Audio Engineering Society

Convention Paper 10188

Presented at the AES 157th Convention
2024 October 8-10, New York, NY, USA

Immersive Voice and Audio Services (IVAS) Codec – The New 3GPP Standard for Immersive Communication

Markus Multrus[1], Stefan Bruhn[2], Juan Torres[2], Eleni Fotopoulou[3], Tomas Toftgård[4], Erik Norvell[4], Stefan Döhla[1], Yuan Gao[5], Huan-yu Su[5], Lasse Laaksonen[6], Adriana Vasilache[6], Takehiro Moriya[7], Stéphane Ragot[8], Marc Emerit[8], Hiroyuki Ehara[9], Marek Szczerba[10], Andrea Genovese[11], Andre Schevciw[11], Václav Eksler[12], and Vladimir Malenovsky[13]

[1] *Fraunhofer IIS, Erlangen, Germany*

[2] *Dolby Laboratories, San Francisco, USA*

[3] *DSP Solutions GmbH & Co. KG, Regensburg, Germany*

[4] *Ericsson AB, Stockholm, Sweden*

[5] *Huawei Technologies Co Ltd., Shenzhen, China*

[6] *Nokia Technologies, Tampere, Finland*

[7] *NTT, Kyoto, Japan*

[8] *Orange Innovation, Lannion / Rennes, France*

[9] *Panasonic Holdings Corporation, Yokohama, Japan*

[10] *Koninklijke Philips N.V., Eindhoven, The Netherlands*

[11] *Qualcomm Incorporated, San Diego, USA*

[12] *Consultant for VoiceAge Corporation, Montreal, Canada*

[13] *VoiceAge Corporation, Montreal, Canada*

Correspondence should be addressed to Markus Multrus (markus.multrus@iis.fraunhofer.de)

ABSTRACT

The recently standardized 3GPP codec for Immersive Voice and Audio Services (IVAS) is the first fully immersive communication codec designed for 5G mobile systems. The IVAS codec is an extension of the mono 3GPP EVS codec and offers additional support for coding and rendering of stereo, multi-channel, scene-based audio (Ambisonics), objects and metadata-assisted spatial audio. The IVAS codec enables completely new service scenarios with interactive stereo and immersive audio in communication, content sharing and distribution. This paper provides an overview of the underlying architecture and new audio coding and rendering technologies. Listening test results show the performance of the new codec in terms of compression efficiency and audio quality.

1 Introduction

The Third Generation Partnership Project (3GPP) has recently finalized its Release 18 ("5G-Advanced") with many new functionalities and features, including a new codec for immersive voice and audio services (IVAS). The standardization of the IVAS codec is the latest step in 3GPP's efforts to uphold its voice and audio services in the cutting edge while maintaining highly competitive service efficiency and quality of experience. A milestone achieved in 2001 was the introduction of AMR-WB coding, enabling wideband (WB) voice services in mobile communications. A further milestone reached in 2014 was the introduction of Enhanced Voice Services (EVS) coding, including super-wideband (SWB) and full-

band (FB) voice/audio, enhanced frame loss resiliency, and high music quality at very low bitrates. The benefits for users were clear: expanding the frequency band increases intelligibility and listening comfort making long conversations less fatiguing. Rendering the complete speech spectrum or even the human audible spectrum ensures true voice naturalness and high music quality.

While the quality promise of this voice service evolution is very high and can be widely experienced in today's mobile telephony, one essential limitation remains: like in traditional telephony, there is the mere provision of monophonic experiences. For full immersion of the user into an audio scene, which is required for true "being there" experiences, 3D spatial audio is needed. At the same time, consumers are already widely accustomed to immersive audio media playback experiences, for example in cinema rooms, on their home cinema or surround sound music systems, and even on their mobiles. For mobile communications, however, immersive audio experiences are still lacking. 3GPP addressed this gap when starting its quest for the new IVAS codec standard.

This paper provides an overview of the IVAS codec standard. It begins with a review of the state of the art of immersive audio coding (section 2), followed by essential features and foreseen IVAS service contexts and use cases (section 3) constituting the terms of reference for the standardization effort that is shortly described in section 4. The main technologies are described in section 5, followed by examples of the performance achieved with the IVAS codec (section 6). A conclusion follows in section 7.

2 State of the Art

In recent years, the popularity of spatial audio has increased significantly, facilitated by technological advancements in audio codecs such as MPEG-H 3D Audio [1], AC-4 [2] and DTS-UHD [3], which support various immersive audio formats and improve the overall user experience. These codecs were primarily developed for entertainment applications, like streaming and broadcasting services, and were not designed for communication applications.

In the realm of communication codecs, there are limited examples that extend support beyond mono. A trivial solution is to code channels separately in a multi-mono or multi-stereo approach; however, this is suboptimal. MPEG AAC-ELDv2 [4, 5] supports stereo and surround multi-channel configurations up to 5.1 loudspeakers for speech and general audio content. Stereo extensions of ITU-T codecs (G.722, G.711.1) have been specified [6]. These codecs are mainly targeting voice services. Another codec, Opus [7, 8], supports coding for stereo, multi-channel and Ambisonics, covering a wide range of bitrates and number of channels.

Despite these technological advancements, current codecs do not fully satisfy all the specific needs of modern mobile networks. Future communication scenarios require support for a variety of spatial audio input formats, across a diverse range of services and devices. It is also essential to have native rendering capabilities to handle various output formats, distinct from the original input format being decoded, for binaural or loudspeaker playback. Furthermore, essential features including Discontinuous Transmission (DTX), bitrate adaptation, Packet-Loss Concealment (PLC), and Jitter Buffer Management (JBM) are crucial for achieving satisfactory performance in mobile network environments. These needs were compiled into a full set of requirements for IVAS operation over 3GPP mobile networks [9]. The IVAS codec offers a solution specifically tailored to meet these requirements while enabling rate-efficient high-quality immersive audio communication.

3 Features and Use-Cases

The IVAS codec is an extension of the 3GPP EVS mono codec [10, 11] towards stereo and immersive coding. It features coding and rendering of immersive signals, a broad range of supported bitrates spanning from low-bitrate efficient coding to excellent quality high-bitrate coding, low delay suitable for communication applications while meeting practical implementation complexity/memory requirements. The key attributes of the IVAS codec are summarized in Table 1. The bitrates given in Table 1 indicate "total bitrate", i.e., the bitrate to represent the complete encoder input, not individual channels or components only. Based on total bitrate and signal characteristics, the codec dynamically allocates portions of the total bitrate to different coding elements.

In addition to EVS mono and stereo coding, IVAS supports the following immersive formats: scene-based audio (2D planar or 3D Ambisonics of order 1 - 3), object-based audio, and channel-based audio (including most common multi-channel configurations from 5.1 up to 7.1+4), and metadata-

Sampling rates	16, 32, 48 kHz
Audio bandwidths	WB (20 – 8,000 Hz), SWB (20 – 16,000 Hz), FB (20 – 20,000 Hz)
Bitrates	13.2, 16.4, 24.4, 32, 48, 64, 80, 128, 160, 192, 256, 384, 512 kbps
Frame length	20 ms (encoder/decoder) 5/10/20 ms (renderer)
Algorithmic delay	32 to 38 ms (incl. rendering delay)
Audio formats	mono, stereo, scene-based audio (Ambisonics order 1- 3), object-based audio (1 – 4 simultaneous objects), multichannel-based audio (5.1, 5.1+2, 5.1+4, 7.1, 7.1+4), MASA, combination of formats
Output configurations	loudspeakers (mono, stereo, 5.1, 5.1+2, 5.1+4, 7.1, 7.1+4, arbitrary layout), Ambisonics (order 1-3), binaural, pass-through

Table 1. IVAS key features in stereo and immersive coding.

assisted spatial audio (MASA) – the latter is a new parametric audio format designed for direct spatial audio pick-up from smartphones.

IVAS also provides rendering functionality for multi-loudspeaker playback and binaural rendering for headphone reproduction including head-tracking, scene rotation, reverberation, and support of external customized Head-Related Transfer Functions (HRTFs). For head-tracked rendering support on lightweight devices, a split rendering solution is provided. The built-in renderer may also be bypassed so that custom renderers can be used.

The transmission of spatial and immersive audio with the IVAS codec enables new communication scenarios, such as:

- Immersive Telephony with Experience Sharing: This allows participants to capture immersive scenes and convey them to one another. The full immersive audio experience of, e.g., an event or outdoor environment can be shared.

- Ad-hoc Conferencing: By placing the capturing device on a conferencing table, a realistic acoustic image of the surrounding participants can be picked up and recreated at one or multiple receivers. Rendering the immersive scene makes it easier to distinguish between the speakers' voices and to separate them from ambient sounds.

- Multi-party Conferencing: For more complex situations, the voices of multiple participants can be transmitted as individual streams and spatially rendered on the receiving device to match the video scene transmitted in parallel. An intermediate call server may combine multiple participants calling from various locations into a (virtual) immersive scene, or forward selected streams.

In these scenarios, immersive audio aims to deliver a more lifelike experience approaching true telepresence, while reducing the effort required for listening and minimizing fatigue. IVAS can further be used for live and non-live streaming of user-generated immersive and Extended Reality (XR) content, immersive messaging (Rich Communication Service, RCS), and advanced XR / metaverse applications.

4 Standardization

3GPP IVAS codec standardization was carried out under a dedicated work item [42] supported by companies from the whole mobile industry, ranging from operators, infrastructure and terminal vendors to chipset and technology providers. Split rendering with head-tracking support was added under a related work item [43].

In accordance with best practices of 3GPP for communication codecs, the IVAS codec standardization followed a rigorous process with definition of terms of reference (performance requirements and design constraints) and a selection process with extensive evaluations against the performance requirements by renowned external labs. The work resulted in a set of core specifications describing the codec algorithm incl. DTX [12] and related essential functions such as rendering [13], PLC [14], and JBM [15]. All functions are fully specified by reference floating-point C code [16] along with test sequences [17] to be used for conformance testing of IVAS implementations. Fixed-point source code permitting more power-efficient implementations on fixed-point DSPs and associated test sequences will be provided by a subsequent Release 19 3GPP work item. To enable IVAS operation over 3GPP networks with guaranteed quality of service (QoS), relevant system specifications were updated [18] including necessary

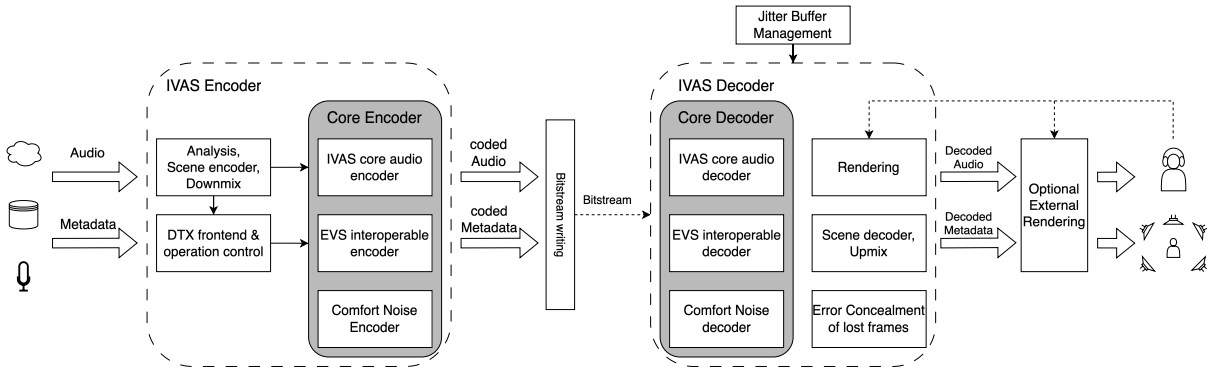

Figure 1. IVAS encoder and decoder architecture overview.

core network specifications. The characteristics of the IVAS codec especially in terms of its rate-distortion performance are documented in 3GPP technical reports [19, 20].

5 IVAS Codec Technologies

5.1 Overview and General Architecture

The IVAS Encoder as depicted in Figure 1 receives an audio signal according to the input audio formats in Table 1. For object-based audio and MASA, the audio signal is accompanied with associated spatial metadata. The spatial metadata is considered an integral part of the immersive audio signal and is included in the further analysis and coding steps. In case of stereo or any immersive audio input, a scene encoder derives from the input signal a set of transport channels with or without spatial metadata. For high-quality/high-rate operation the channel count is typically retained while in other cases, the input channels are downmixed to fewer transport channels. In the latter case, the spatial metadata is used by the decoder to reconstruct the immersive audio signal.

The full set of transport channels is encoded by the IVAS core encoder. When DTX is enabled, voice activity detection is performed to distinguish between active speech/music signals and inactive audio segments. For inactive portions, the comfort noise encoder provides a compact representation of the (immersive) signal.

The transport channels are processed by the IVAS core audio encoder: A *Single Channel Element* (SCE) encoder is based on one core encoder which represents each transport channel independent of each other. A pair of transport channels can be encoded in a *Channel Pair Element* (CPE) encoder that represents both channels using one or two core encoders and stereo coding techniques. Finally, the *Multichannel Coding Tool* (MCT; derived and adapted from [21]) encoder is used for joint core encoding of multiple transport channels. The MCT performs a preprocessing and analysis across all transport channels and a subsequent pairwise grouping of the channels based on correlation. The resulting channel pairs are jointly coded exploiting stereo techniques. Both CPE and MCT coding aim to exploit redundancies between channels and avoid spatial unmasking artifacts. The core encoder itself is a derivative of the EVS codec, with additional flexibility, variable bitrate and optimized complexity.

The IVAS Decoder first decodes the core audio transport channels or comfort noise representation. The scene decoder reconstructs the stereo or immersive signal; in case a downmix was coded by the core audio encoder, an upmix inverting the downmix is performed using the decoded metadata. For immersive signals, thisparametric upmix is performed in a Complex Low-Delay Filterbank (CLDFB) domain, which is applied after the core audio decoder.

Integrated rendering algorithms are tightly coupled to the scene decoder/upmixer, where the IVAS decoder provides rendered loudspeaker or binaural signal output. Alternatively, unrendered decoded audio signals and metadata may be provided to the external IVAS renderer or other custom rendering solutions.

5.2 EVS Compatible Mono Coding and Stereo Downmix

The IVAS codec supports mono coding with the EVS codec by implementing all EVS functionalities (incl. AMR-WB interoperable mode). The implementation is bit-exact to the EVS codec. This functionality is suitable for applications that require low bitrate and

high-quality mono speech and audio. The EVS mono coding in IVAS is thus compatible with other implementations of the EVS codec, which enables seamless integration with existing systems and devices that already use EVS.

In addition, a stereo downmix is supported to generate a mono signal for an EVS interoperable bitstream with no extra delay. Depending on the characteristics of the input stereo signal, one of two downmix tools is selected in every frame. The first tool carries out a weighted sum of two channels with an adaptive weight depending on the inter-channel time difference. The second tool performs a sum of two channels modified by a time domain phase compensation algorithm. This EVS-compatible stereo downmix capability is particularly useful for multi-party conference systems with stereo capture when some attendants need to connect from legacy smartphones.

5.3 Stereo Audio Coding

The IVAS codec is capable of coding stereo signals at bitrates ranging from 13.2 kbps to 256 kbps. At low bitrates, from 13.2 kbps to 32 kbps, the IVAS codec employs a stereo module with parametric spatial representation that dynamically switches between *Time-Domain (TD) stereo* mode and *Discrete Fourier Transform (DFT) stereo* mode.

The TD stereo coder relies on low-bitrate, low-complexity parametric stereo technology that encodes the stereo input signal in the time domain. It is particularly effective in scenarios where the correlation between the left and right channels of the input audio signal is low, when background noise fluctuates, or when an interfering talker is present. In TD stereo, the left and right channels are encoded independently using two mono core coders, with small side information.

The DFT-based stereo performs joint Mid/Side stereo coding exploiting spatial cues, where the mid-channel is encoded by a primary mono core coder. The side channel is predicted parametrically, and the residual is optionally coded by a secondary core coder.

The TD/DFT stereo switching mechanism is driven by a stereo classifier, which evaluates each audio frame to determine the most appropriate mode for optimal stereo image representation and audio fidelity, based on inter-channel correlation.

As the bitrate increases above 32 kbps, the IVAS stereo codec transitions to discrete coding using the *Modified Discrete Cosine Transform (MDCT) stereo* mode. This mode is better suited for higher bitrates due to its advanced capabilities in handling complex audio signals and maintaining high audio quality. The MDCT stereo mode leverages the additional bitrate to provide a richer and more detailed stereo representation, thus enhancing the overall listening experience.

To enhance energy compaction for both the low-rate and mid-rate stereo modes, the input signals are time-aligned based on an estimation of inter-channel time difference.

5.4 Multichannel Audio Coding

The IVAS codec supports coding of *multi-channel* (MC) inputs at bitrates from 13.2 kbps to 512 kbps for channel layouts 5.1, 7.1, 5.1+2, 5.1+4, and 7.1+4. The coding method is selected from four MC coding modes based on the bitrate and input channel layout.

At lowest bitrates, the *Multi-channel MASA* (McMASA) coding mode is used. This coding mode is based on a perceptually motivated parameterization, largely shared with the MASA coding (see Section 5.6). Depending on bitrate, McMASA utilizes 1 – 3 transport channels, where the third channel is the center channel that can be separated before the downmix. The *Low Frequency Effects* (LFE) channel is always coded parametrically.

At mid bitrates (48 – 160 kbps) two alternative *parametric coding* strategies are used. *Parametric multi-channel coding* mode represents the audio signal by 2 or 3 downmix channels (depending on bitrate) and inter-channel level differences and correlations. The decoder reconstructs the audio signal following a covariance synthesis method [22]. Another coding strategy is used in the *parametric multi-channel upmix coding* mode, which applies a hybrid parametric/waveform linear predictive approach in sub-band domain with para-metric encoding of some channels. Non-predictable components are reconstructed by decorrelators.

At high bitrates, a fully *discrete coding* mode using the MCT algorithm is employed. Depending on the format, it is used starting from 96 kbps (5.1) to 192 kbps (7.1+4).

Except for the McMASA coding mode, the LFE channel coding is based on a separate waveform

coding method in MDCT domain. This technique encodes frequencies of up to 400 Hz with a focus on very low frequencies up to about 130 Hz. For McMASA coding mode, the LFE signal is coded together with another channel. The LFE is then extracted from the mix using LFE-to-channel ratio and low-pass filtering.

5.5 Scene Based Audio (Ambisonics) Coding

Scene-based audio (SBA) or Ambisonics represents a sound field at a reference point by means of spherical harmonics [23]. The spatial precision is only limited by the order of the spherical harmonics, which also impacts the perceptual sweet spot, localization, source-width and coloration [24]. Orders 1 – 3, as supported by IVAS, suffice for a high perceptual quality with a manageable number of input signals. Ambisonics audio can be captured using irregular microphone arrays, such as those found on smartphones followed by an upmix process [25, 26] or by external microphone input. Scene-based audio is suitable for representing immersive audio scenes with potentially complex audio environments, including point audio sources, e.g., stemming from multiple talkers in a conference scenario, or diffuse/ reverberant sources such as acoustic environments.

The IVAS codec supports coding of SBA inputs at bitrates from 13.2 kbps to 512 kbps. IVAS SBA coding is based on a combination of *Spatial Reconstruction* (SPAR) [27] and *Directional Audio Coding* (DirAC) [28] technologies. SPAR coding, applied to lower frequencies up to 4 kHz, relies on a hybrid parametric/waveform predictive approach in the sub-band domain that is able to faithfully reconstruct the input signal when the number of input channels is equal to the number of transport channels. Cross-prediction of the Ambisonics component channels is followed by either a parametric or a waveform representation of the residual signals. Non-predictable components are reconstructed by decorrelators. DirAC coding, applied to higher frequencies, is based on findings from psychoacoustics and on the underlying assumption that humans localize only one sound object per critical band at a time.

In the encoder, a first-order Ambisonics representation of the input signal is analyzed per frequency band in terms of Direction of Arrival (DoA) and diffuseness. Signal re-synthesis aims at reproducing DoAs and diffuseness in the respective frequency bands. At 384 and 512 kbps, a variant of higher-order DirAC [29] is employed to encode two DoAs and a global diffuseness parameter at the same time. This improves the spatial resolution of complex scenes, e.g., with overlapping talkers. In addition, DirAC enables upmixing to Ambisonics orders higher than that of the input signal or the signal encoded in the bitstream.

In first-order Ambisonics operation at 256 kbps, specific tools for adaptive decorrelation using *principal component analysis* [30] are also supported.

5.6 Metadata Assisted Spatial Audio (MASA) Coding

Metadata-Assisted Spatial Audio (MASA) is a parametric spatial audio format that can be used with any multi-microphone array by implementing a suitable capture analysis. The MASA format [16, 31] developed as part of the IVAS work item, is specifically optimized for direct spatial audio pick-up from smartphones.

The MASA format is based on a combination of 1 – 2 audio channels and an associated set of metadata parameters. The metadata includes spatial metadata that provides information about the captured spatial audio scene and descriptive metadata that provides further description about the capture configuration that can be used, e.g., to aid rendering.

The IVAS codec supports MASA coding at bitrates between 13.2 kbps and 512 kbps. For MASA format, one 20 ms frame consists of 1 or 2 transport audio channels and the corresponding metadata. The MASA transport audio channels are coded using the SCE and CPE encoders based on a bit allocation available after the encoding of the spatial metadata.

The spatial metadata includes the following parameters: spatial direction index [32], direct-to-total energy ratio, diffuse-to-total energy ratio, remainder-to-total energy ratio, spread coherence, and surround coherence. These are provided for 4 temporal sub-frames and 24 frequency bands, i.e., 96 time-frequency (TF) tiles. The uncompressed spatial metadata is 272.8 kbps or 422.4 kbps depending on whether there are 1 or 2 spatial directions per TF tile. To respect the maximum number of allowed bits, the metadata parameters are reduced prior to coding by merging values in time and frequency based, e.g., on detected metadata composition, total bitrate, and number of transport channels. Following the reduction of metadata, the MASA metadata is quantized and encoded with variable bitrate.

The encoded MASA metadata uses at most 19% of the total bitrate.

The direction encoding consists of encoding of azimuth and elevation values with one of three entropy coders selected based on accuracy and bit consumption. For the energy ratio parameters, non-uniform quantization is used with different encoding methods, e.g., for encoding one or two concurrent directions. The coherence values, when present, are quantized and encoded only starting at 48 kbps.

5.7 Object Based Audio Coding

The IVAS codec supports simultaneous coding of up to four independent audio objects, each object combined with associated metadata, at bitrates from 13.2 kbps to 512 kbps. The coding is based on input objects analysis, metadata coding, and inter-object core-coder SCEs bitrate adaptation. It employs two coding schemes: the *discrete mode* in which each audio object is coded by one SCE, and the *parametric mode*. On the decoder side, the objects can be rendered to the requested output configuration while a customization of the objects' spatial properties can be done.

A performance advantage compared to conventional uniform bitrate assignment is achieved by using an efficient inter-object core-coder bitrate adaptation method. The method distributes an available codec bit budget to encode waveforms of the individual audio objects based on a classification of the objects' subjective importance in particular frames.

A parametric mode is used at low bitrates (32 kbps and 48 kbps) for coding 3 or 4 audio objects. In this mode, the most relevant 2 objects are identified per frequency band while spatial parameters including direction information are extracted. The waveforms of the 3 or 4 audio objects are then downmixed into 2 transport channels coded by two SCEs.

In IVAS, two groups of object-based audio metadata are analysed, coded, and transmitted in the bitstream. A first group, supported at all bitrates, comprises the object position in spherical coordinates using azimuth and elevation angles. A second group, optionally supported at total bitrates equal or higher than 64 kbps, adds radius to the position, and the object orientation using yaw and pitch. There is also an option that some or all input objects are sent without metadata; in this case associated metadata may be available by other means. Finally, there is a support of coding one non-diegetic audio object for which metadata consists of a panning gain parameter.

5.8 Combined Objects and SBA/MASA Coding

The IVAS codec also supports two format combinations consisting of up to 4 independent audio objects and an audio scene represented as either SBA or MASA. These combinations are called OSBA and OMASA, and they both cover bitrates between 13.2 kbps and 512 kbps. The OSBA coding consists of two internal coding modes, while the OMASA coding has four different modes. At the lowest bitrates the encoder pre-renders the objects into the respective SBA and MASA audio scene using the object metadata, and with increasing bitrate more objects are discretely coded. At high bitrates, all objects are discretely coded in addition to the SBA or MASA audio scene.

5.9 Additional Features

The IVAS codec provides additional features which are essential for mobile networks:

- DTX with *Comfort Noise Generation* (CNG) for rate-efficient stereo and immersive conversational voice and audio transmissions. The feature is supported for stereo, SBA, MASA, and object-based audio coding. For inactive signal portions, a Silence Insertion Description (SID) is transmitted. The IVAS SID is composed of a single transport channel (downmix or single object) and spatial parameters. The resulting IVAS SID payload size is 104 bits.The default SID transmission interval is once per 8 frames, but other update intervals are also supported.

- *Error concealment* mechanisms to combat the effects of transmission errors and lost packets [14]. A major part of PLC resides in the core decoder based on EVS. The spatial parameters are then either extrapolated from the most recently received frame or subject to further concealment operations performed on these parameters.

- Support for instantaneous *bitrate switching* at any frame boundaries upon command. This is primarily used to adapt the bitrate to varying network conditions.

- JBM mechanism [15] that allows operation over networks with packet interarrival jitter by applying time-scaling to audio transport channels

and parametric representations within the decoder to allow operation with low delay and low packet loss at low complexity overhead.

- *Real-time Transport Protocol* (RTP) support and Session Description Protocol (SDP) parameters for VoIP operation with the specific demands of 3GPP networks in mind. The defined payload format and SDP parameters ([12] Annex A) builds on top of the EVS payload format for backwards interoperability and supports all features of the codec plus a transport mechanism for additional processing information to enhance the rendering.

5.10 Rendering to Loudspeakers and Format Conversion

The IVAS codec supports predefined loudspeaker layouts as detailed in Table 1. In addition, the renderer also supports flexible rendering to arbitrary loudspeaker layouts up to a total of 16 loudspeakers (including one LFE channel).

Format conversion for known and supported loudspeaker layouts is performed via predefined up- and downmix tables. For conversion to an arbitrary loudspeaker layout the Edge Fading Amplitude Panning (EFAP) [33] algorithm is used. For rendering of Ambisonics or object-based audio to loudspeakers, All-Round Ambisonic Panning and Decoding (ALLRAD) [34] with EFAP is used to compute virtual loudspeaker panning gains. Rendering of MASA-based representations to loudspeakers is based on Vector Base Amplitude Panning (VBAP) [35].

5.11 Binaural Rendering

The binaural rendering process uses HRTFs and acoustic synthesis to create spatial immersive audio over headphones [13]. The system allows for head-tracking and scene orientation control which enable an interactive sound field responsive to head movements.

In IVAS there are four binaural renderers, optimized depending on input audio format, bitrate, coding mode, and binaural rendering output mode. Two of these renderers (low-complexity parametric renderer, convolution-based renderer) operate in the CLDFB domain. The other two renderers operate in the time-domain: a multi-channel convolution-based renderer and an object-based audio renderer, which also supports rendering in 6 degrees-of-freedom (6-DoF). The decoder output is routed to the renderer, optimized for the input format and mode. In specific cases, the data stream is transformed to fit a different renderer if there are computational advantages. More information on the renderers can be found in [12, 13].

The synthesis of *room acoustics* allows for a realistic immersive effect. Room acoustics can be synthesized using Binaural Room Impulse Response (BRIR) convolution or using a reverberator, that can be combined with spatialized early reflections synthesis for achieving high quality with acceptable complexity. There are two reverberator algorithms available, which apply according to the rendering domain: a sparse frequency-domain reverberator and a feedback-delay-network reverberator. The acoustic parameters that control the reverb synthesis consist of RT60 and DSR (diffuse-to-source energy ratio) per band, and a pre-delay (at which the DSR was estimated). The early-reflections synthesis is based on the first-order reflections of a shoebox model, which includes room dimensions and absorption coefficients per wall. Optionally the listener's coordinates within the virtual room can be also specified to simulate wall proximity effects. A low-complexity mode allows for reduced computational cost in exchange for lower spatial accuracy.

The *rendering controls* allow for real-time listener and scene orientation control, as well as for rendering customization. The orientation controls include capabilities, e.g., for processing scene orientation, handling the listener pose through several orientation tracking algorithms, and combinations of the above. The means of rendering customization include support for custom HRTF and BRIR sets, the setting of room acoustic parameters, object directivity and distance attenuation.

5.12 Split Rendering for Head-tracked Binaural Audio

The IVAS binaural renderer supports *split operation* with head-tracked pre-rendering and transcoding to a head-trackable intermediate representation that can be transmitted to a lightweight end-device such as earbuds or XR glasses that carry out head-tracked post-rendering. This allows moving a large part of the processing load and memory requirements for IVAS decoding and rendering to a (more) capable entity (network node or mobile phone) while offloading the final rendering end-device. The core principle is that correction metadata is transmitted along with the pre-rendered binaural audio signal, allowing adjustments of the binaural signal

in response to the actual head pose at the end-device. In split rendering scenarios [36] with substantial latency on the interface between the pre-rendering entity and end-device (e.g. Bluetooth interface), such adjustments are desirable to avoid impacts on quality of experience due to a build-up of motion-to-sound latency.

Operations at the pre-renderer involve binaural rendering to the potentially outdated ("assumed") head pose obtained via the delay-prone interface between pre-rendering entity and end-device. The correction metadata is derived under an MMSE criterion based on additional binaural renditions at probing poses different from the assumed pose. The adjustments in the post-renderer involve CLDFB-domain filtering of the binaural audio signal using the metadata and interpolations with respect to the deviation between assumed and actual head pose. Coding of the pre-rendered binaural audio signal that is transported between the two devices is done either directly in CLDFB domain using a *Low-Complexity Low-Delay* (LCLD) codec specified as part of the IVAS codec standard or using the integrated *LC3plus* codec [37]. There is also a time-domain PCM interface that allows another transport codec (or no codec) to be used. Split rendering can operate at various DoFs, ranging from 0-DoF (no pose correction) to 3-DoF (pose correction on the three rotational axes yaw, pitch, roll) at bitrates from 256 kbps (0-DoF) to 384 – 768 kbps (1 to 3-DoF).

6 Performance Evaluation

The performance of the IVAS codec was evaluated during the Selection phase of the 3GPP IVAS codec standardization. In total, 9 P.800 DCR [38, 39] and 14 BS.1534 (MUSHRA) [40] experiments were conducted. Each experiment was conducted twice in two independent laboratories.

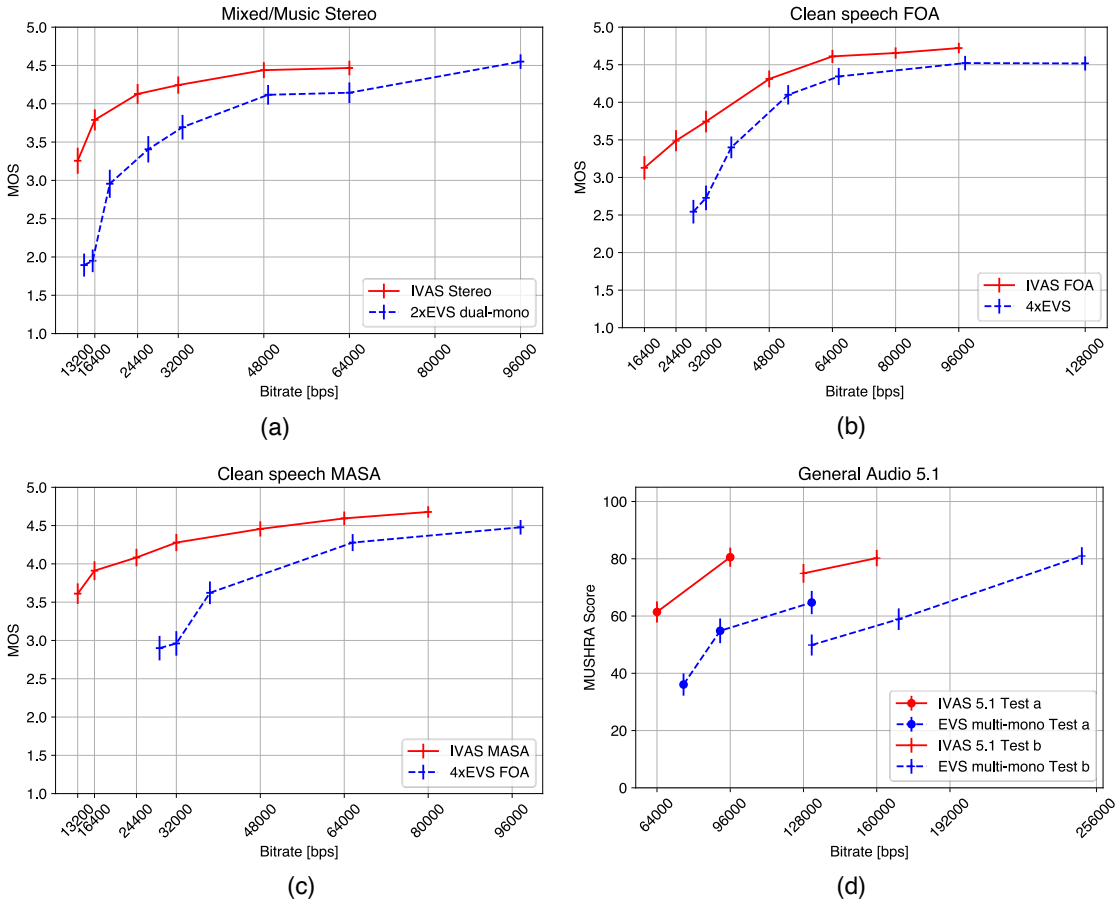

Figure 2. Results of IVAS Selection Tests (a) P800-3, (b) P800-4, (c) P800-8 and (d) BS1534-2a/b for laboratory a. Tests (a) – (c) were performed using headphone presentation, Test (d) using 5.1 loudspeaker presentation. In (d), results of two separate tests (BS1534-2a and BS1534-2b) are depicted.

The experiments were set up to compare the performance of IVAS against EVS operated in a multi-mono fashion. Multi-mono EVS was operated at multiples of its native bitrates, thus for certain comparisons the bitrates of IVAS and N x EVS are either not available or don't exactly match.

Figure 2 illustrates examples of the performance of the IVAS codec for selected operating points. The plots include mean opinion scores (MOS) or MUSHRA scores with 95% confidence intervals using Student's t-distribution. A detailed report on the performance of all tested operating points can be found in [19]. The results indicate significant quality benefits of IVAS over N x EVS.

Evaluations of the IVAS split rendering solution [20] indicate that it provides a quality level very close to the "ideal" alternative of performing full IVAS decoding and rendering in the end-device.

7 Conclusion

The new 3GPP IVAS standard specifies a highly efficient and versatile codec, for the first time bringing immersive audio experiences to mobile communications. This will enable 3GPP and the mobile industry to continue offering and deploying cutting-edge mobile voice services as well as new services relying on immersive audio experiences as one essential attribute. The IVAS standard describes a codec with backwards interoperability to the widely deployed AMR-WB and EVS codecs. It supports a wide range of stereo and immersive audio formats across a large range of bitrates that can be used for rate adaptation on a frame-by-frame basis. Beyond that, the IVAS standard supports a multitude of rendering options ranging from head-tracked binaural rendering including split rendering support, acoustic room modelling, advanced renderer control, multi-loudspeaker rendering to arbitrary loudspeaker layouts, audio format conversions, PLC, DTX, JBM, RTP payload format and SDP parameter definitions. The available reference codebase has already been used in IVAS service demonstrations over 5G networks [41], meaning that the practical hurdles to implement IVAS services are relatively low. As IVAS services are deployed within the 3GPP IP Multimedia Subsystem (IMS) framework (with QoS guarantees) [18] or over-the-top, the authors expect that the public will soon be able to experience the immersive use-cases that are enabled by the codec.

8 Acknowledgements

The authors would like to thank the following people for their valuable contributions to this project: Stefan Bayer, Reinhold Boehm, Alexandre Bouthéon, Jeroen Breebaart, Jan Brouwer, Stefanie Brown, Jan Büthe, Serdar Buyuksarac, Venkata Chebiyyam, Catherine Colomes, Dan Darcy, Grant Davidson, Frans De Bont, Sumeyra Demir Kanik, Martin Dietz, Paul Dillen, Sascha Disch, Michael Eckert, Bernd Edler, Andrea Eichenseer, Per Ekstrand, Tommy Falk, Tomas Frankkila, Guillaume Fuchs, Rory Gamble, Florin Ghido, Alexandre Guérin, Noboru Harada, Akira Harada, Dominik Häußler, Christian Helmrich, Jürgen Herre, Christoph Hold, Jiaquan Huo, Wolfgang Jaegers, Fredrik Jansson, Milan Jelinek, Noboru Kamamoto, Yuichi Kamiya, Erlendur Karlsson, Patrick Kechichian, Jan Kiene, Jenny King, Charles Kinuthia, Srikanth Korse, Fabian Küch, Mikko-Ville Laitinen, Brian Lee, Arnaud Lefort, Arvi Lintervo, Manfred Lutzky, Pallavi Maben, Goran Markovic, Sujeet Mate, Benjamin McDonald, David McGrath, Adam Mills, Chamran Moradi Ashour, Harald Mundt, Srikanth Nagisetty, Werner Oomen, Lauros Pajunen, Jouni Paulus, Nils Peters, Tapani Pihlajakuja, Simon Plain, Harald Pobloth, Karin Prebeck, Shanush Prema Thasarathan, Heiko Purnhagen, Anssi Rämö, Emmanuel Ravelli, Stefan Reuschl, Franz Reutelhuber, Kacper Sagnowski, Markus Schnell, Michael Schug, Erik Schuijers, Martin Sehlstedt, Leif Sehlstrom, Amber Shakespeare, Dan Sinder, Ripinder Singh, Ryosuke Sugiura, Jonas Svedberg, Nathan Swedlow, Archit Tamarapu, Mikko Tammi, Oliver Thiergart, Henri Toukomaa, Anika Treffehn, Rishabh Tyagi, Tommy Vaillancourt, Imre Varga, Vinit Veera, Atti Venkatraman, Juha Vilkamo, Lars Villemoes, Mark Vinton, Jussi Virolainen, Dominik Weckbecker, Przemyslaw Wojciga, Oliver Wübbolt, Mengqiu Zhang, Ke Zhao.

References

[1] J. Herre et al., "MPEG-H Audio—The New Standard for Universal Spatial/3D Audio Coding," *Journal of the AES*, vol. 62, no. 12, pp. 821—830, Dec. 2014.

[2] K. Kjörling et al., "AC-4 – The Next Generation Audio Codec," *140th AES Conv.*, paper #9491, Jun. 2016.

[3] ETSI TS 103 491, "DTS-UHD Audio Format; Delivery of Channels, Objects and Ambisonic Sound Fields," v1.2.1, 2019.

[4] M. Valero et al., "A New Parametric Stereo and Multichannel Extension for MPEG-4 Enhanced Low Delay AAC (AAC-ELD)," *128th AES Conv.*, paper #8099, May 2010.

[5] M. Lutzky et al. "AAC-ELD V2 - The New State of the Art in High Quality Communication Audio," *138th AES Conv.*, paper #8516, Oct. 2011.

[6] D. Virette et al., "G.722 annex D and G.711.1 annex F - New ITU-T stereo codecs," in *Proc. of IEEE ICASSP*, pp. 528—532, May 2013.

[7] J.M. Valin et al. "Definition of the Opus Audio Codec," RFC 6716, Sept. 2012.

[8] J. Skoglund and M. Graczyk, "Ambisonics in an Ogg Opus Container," RFC 8486, Oct. 2018.

[9] 3GPP TSG SA WG4, "IVAS Design Constraints (IVAS-4)," 3GPP agreed Permanent Document S4-231031 (https://www.3gpp.org/ftp/TSG_SA/WG4_CODEC/TSGS4_124_Berlin/Docs/S4-231031.zip), *3GPP TSG-SA WG4 Meeting #124*, 2023.

[10] M. Dietz et al., "Overview of the EVS codec architecture," in *Proc. of IEEE ICASSP*, Apr. 2015.

[11] 3GPP TS 26.445, "EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification," 2014.

[12] 3GPP TS 26.253, "Codec for Immersive Voice and Audio Services (IVAS); Detailed Algorithmic Description including RTP payload format and SDP parameter definitions," 2024.

[13] 3GPP TS 26.254, "Codec for Immersive Voice and Audio Services (IVAS); Rendering," 2024.

[14] 3GPP TS 26.255, "Codec for Immersive Voice and Audio Services (IVAS); Error concealment of lost packets," 2024.

[15] 3GPP TS 26.256, "Codec for Immersive Voice and Audio Services (IVAS); Jitter Buffer Management," 2024.

[16] 3GPP TS 26.258, "Codec for Immersive Voice and Audio Services (IVAS); C code (floating-point)," 2024.

[17] 3GPP TS 26.252, "Codec for Immersive Voice and Audio Services (IVAS); Test sequences," 2024.

[18] 3GPP TS 26.114: "IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction", 2024.

[19] 3GPP TR 26.997, "Codec for Immersive Voice and Audio Services (IVAS); Performance characterization," 2024.

[20] 3GPP TR 26.996, "Immersive Audio for Split Rendering Scenarios; Performance characterization", 2024.

[21] F. Schuh et al., "Efficient Multichannel Audio Transform Coding with Low Delay and Complexity," *141st AES Conv.*, paper #9660, Sept. 2016.

[22] J. Vilkamo, T. Bäckström, and A. Kuntz,

"Optimized covariance domain framework for time–frequency processing of spatial audio," *Journal of the AES*, vol. 61, no. 6, pp. 403—411, 2013.

[23] J. Daniel and S. Moreau, "Further Study of Sound Field Coding with Higher Order Ambisonics," *116th AES Conv.*, paper #6017, May 2004.

[24] M. Frank, "How to make Ambisonics sound good," *Forum Acusticum,* 2014.

[25] M. Heikinnen et al., "Neural Ambisonics Encoding For Compact Irregular Microphone Arrays," in *Proc. of IEEE ICASSP*, pp. 701—705, April 2024.

[26] Dolby Sweden AB, "Example design of spatial audio capture for multi-microphone UE devices," 3GPP Tdoc (written contribution) S4-240231 (https://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/TSGS4_127_Sophia-Antipolis/Docs/S4-240231.zip), *3GPP TSG-SA WG4 Meeting #127*, 2024.

[27] D. McGrath et al., "Immersive Audio Coding for Virtual Reality Using a Metadata-assisted Extension of the 3GPP EVS Codec," in *Proc. of IEEE ICASSP*, pp. 730—734, May 2019.

[28] V. Pulkki: "Spatial sound reproduction with directional audio coding," *Journal of the AES*, vol. 55, no. 6, pp. 503—516, 2007.

[29] C. Hold et al., " Compression of Higher-Order Ambisonic Signals Using Directional Audio Coding", *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 651—665, 2024.

[30] P. Mahé, S. Ragot, S. Marchand, "First-order ambisonic coding with quaternion-based interpolation of PCA rotation matrices," in *Proc. EAA Spatial Audio Signal Processing Symposium*, Sept. 2019.

[31] J. Paulus et al., "Metadata Assisted Spatial Audio (MASA) - An Overview" in *Proc. of IEEE International Symposium of Internet of Sounds*, 30 Sept.-2 Oct 2024, Erlangen, Germany, in press.

[32] S. Ragot and A. Vasilache, "Spherical Vector Quantization for Spatial Direction Coding," in *Proc. of IEEE ICASSP*, 2023.

[33] C. Borß, "A polygon-based panning method for 3D loudspeaker setups," *137th AES Conv.*, paper #9106, Oct. 2014.

[34] F. Zotter, and F. Matthias, "All-round ambisonic panning and decoding," *Journal of the AES*, vol. 60, no. 10, pp. 807—820, 2012.

[35] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Journal of the AES* , vol. 45, pp. 456–466, June 1997.

[36] 3GPP TR 26.865, "Immersive Audio for Split Rendering Scenarios; Requirements"

[37] M. Schnell et al., "LC3 and LC3plus: The new audio transmission standards for wireless communication", *150th AES Conv.*, paper #10491, May 2021.

[38] Recommendation ITU-T P.800 Rec., "Methods for subjective determination of transmission quality," 1996.

[39] Recommendation ITU-T P Suppl 29, "ITU-T P.800 – Use Cases," 2023.

[40] Recommendation ITU-R BS.1534-3, "Method of the subjective assessment of intermediate quality level of audio systems," 2015.

[41] Nokia Press Release, "Nokia makes world's first immersive voice and audio call," https://www.nokia.com/about-us/news/releases/2024/06/10/nokia-makes-

worlds-first-immersive-voice-and-audio-call (Accessed June 19, 2024.)

[42] 3GPP TSG SA, "Revised WID on EVS Codec Extension for Immersive Voice and Audio Services", 3GPP approved document SP-220608 (https://www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGS_96_Budapest_2022_06/Docs/SP-220608.zip), *3GPP TSG SA Meeting #96*, 2022.

[43] 3GPP TSG SA, "Revised WID on Immersive Audio for Split Rendering Scenarios", 3GPP approved document SP-231291 (https://www.3gpp.org/ftp/TSG_SA/TSG_SA/TSGS_102_Edinburgh_2023-12/Docs/SP-231291.zip), *3GPP TSG SA Meeting #102*, 2023.