelihood that something will happen as a result of action taken by some agency the behaviour of which is outside the scope of the present document

**might not** indicates a likelihood that something will not happen as a result of action taken by some agency the behaviour of which is outside the scope of the present document

In addition:

**is** (or any other verb in the indicative mood) indicates a statement of fact

**is not** (or any other negative verb in the indicative mood) indicates a statement of fact

The constructions "is" and "is not" do not indicate requirements.

# 1 Scope

The present document is an introduction to the audio processing parts and auxiliary functions of the codec for Immersive Voice and Audio Services (IVAS codec). A general overview of the audio processing and auxiliary functions is given, with reference to the documents where each function is specified in detail.

# 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 TS 26.441: "Codec for Enhanced Voice Services (EVS); General Overview".

[3] 3GPP TS 26.253: "Codec for Immersive Voice and Audio Services (IVAS); Detailed Algorithmic Description incl. RTP payload format and SDP parameter definitions".

[4] 3GPP TS 26.254: "Codec for Immersive Voice and Audio Services (IVAS); Rendering".

[5] 3GPP TS 26.251: "Codec for Immersive Voice and Audio Services (IVAS); C code (fixed-point)".

[6] 3GPP TS 26.258: "Codec for Immersive Voice and Audio Services (IVAS); C code (floating point)".

[7] 3GPP TS 26.252: "Codec for Immersive Voice and Audio Services (IVAS); Test Sequences".

[8] 3GPP TS 26.255: "Codec for Immersive Voice and Audio Services (IVAS); Error concealment of lost packets".

[9] 3GPP TS 26.256: "Codec for Immersive Voice and Audio Services (IVAS); Jitter Buffer Management".

[10] 3GPP TR 26.997: ["Codec for Immersive Voice and Audio Services (IVAS); Performance characterization"].

[11] 3GPP TS 26.131: "Terminal acoustic characteristics for telephony; Requirements".

[12] 3GPP TS 26.261: "Terminal audio quality performance requirements for immersive audio services".

[[13] 3GPP TS 26.114: "IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction".]

# 3 Definitions of terms, symbols and abbreviations

## 3.1 Terms

For the purposes of the present document, the terms 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].

## 3.2 Symbols

void

## 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].

| | |
|---|---|
| EVS | Enhanced Voice Services |
| IVAS | Immersive Voice and Audio Services |
| JBM | Jitter Buffer Management |
| MASA | Metadata-Assisted Spatial Audio |
| SID | Silence Insertion Descriptor |

# 4 General

The codec for Immersive Voice and Audio Services is part of a framework comprising besides encoder and decoder, renderer and a number of auxiliary functions associated with the support of stereo and immersive audio formats.

The IVAS codec is an extension of the 3GPP Enhanced Voice Services (EVS) codec; it provides full and bit exact EVS codec functionality for mono speech/audio signal input.

On top of that the IVAS codec is optimized for encoding and decoding of stereo and immersive audio formats, using tools such as Single Channel Element (SCE) coding, Channel Pair Element (CPE) coding and multi-channel coding by means of the Multi-channel Coding Tool (MCT). The stereo modes comprise a hybrid time-domain/DFT-domain/MDCT-domain coding scheme including inter channel alignment (ICA). Immersive audio formats comprise multi-channel audio (5.1, 5.1.2, 5.1.4, 7.1, 7.1.4 setups), scene-based audio (Ambisonics up to order 3), metadata-assisted spatial audio (MASA), and object-based audio (Independent Stream with Metadata (ISM) up to 4 ISMs). In addition, the following combined immersive audio formats are supported: object-based audio with scene-based audio (OSBA, up to 4 ISMs with Ambisonics) and object-based audio with metadata-assisted spatial audio (OMASA, up to 4 ISMs with MASA).

The codec features VAD/DTX/CNG for rate efficient stereo and immersive conversational voice transmissions, an error concealment mechanism to combat the effects of transmission errors and lost packets. Jitter buffer management is also provided.

The IVAS codec operates on 20 ms audio frames. It is capable of switching its bit rate upon command instantly at (active) frame boundaries.

A reference configuration where relevant interface signals and various relevant send side processing functions are identified is given in Figure 1. A corresponding reference configuration for receive side identifying relevant interface signals and processing functions is given in Figure 2. In the figures, the relevant specifications for each function are also indicated.

In Figures 1 & 2, the UE Send and Receive Audio processing are included, to show the complete path between the audio input/output in the User Equipment (UE) and a possible digital interface in network (all excluding A/D or D/A conversion). The detailed specification of the audio parts is not within the scope of the present document. These aspects are only considered to the extent to highlight that the function of the audio parts and the operation of the IVAS codec are closely dependent on each other.

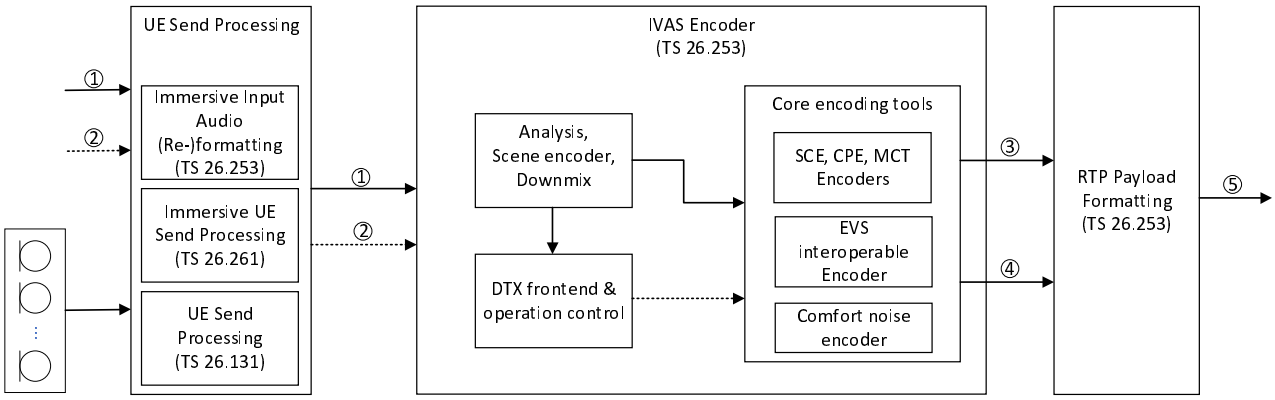


**Figure 1: Overview of audio processing functions - Transmit Side**

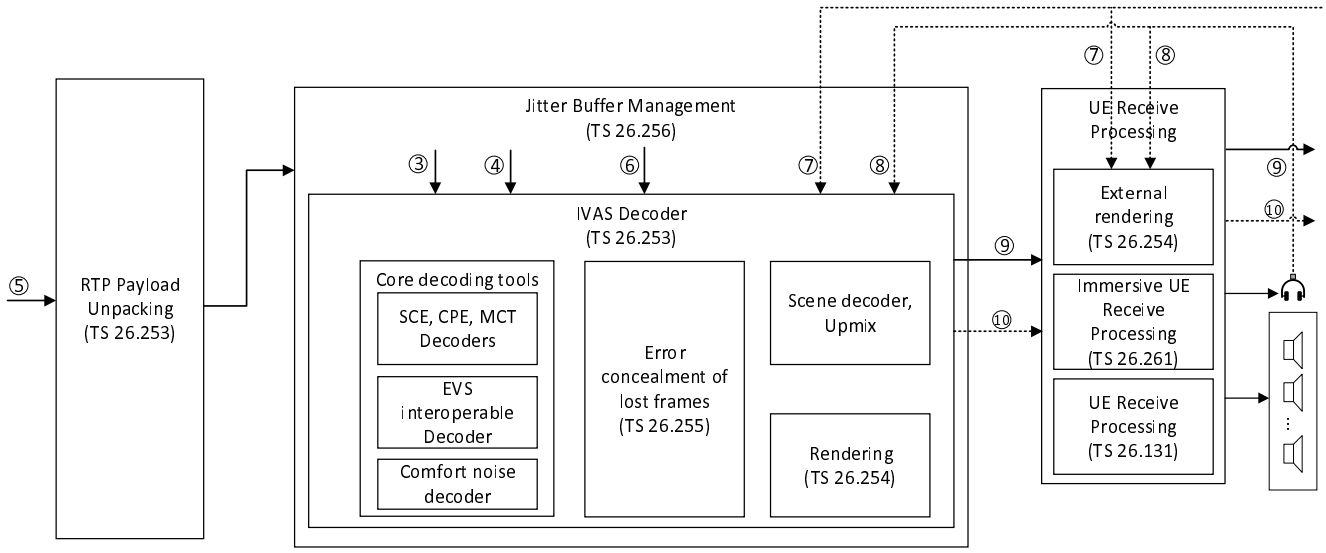


**Figure 2: Overview of audio processing functions - Receive Side**

Interfaces:

1: Audio input channels (16-bit linear PCM, sampled at 8 (only EVS), 16, 32, or 48 kHz)

2: Metadata associated with input audio

3: Encoded audio frames (50 frames/s), number of bits depending on IVAS codec mode

4: Encoded Silence Insertion Descriptor (SID) frames

5: RTP Payload packets

6: Lost Frame Indicator (BFI)

7: Renderer config data

8: Head-tracker pose information and scene orientation control data

9: Audio output channels (16-bit linear PCM, sampled at 8 (only EVS), 16, 32, or 48 kHz)

10: Metadata associated with output audio

# 5 Transcoding functions

An algorithmic description of the IVAS codec is provided in [3].

As shown in Figures 1 & 2, the audio encoder takes its input, and can produce an output at the decoder/renderer, in various audio output formats. Input and output audio signals consist of one or multiple constituent channels of the respective audio format and in some cases metadata. The constituent channels are in the form of 16-bit uniform Pulse Code Modulated (PCM) signals at sampling frequencies of 8 kHz (only EVS interoperable coding), 16 kHz, 32 kHz or 48 kHz. The audio may typically originate from and terminate within the audio part of the UE or from the network side.

The detailed mapping between blocks of input audio to encoded blocks (in which the number of bits depends on the presently used codec mode) and from these to output blocks of reconstructed audio is described in [3]. The supported bit rates of the EVS interoperable coding are provided in [2]. Stereo and immersive audio coding is offered at the following discrete bit rates [kbps]: 13.2, 16.4, 24.4, 32, 48, 64, 80, 128, 160, 192, 256, 384, and 512, with supported bit rate ranges and supported source-controlled rate operation (DTX) listed in Table 1.

**Table 1: Ranges of supported source codec bit rates for stereo and immersive audio coding modes of the IVAS codec**

| Input audio format | Range of supported bit rates (kbps) | Source Controlled Rate Operation Available |
|---|---|---|
| Stereo, binaural audio(1 | 13.2 – 256 | Yes, up to 256 kbps |
| Scene-based audio (Ambisonics: FOA, HOA2, HOA3) | 13.2 – 512 | Yes, up to 80 kbps |
| Metadata-assisted spatial audio (MASA) | 13.2 – 512 | Yes, up to 512 kbps |
| Object-based audio (ISM)(2 | 13.2 – 512(3 | Yes, up to 512 kbps |
| Multi-channel audio | 13.2 – 512 | No |

Notes:

(1 A head-trackable binaural audio format at rates ranging from 256 kbps to 768 kbps is additionally supported as intermediate split rendering representation.

(2 Combined input audio format combining scene-based audio with ISM (OSBA) is supported. Metadata-assisted spatial audio with ISM (OMASA) is supported.

(3 13.2 kbps – 128 kbps for 1 ISM, 16.4 kbps – 256 kbps for 2 ISMs, 24.4 kbps – 384 kbps for 3 ISMs, resp. 24.4 kbps – 512 kbps for 4 ISMs.

For stereo input, a downmix tool is provided to generate a mono signal for EVS interoperable stream without additional delay.

# 6 Rendering

IVAS rendering is the process to generate the IVAS audio output in the same or a different audio format than the input format, whereby in some cases, such as stereo-to-stereo, there is no particular rendering processing other than the decoding. The IVAS decoder provides integrated binaural rendering functionality for headphone reproduction including head-tracking and scene orientation control and integrated rendering for loudspeaker reproduction. IVAS binaural rendering also supports a reverberation effect. There is also the possibility to feed the IVAS decoder output to a customized external renderer while bypassing the integrated renderer. A special feature of the renderer is that it supports split operation with pre-rendering and transcoding to a head-trackable intermediate representation that can be transmitted to a post-rendering end-device. This enables moving a large part of the processing load and memory requirements for IVAS decoding and rendering to a (more) capable node/UE while offloading the final rendering end-device.

IVAS rendering can also be operated stand-alone, i.e., without prior IVAS encoding/decoding of the input audio signal.

IVAS rendering is described in [3] and [4].

# 7 C-code

The C-code of the IVAS codec including VAD/DTX/CNG functionality, rendering, error concealment of lost frames, Jitter Buffer Manager (JBM) are described in [5] for fixed point arithmetic operation and are described in [6] for floating point arithmetic operation.

The C-code is mandatory.

# 8 Test sequences

A set of digital test sequences is specified in [7], thus enabling the verification of compliance, i.e., bit-exactness, to a high degree of confidence.

The IVAS encoder, decoder and renderer (see Figures 1 and 2) are defined in bit-exact arithmetic. Consequently, they shall react on being presented with a given input sequence always with the corresponding bit-exact output sequence, provided that the internal state variables are also always exactly in the same state at the beginning of the test.

The input test sequences shall produce the corresponding output test sequences, provided that the codec is operated from reset state.

# 9 Discontinuous transmission (DTX)

The discontinuous transmission (DTX) functionality of the IVAS codec including voice activity detection (VAD) and comfort noise generation (CNG) is defined in [3]. DTX functionality is supported for IVAS operation modes, i.e., audio formats and bit rates, that are especially optimized for efficient stereo and immersive conversational voice transmissions (see Table 1).

During a normal telephone conversation, the participants alternate so that, on the average, each direction of transmission is occupied about 50 % of the time. Source-controlled rate operation is a mode of operation where the encoder encodes speech frames containing only background noise with a lower bit rate than normally used for encoding speech. A network may adapt its transmission scheme to take advantage of the varying bit rate. This may be done for the following two purposes:

1) In the UE, battery life will be prolonged, or a smaller battery could be used for a given operational duration.

2) The average required bit rate is reduced, leading to a more efficient transmission with decreased load and hence increased capacity.

The following functions are provided by the IVAS codec for the source-controlled rate operation:

- a Voice Activity Detector (VAD) or more accurately Sound Activity Detector (SAD) on the TX side;
- evaluation of the background acoustic noise on the TX side, in order to transmit characteristic parameters to the RX side;
- generation of comfort noise on the RX side during periods when no normal speech frames are received.

The transmission of comfort noise information to the RX side is achieved by means of a Silence Descriptor (SID) frame, which is sent at regular intervals.

The non-EVS IVAS SID frames are represented by 104 bits.

# 10 Error concealment of lost frames

The IVAS codec error concealment of erroneous or lost frames is described in [3] and [8].

Frames may be erroneous due to transmission errors or frames may be lost or delayed due to packet loss in a transport network.

In order to mask the effect of erroneous/lost frames, the decoder shall be informed about such frames, which shall initiate error concealment actions leading to generation of substitution frames for the decoded/rendered audio output.

# 11 Frame structure

The IVAS codec frame structure is described in [3].

# 12 RTP Payload Format

The IVAS coder RTP Payload Format for media handling and interaction is described in [3].

# 13 Jitter Buffer Management

The IVAS codec Jitter Buffer Management is described in [9].

# 14 Performance characterization

The IVAS codec performance characterization is described in [10].

# Annex A (informative): Change history

| Change history | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Date** | **Meeting** | **TDoc** | **CR** | **Rev** | **Cat** | **Subject/Comment** | **New version** |
| 08-2023 | Post SA4#124 telco | S4aA230093 | | | | Presented to Audio SWG for information | 0.0.1 |
| 08-2023 | SA4#125 telco | S4-231250 | | | | Presented to SA4 meeting as part of IVAS codec selection deliverables | 0.1.0 |
| 08-2023 | SA4#125 | S4-231441 | | | | Presented to SA4 meeting for agreement to be presented for information to TSG SA#101 | 0.2.0 |
| 09-2023 | SA#101 | | | | | Version 1.0.0 created by MCC | 1.0.0 |
| 04-2024 | SA4#127-e-bis | S4-240710 | | | | Implementation of SA4-agreed pCR on adding ISAR track-a split rendering feature | |
| 05-2024 | SA4#128 | S4-241048 | | | | Editorial updates | 1.0.1 |
| 05-2024 | SA4#128 | S4-241299 | | | | Removal of language on levels | 1.1.0 |
| 05-2024 | SA4#128 | S4-241338 | | | | Editorial updates | 1.2.0 |
| 06-2024 | SA#104 | | | | | Version 2.0.0 created by MCC | 2.0.0 |
| 06-2024 | SA#104 | | | | | Version 18.0.0 created by MCC upon TSG approval | 18.0.0 |
| 06-2024 | SA#104 | | | | | Change of spec title as approved by TSG SA in SP-240917 | 18.0.0 |

# History

| Document history | | |
|---|---|---|
| V18.0.0 | July 2024 | Publication |
| | | |
| | | |
| | | |
| | | |