**Source: Editor[[1]](#footnote-1)**

**Title: IVAS Usage Scenarios (IVAS-9)**

**Version: 0.2.0**

**Agenda Item: 7.5**

1. **Scope**

This document provides a collection of example usage scenarios for the EVS Codec Extension for Immersive Voice and Audio Services (IVAS). The purpose of the collection is to create industry awareness of IVAS and to trigger interest already at an early stage, even prior to IVAS standard finalization. After successful standardization and characterization, these example usage scenarios could be considered to be incorporated into the IVAS TR.

Additional information on the codec development project can be found in the other IVAS permanent documents, for which the latest versions can be found at:

<https://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/IVAS_Permanent_Documents>.

1. **Introduction**

According to the IVAS\_Codec work item description [1], Immersive Voice and Audio Services are expected to cover both conversational and non-conversational use cases where immersive content originates and is consumed in end-user devices. The overall objective is to develop a single general-purpose audio codec for immersive 4G and 5G services and applications including the XR use cases envisioned in 3GPP TRs 26.918 and 26.928 and possibly relying on devices described in 26.998.

The following is a normalized description of IVAS example usage scenarios specifically collected in the IVAS\_Codec WI.

1. **IVAS Example Usage Scenarios**

In order to collect relevant example usage scenarios for IVAS, this clause documents the collected usage scenarios and the common approach for collecting them. The example usage scenarios are expected to be categorized in terms of delivery, media components, and devices, provided with considerations on QoE/QoS requirements, and summarized in terms of feasibility with existing or emerging technologies and equipment. Further considerations relating to the example usage scenario impact on 3GPP network interfaces and devices and any other requirements are welcome.

The template provided in Table 1 is recommended to be used for the collection.

* 1. **Usage scenario template**

Table 1 Proposed Usage Scenario Collection Template

|  |
| --- |
| **Usage Scenario Name** |
| <add usage scenario name> |
| **Description** |
| <add detailed usage scenario description> |
| **Categorization** |
| **Type: <Mono, Stereo, Immersive, AR, VR, XR, MR>****Degrees of Freedom: <0DoF, 3DoF, 3DoF+, OD 6DoF, 6DoF>****Delivery: <Local, Streaming, Interactive, Conversational>****Media Components: <Audio-only, Audio-Visual>****Device: <UE, HMD, Glasses, Automotive, …>** |
| **Preconditions** |
| <provides conditions that are necessary to run the usage scenario, for example support for functionalities on the end device or network> |
| **QoS/QoE Considerations** |
| <provides a summary on potential service considerations (including KPIs/QoE and QoS aspects)> |
| **Feasibility** |
| <provides a summary on how the implementation of such a usage scenario using the IVAS codec is anticipated> |
| **Potential Standardization Status and Needs** |
| <identifies potential standardization needs>Editor’s note: It is invited input on the title and content to clarify the scope of this box. |

* 1. **Telephony Usage Scenarios**

The following is a collection of IVAS telephony usage scenarios.

* + 1. **Stereo and Immersive Telephony**

|  |
| --- |
| **Usage Scenario Name** |
| Immersive audio call with experience sharing |
| **Description** |
| **Usage scenario short description:*** Audio call is established between two participants
* During the call, participant with immersive audio capture capability wishes to share their experience (its atmosphere/ambience) with the other party
* A spatial sound scene is transmitted and rendered to the other party

**User story:**Alice is at home sitting on the living room couch. Alice calls her son Bob wishing to hear what Bob and Alice’s grandchildren are up to. Bob is at the race track, where his children are racing go-karts. Bob is standing in the middle of a hairpin corner with go-karts driving around him. Bob answers the call, greets Alice, and wishes to share the race track atmosphere with her. Bob extends his hand holding his smartphone towards the track corner. Alice now clearly hears the go-karts driving around her, with Bob excitedly cheering his children to go faster, over her headphones. This brings a smile on Alice’s face.Figure 1 presents an illustration of the spatial audio capture and presentation according to the usage scenario.Figure 1. Illustration of the spatial audio capture and presentation. |
| **Categorization** |
| **Type: Immersive****Degrees of Freedom: 0DoF/3DoF****Delivery: Conversational****Media Components: Audio-only** (Audio-Visual possible)**Device: UE (Smartphone), Headphones** |
| **Preconditions** |
| * Alice is wearing headphones (potentially with head-tracking)
* Bob’s smartphone implements multi-microphone immersive audio capture or Bob utilizes an immersive audio capture accessory with his device
 |
| **QoS/QoE Considerations** |
| * QoS: Codec bit rate (16.4 or) 24.4 kbps and higher for high-quality encoding of immersive audio captured by smartphones
* QoE: Immersive audio rendering/binauralization quality
 |
| **Feasibility** |
| Multi-microphone capture of immersive audio on smartphones is getting more common. The smartphone currently dominates the entertainment and communications device categories. Immersive audio can be captured also using a dedicated accessory.Use of headphones for entertainment and communications purposes is increasingly popular. This includes mobile use.Head-tracking technologies are currently not common in consumer devices (such as headphones). There is however growing interest in this capability. On the other hand, while head-tracking is beneficial for the current use case, it is not strictly required. |
| **Potential Standardization Status and Needs** |
| [Required:* Support for high-quality encoding of immersive audio captured by smartphones (including smartphone accessories)
* Binaural rendering of immersive audio to headphones

Potentially required:* Head-tracking information interface for decoder/renderer
* Mode switching between capture/input formats

] |

[

* + 1. **Stereo and Immersive equipment for representative telephone**

|  |
| --- |
| **Usage Scenario Name** |
| Representative telephone and internal business phone equipment |
| **Description** |
| Representative telephone is preferable to have functions of 1. multiple connections to networks including mobile and fixed line.
2. facilities to minimize tandem coding and mixing
3. handling multiple streams of speech
4. stereo/multiple channel rendering or binaural rendering

Internal business phone equipment is preferable to have functions of1. handling multiple streams of speech.
2. stereo/multiple channel rendering or binaural rendering

There are huge number of customers who make phone calls from smartphone coded by EVS. In addition, we can assume internal IP connections between representative phones and internal business phones have reliable and high-speed networks. And both are assumed to be conformant to IVAS, which enables immersive conversation and interoperable mode of EVS. **User story:**Alice, a customer makes a phone call from smartphone through VoLTE to a representative telephone number.Bob, an operator picks up the call from Alice on representative phone equipment and makes direct conversation with Alice with EVS over VoLTE. Bob can transfer the call from Alice to a suitable person, John at internal business phone. Since both the representative phone and the business phone are assumed to be conformant to IVAS, which has interoperable mode of EVS, Alice and John can make conversation without quality degradation.In some cases, Bob can continue talking with Alice under supervision of John, where John’s voice is audible to only Bob, and John can secretly listen to the conversation between Alice and Bob. When prioritized fixed-line higher bit-rate connections between Bob and John are assumed, higher bit rate fully functional IVAS can be used. Bob’s phone has binaural listening environment and can listen to the voice of Alice from left ear and may hear John’s advice from right ear. John has also binaural listening facility, and can listen to Bob’s voice from left ear and Alice’s voice from right ear. Sometimes, Bob and John can make rich communication with documents, texts, video, as well as immersive audio.It is obvious that these types of connection (legacy UE, representative phone and internal business phone) can be used for 3-party teleconference, where all participants can enjoy high-quality communication of EVS or higher quality coding of IVAS without using tandem coding.Fig. 1 Simple construction of teleconference or call center by representative phone and internal business phone |
| **Categorization** |
| **Type: <Mono, Stereo, Immersive>****Degrees of Freedom: <0DoF, 3DoF>****Delivery: <Conversational>****Media Components: <Audio-only, Audio-visual>****Device: <UE, PSTN, tablet, business phone>** |
| **Preconditions** |
| need connections to mobile network and fixed network |
| **QoS/QoE Considerations** |
| QoS; controlled network such as VoLTE, NR and prioritized fixed line should be used for high-quality conversationQoE: simple immersive audio rendering/binauralization quality from multiple parties |
| **Feasibility** |
| Telephone set should take care of accepting and producing multiple stream of speech from various sources |
| **Potential Standardization Status and Needs** |
| Tbd |

]

[

* + 1. **VR Telephony**

[TBD.]

]

* 1. **Conferencing Usage Scenarios**

The following is a collection of IVAS conferencing usage scenarios.

* + 1. **Spatial conferencing**
			1. **Ad-hoc spatial audio telco**

|  |
| --- |
| **Usage Scenario Name** |
| Ad-hoc spatial audio telco |
| **Description** |
| **Usage scenario short description:*** Audio call is established between two participating sites with several participants each
* One site utilizes a mobile device for audio capture and rendering, and the other site features a meeting room with a spatial audio system
* A spatial sound scene is rendered in the meeting room, and a mono audio downmix of the spatial audio is rendered for the mobile device participants

**User story:**Bob and his colleagues have scheduled a teleconference with their clients. They are returning from a business trip, and their flight has been delayed. Bob and his colleagues will therefore not make it to the office in time. Instead of cancelling the call, Bob finds a meeting room at the airport and uses his smartphone to join the teleconference. Bob greets the clients who are seated in a meeting room with spatial audio capture and loudspeaker presentation and introduces the participants joining with him. Bob’s smartphone captures the spatial audio allowing for the voices of Bob and his colleagues to be spatially rendered in the meeting room. The meeting room audio is played back in mono on the smartphone loudspeaker. The meeting is a great success.Figure 2 illustrates the spatial audio capture according to the usage scenario.Figure 2. Illustration of spatial audio capture. |
| **Categorization** |
| **Type: Immersive****Degrees of Freedom: 0DoF****Delivery: Conversational****Media Components: Audio-only****Device: UE (Smartphone), Loudspeakers (meeting room spatial audio capture and presentation system)** |
| **Preconditions** |
| * Bob’s smartphone implements multi-microphone immersive audio capture or Bob utilizes an immersive audio capture accessory with his device
 |
| **QoS/QoE Considerations** |
| * QoS: Codec bit rate (16.4 or) 24.4 kbps and higher for high-quality encoding of immersive audio captured by smartphones
* QoE: Immersive audio loudspeaker rendering quality, accuracy of talker separation in spatial audio capture, quality of echo cancellation in meeting room spatial audio system
 |
| **Feasibility** |
| Multi-microphone capture of immersive audio on smartphones is getting more common. The smartphone currently dominates the (entertainment and) communications device category(ies). Immersive audio can be captured also using a dedicated accessory.Spatial audio capture in a meeting room can be achieved using various microphone array configurations or discrete channels can be obtained from several microphones (worn by or otherwise allocated to talkers). |
| **Potential Standardization Status and Needs** |
| [Required:* Support for high-quality encoding of immersive audio captured by smartphones (including smartphone accessories)
* Spatial audio signal rendering to loudspeakers

Potentially required:* Rendering scalability, i.e., a spatial signal is rendered at lower spatial dimension (e.g., as mono or stereo only)

] |

* + - 1. **Server-based spatial voice conferencing**

|  |
| --- |
| **Usage Scenario Name** |
| Server-based Spatial Voice Conferencing |
| **Description** |
| Figure 3 illustrates one typical realization of the spatial conferencing scenario. It demonstrates the fact that conferencing situations may consist of many endpoints with significantly different capabilities, for instance in terms of supported codec, bit rate, audio bandwidth, capture and render. Some of the endpoints may be 3GPP UEs, some may be PSTN or generic VoIP clients. Among the 3GPP UEs, not all will necessarily support IVAS but only legacy codecs like AMR, AMR-WB or EVS. There is also some capability variety among the IVAS-enabled UEs. In this specific scenario, all participants are connected through a call server (MRFP) that handles mixing, forwarding, and floor control functions. Whilst other conferencing architectures are possible, the use of a centralized server architecture is practical when a diverse set of endpoints with very different capabilities participate in the same conference. Figure 3. Illustration of a multi-party conference call for a spatial conferencing service.The following text discusses in detail the functionality of the nodes of the spatial conference scenario depicted in Figure 3.**1: Call server:** While multi-party calls are sometimes envisioned as peer to peer, practically speaking, when many different types of endpoints need to be serviced as part of the same conference, a server-based architecture becomes attractive. In this usage scenario, conference participants join using a range of endpoints, from PSTN lines, mobile connections, mobile apps (over the top), computer apps, and room systems. Since PSTN connections will continue to represent a significant connection type, transcoding efficiency in the server is still of relevance for the operation of such a service, making decode and re-encode efficiency important. While a Selective Forwarding Unit (SFU) style conferencing server is efficient and may look very appealing, this is infeasible when transmitting to PSTN endpoints and other endpoints that can only receive a single mixed stream. The endpoints may have a wide range of rendering capabilities and the service will need to support endpoint rendering spanning from mono, stereo, headphone rendering (binaural), to multi-channel room systems. Likewise, the conference needs to support diverse audio format capabilities upstream. Clients most commonly use mono as upstream audio format, but endpoints with stereo or multi-channel (e.g., 1st order B-format) capture exist today. The server must be capable of managing these various formats, mix in a common domain, and forward appropriately. **2: Standard PSTN/other PLMN endpoint, whether from an IP-PBX, home phone, or any endpoint residing in a different operator network (in case the interconnect involves transcoding at network edge):** This provides the most ubiquitous access and fallback for connectivity and is supported as a matter of course. This is likely to represent a significant fraction of conference attendees for the foreseeable future, though it is trending downward with time. The key element here is that the PSTN/other PLMN endpoint must receive a mono downmix of the signal and encode in the appropriate format. Thus, the concept of a stream forwarding unit is not applicable here – the server must mix in some manner and re-encode to the required PSTN/PLMN interconnect codec (G.711 being mandatory and always a fallback).**3: Legacy UE running a mono codec (EVS, AMR-WB, AMR-NB):** While currently it is uncommon for mobile service providers to terminate AMR-NB or AMR-WB on a conference server, such termination is both feasible and under consideration by conference service providers. This can be viewed as a higher quality mono endpoint than is feasible with G.711 transcoding and connect. A shift to an EVS mono codec does not fundamentally change the fact that mono up and mono down, while of good quality, does not provide a spatial conferencing experience to the legacy UE. However, it is worth noting that in a server-based architecture, mono sources can be spatially mixed into an audio scene that is delivered to a spatial endpoint. Thus, a user on a spatial endpoint, conversing with multiple mono endpoints, can still receive a spatial experience.**4: UE:** This represents the target of IVAS in the conferencing scenario. The access over a mobile network (LTE, 5G) provides extensive coverage. Other access techniques (WiFi, even OTT via mobile access) may be included. Typically, the audio output is binaurally rendered to headphones but could also be stereo rendered for the increasing number of mobile phones and tablets that have stereo speakers. Mono playout may also frequently occur. Thus, IVAS-enabled UEs should support and apply one out of multiple render types (binaural, stereo, mono). Since the UE is typically connected over a mobile network, there are likely to be bandwidth limits, both upstream and downstream, which may be time varying. According to the nature of mobile communications, the transmission may be prone to errors. Thus, bandwidth efficiency and error robustness are important in this usage scenario.**5,6: Conference room clients.** These two represent the same endpoint type (endpoint in conference room or home), where the endpoint has the capability of spatial capture and spatial playout. The clients may be with or without 3GPP connectivity, natively connected or over the top. The playout could be 1,2,3,..N channel loudspeaker render. The reason that two endpoints are shown is to emphasize the fact that conferencing with two or more rooms is a very common usage scenario and mixing and merging of multiple spatial captures at the server can represent a usage scenario where the computational complexity is relatively high. The key thing to recognize here is that spatial capture and coding may require a significantly higher bandwidth (both upstream and downstream) than other examples, so bit rate efficiency is critical. **7. Computer client:** Typically with headset, running a conferencing client and with or without 3GPP connectivity. This may be an application running on the computer, connected over the top via a reasonably high bandwidth upstream and downstream link. In this example, the computer is playing out through a headset, so that the signal should be binaurally rendered. Capture is typically mono because in this usage scenario there is usually only one talker.**8. Stereo device:** Devices like 5,6 or 4 but with stereo capture and stereo render capability. Stereo capture might be obtained using microphone pairs in various possible configurations such as XY, M/S, ORTF or A/B.**9. Content ingest:** This is an example of pre-recorded or live streaming content that might be fed into and delivered over the system. This could be served for instance while the conference has not yet started or is on hold. Announcements, jingles, etc. to conference participants also fall into this category. Typically, it involves upstream transmission to the conference server only (i.e., there is no downstream audio path from the server to this node).**10. Recording/analytics:** This is an example of a recording function of the conference. Audio analytics features like automatic transcription, speaker annotation, conference control keyword spotting, etc. could also be part of this function. Typically, unless this functionality is realized physically in the call server, it involves downstream transmission without strict bit rate limitations. |
| **Categorization** |
| **Type: Immersive****Degrees of Freedom: 3DoF****Delivery: Interactive, Conversational****Media Components: Audio-only, Audio-Visual****Devices: Legacy UE, UE, PSTN phones, Call server, Desktop/Laptop computer, Conference systems, Content ingest server, Recording / analytics server** |
| **Preconditions** |
| TR 26.980 describes many elements of conferencing using a server (including describing a mixing server and stream forwarding server and combinations of the two). However, the emphasis here is that a spatial conferencing service will have endpoints with different spatial capabilities and thus a key element of the system (and server) is to ensure that all endpoints can deliver high quality audio output to their users, according to their capabilities and the constraints imposed by the network and by the operator offering the service. To meet this general precondition, server-based mixing is practical.While server-based mixing is often a precondition to support a broad array of endpoints, the server design that maximizes flexibility and quality is one that combines both mixing and stream forwarding. For example, it may be beneficial to stream forward a small number of mono voice endpoints and at the same time mix other endpoints. This is realized with codecs allowing both a main audio stream (which is mixed and may be immersive) as well as substreams (which are forwarded but not mixed) and are typically mono streams. Though in practical systems, there are also server designs that require multiple immersive substreams to be sent in parallel.In the industry, it is common to formulate a conferencing service using geographically cascaded servers to form a single “virtual” server. This geographical cascade allows geographically localized participants to speak with short round trip times across their local server while geographically distant participants experience a longer round-trip time. This approach also reduces long backhaul traffic because each local server may mix some traffic, prune some traffic, and selectively forward some traffic. To this end, multiple audio substreams are assembled into common frames for between-server traffic. In these cases, it is common for a secondary server to mix some of the previously forwarded traffic into the main stream for delivery to endpoints, to drop some traffic, and to retain some traffic as independent substreams. Ideally, substreams should be extracted or inserted by parsing of the codec bit stream rather than doing a full decode. |
| **QoS/QoE Considerations** |
| The following considerations related to QoS/QoE of the server-based spatial voice conferencing usage scenario can be made:1. In general, it can be expected that spatial multi-party conferencing will be dominated by mono, two-channel, and 1st order B-format representations. The 1st order B-format representation in the considered usage scenario may typically omit the vertical component because that direction is not particularly useful in audio voice conferencing.
2. **Mono coding performance:** High quality/rate performance should be provided for voice (clean, noisy, reverberant), generic audio and music. It should be assumed that noise reduction to a certain degree is applied in most cases. “Far voice” or capture and encoding of reverberant speech is a frequent occurrence and measures like applying a sufficiently high bit rate should be taken to avoid coding artefacts introduced by room reverberation. Concurrent talkers (overtalking) is a frequent occurrence and, accordingly, measures should be taken to maintain high quality when two or more voices are mixed and encoded.
3. **Two-channel coding performance:** High quality/rate performance should be provided for voice (clean, noisy, reverberant), generic audio and music captured with various setups and in various environments. Capture in conference rooms is a relevant scenario for which it should be assumed that noise reduction to a certain degree is applied and that multiple talkers talking on top of each may frequently occur. Thus, mixed voices coming from different angles are common and should be coded with high quality. Various relevant microphone layouts should be supported, such as XY, M/S, ORTF, A/B. “Far voice” or capture and encoding of reverberant speech is a frequent occurrence and measures like applying a sufficiently high bit rate should be taken to avoid coding artefacts introduced by room reverberation. High quality should also be achieved for ambient sounds, including speech, music or nature sounds, and not only for the (foreground) voice. It should be ensured that the codec does not create onerous artefacts from common audio events in real rooms or real environments.
4. **Spatial coding performance:** High quality/rate performance should be provided for voice (clean, noisy, reverberant), generic audio and music captured over a planar B-Format microphone system. Capture in rooms for conferencing typically have noise reduction to a certain degree applied, but there may be multiple simultaneous talkers from different room directions. The spatial coding of reverberant speech and speech in noise should not cause significant degradations. Reverberant speech capture is an extremely common (or just normal) situation in spatial capture of speech. It should be ensured that annoying artefacts introduced into the ambiance by the encoding process are avoided, e.g. by using a sufficiently high bit rate.High quality should also be achieved for ambient sounds, including speech, music or nature sounds, and not only for the foreground voice. The codec should not create onerous artefacts from common audio events in real rooms or real environments.High quality/rate performance should also be achieved for coding of B-format content with additional vertical component.
5. If supported by the codec, for complexity reasons it may be beneficial for receiving endpoints with mono only rendering capabilities to use coding modes that allow high-quality mono decoding of multiple-channel representations (two-channel or spatial) without the need to decode and downmix the multiple-channel representation.
6. The service should be set up such that it supports high-quality two-channel rendering (binaural/stereo) derived from mono, two-channel or spatial ingests .
7. The service should be set up such that it supports high-quality multi-channel (>= 3 channels) (spatial) rendering derived from mono, two-channel or spatial ingests.
8. High complexity decodes and encodes on the server may result in low scalability (and high cost) for the server-based spatial voice conferencing service. Encoding and decoding complexity of the used codec modes may thus be an important criterion when setting up the service.
 |
| **Feasibility** |
| Spatial audio conferencing systems already exist using Internet-based (IP) protocols, therefore implementation as a 5G service is feasible. |
| **Potential Standardization Status and Needs** |
| It is expected that there will be a specification need for the handling and interaction of the media (audio/video) associated with the conferencing server as well as of the service setup parameters.  |

[

* + 1. **VR Conferencing**

[TBD.]

* + 1. **Virtual Meeting**

[TBD.]

* + 1. **Remote class participation**

[TBD.]

* + 1. **In-Game communications**

[TBD.]

* + 1. **XR Meeting**

[TBD.]

* + 1. **XR Convention / Poster Session**

[TBD.]

* 1. **User-generated content distribution Usage Scenarios**

The following is a collection of IVAS usage scenarios pertaining to user-generated content distribution.

* + 1. **Immersive and VR content distribution**

[TBD.]

# ]

# **References**

[1] Tdoc SP-220608: EVS Codec Extension for Immersive Voice and Audio Services.

1. Lasse Laaksonen – Nokia Corporation [↑](#footnote-ref-1)