Object-Based Audio Coding in Immersive Mobile Communications

Václav Eksler
Consultant for VoiceAge Corporation
Brno, Czechia
vaclav@eksler.cz

Milan Jelínek
Speech and Audio Coding Research Group
University of Sherbrooke / VoiceAge Corporation
Sherbrooke, QC, Canada
milan.jelinek@usherbrooke.ca

***Abstract*—Object-based audio is a spatial audio representation in which all individual sounds are distributed independently and accompanied by metadata. The sounds are then mixed during reproduction, providing a flexible and personalized immersive audio experience. Object-based audio can be found in many audio reproduction, streaming, or broadcasting systems. However, its use in interactive communications is a challenge as it faces many constraints like limited transmission bitrate, low delay, limited computational and memory resources, packet losses, or support of discontinuous transmission. This paper discusses trade-offs to overcome these constraints and presents novel methods that enable the use of object-based audio in modern 5G mobile communications services. These methods have been adopted in the recently standardized 3GPP codec for Immersive Voice and Audio Services (IVAS). Several implementation and performance details of IVAS object-based audio coding are also provided.**

Keywords—object-based audio, 5G, immersive audio, metadata, IVAS

I. Introduction

Historically, mobile communications have been used in conjunction with mono handsets to output sound to only one of the user's ears. In the last decade, users have started to use their portable devices in conjunction with headphones to receive the sound in both ears. The evolution of mobile network services standardized within the Third Generation Partnership Project (3GPP) follows this trend. While audio services in 3G or 4G networks target mono speech and audio communications, 5G networks services can deliver stereo or even immersive spatial audio content.

One of the immersive audio formats is Object-Based Audio (OBA), which represents a complex audio scene as a collection of individual sounds, known as audio objects. Each audio object is defined by an audio channel (typically a mono audio channel of speech, music, or a general audio sound) and accompanying metadata that describes the spatial properties of the audio object. The metadata ensures flexible playback in any reproduction scenario, allows interactivity and setting of artistic or listeners' personal preferences, and use of enhanced audio rendering techniques. This flexibility may include, for example, the ability to adjust sound levels, change the positions of sound sources, or select different languages for reproduction.

Over the past years, OBA representation has been introduced in cinema, for downloading, streaming, or in the broadcasting industry. Well-known examples are the Motion Picture Experts Group (MPEG) standards: MPEG Spatial Audio Object Coding (SAOC) [1], which defines parametric storage and transport of OBA, or MPEG-H [2] that defines coding of 3D audio channels, audio objects, or ambisonics. An amendment to the MPEG-H audio standard later introduced the Multichannel Coding Tool (MCT) [3], which enables signal-adaptive joint coding of multiple audio channels. Furthermore, MPEG-I immersive audio codec [4] provides coding of interactive OBA, Channel-Based Audio (CBA), and Scene-Based Audio (SBA). Another example is the ETSI Next Generation Audio, known as AC-4 [5], which supports Advanced Joint Object Coding (A-JOC) [6], a tool based on a multi-channel downmix of immersive content using perceptual audio coding algorithms along with parametric side information. However, none of these codecs is suitable for interactive immersive communications services which face many constraints like limited transmission bitrate, low delay, limited computational and memory resources, packet losses, or support of discontinuous transmission. Furthermore, those codecs generally employ joint parametric coding of audio objects at low bitrates, which significantly limits playback flexibility. On the contrary, a discrete coding of audio objects (i.e. coding each object separately) is essential for full playback interactivity in conversational communications systems.

This paper discusses the challenges in delivering immersive OBA content in real-time communications systems in section II. The newly developed methods to overcome these challenges are described in sections III and IV. These methods have been adopted and integrated into the recently standardized 3GPP codec for Immersive Voice and Audio Services (IVAS) [7], thereby helping to deliver a 3D audio and voice experience in 5G mobile communications. Some IVAS OBA performance results are provided in section V. Finally, section VI mentions the use of OBA in communications codecs in combination with other audio formats before the paper is concluded in section VII. The term *communications* is further used in this paper to describe conversational communications for real-time, interactive services.

Although IVAS represents a standardized framework primarily focused on 5G mobile communications, its potential use goes far beyond traditional mobile services. It can accommodate various sound-based solutions in novel ecosystems of the Internet of Sounds (IoS), of which a comprehensive review can be found in [8]. IVAS capabilities beyond typical communication scenarios were discussed and demonstrated across three different immersive audio use-cases through a web demo implementation in [9]. OBA in IVAS, with its flexibility, performance, and extensive features, might then represent an attractive spatial audio coding and rendering solution for Networked Music Performance (NMR) systems. These systems constitute an essential component of the emerging field of the Internet of Musical Things (IoMusT) [10] for which 5G is a fundamental enabler [11].

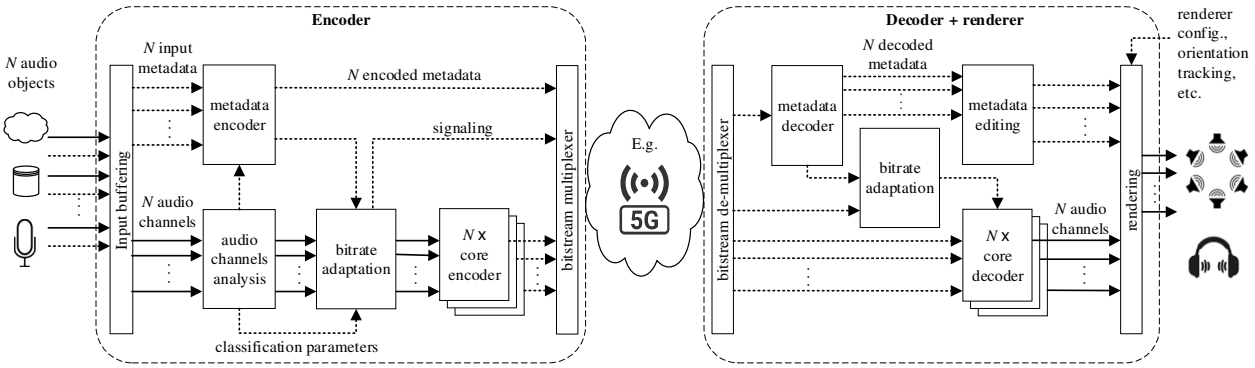

Fig. 1. Block diagram of the presented object-based audio codec in a mobile communications system. Audio channels are denoted by solid lines while metadata and parameters are denoted by dotted lines.

II. CHALLENGES IN COMMUNICATIONS SYSTEMS

Source coding in conversational immersive communications systems faces many constraints compared to codecs targeted for storage, streaming, or broadcasting. The main challenges in designing an attractive communications codec that fully satisfies the specific needs of modern communications networks are summarized below.

- *Transmission bitrate* is the amount of coded data per unit of time. It is related to the compression efficiency while a higher compression results in a higher data loss and thus a lower decoded and rendered audio quality. The bitrate of streaming or broadcast codecs supporting discrete coding of audio objects usually ranges from higher tens or hundreds of kbps to several Mbps (e.g., ~256 kbps – 1.5 Mbps in MPEG-H streams with objects). On the other hand, the transmission bitrate in mono mobile communications systems is usually between 10 and 20 kbps. For example, typical bitrates of communications codecs are 12.65 kbps for the 3G AMR-WB codec [12] or 13.2 kbps for the 4G EVS codec [13]. Keeping the bitrate low in immersive multichannel coding systems is a real challenge.
- *Algorithmic delay* refers to the time needed by the coding algorithm for encoding, decoding, and rendering the audio signal and is typically composed of a frame length, look-ahead, pre- and post-processing delays, and resampling or filter-bank delays. The longer the delay, the higher the coding quality is generally achieved. Broadcasting audio codecs have delays often exceeding 100 ms (e.g., ~100 – 200 ms in MPEG-H OBA) while communications codecs or live-streaming codecs have delays of lower tens of milliseconds (e.g., 32 ms in EVS). Note that this delay does not account for the CPU time or the time for transferring data over a network.
- *Frame or packet losses* commonly occur in packet-based communications networks. Broadcasting codecs often take advantage of frame buffering, redundancy, or retransmission of lost packets to combat lost frames. On the other hand, communications codecs cannot benefit from these strategies due to their low bitrate and low algorithmic delay and thus need to be equipped with error concealment mechanisms to combat the effects of transmission errors with no additional delay. In addition to detected lost frames, communications codecs need to be also robust to residual bit errors. To limit error propagation in decoded audio, communications codecs also aim to limit inter-frame prediction and dependency of coded parameters.
- *Jitter* naturally occurs in packet-based networks and is dealt with by a Jitter Buffer Management (JBM) mechanism that allows for an adjustment of the playout delay by applying time-scale modifications to received audio. The goal of communications codecs is to provide a JBM with a low delay (i.e., using a short JBM buffer length), low complexity, and as low packet loss as possible.
- *Discontinuous Transmission* (DTX) is used in mobile communications systems to switch a radio transmitter off during speech or general audio pauses. The use of DTX saves power in mobile devices and increases the time required between battery recharging. Thus, DTX is an essential part of communications codecs. For inactive audio signal portions, a Silence Insertion Description (SID) is transmitted at certain transmission intervals at a very low bitrate. The SID frame contains Comfort Noise (CN) parameters that are used at the decoder to regenerate the background noise as similar as possible to the background noise at the encoder input. For example, in AMR-WB and EVS codecs, the CN parameters are by default transmitted in one out of 8 inactive frames at a fixed transmission bitrate of 1.75 kbps and 2.4 kbps, respectively. The other 7 out of 8 inactive frames are so-called NO_DATA frames with no transmission (0 kbps). The challenge is how to deal with the CN coding and regeneration in case of an immersive codec, with e.g. several objects, while keeping the CN bitrate low.
- *Computational and memory resources* are generally limited in mobile devices. High-complexity codecs negatively affect application multitasking and shorten battery recharging intervals. Thus, communications codecs are developed with a high emphasis on their low computational complexity and ROM/RAM consumption. The methods described in this paper represent efficiently implemented algorithms with complexity and memory requirements being carefully taken into account in the design choices.
- *Metadata* is an essential part of OBA and ensures that audio objects are correctly rendered, processed, or distributed. A general definition of a high number of different metadata parameters ensuring compatibility

across different systems can be found in the metadata specification of the Audio Definition Model [14]. However, in practice, audio codecs use only a subset of metadata parameters to transport them efficiently at a reasonable bitrate. For example, MPEG-H defines a metadata scheme based on differential coding that provides the spatial positions (azimuth, elevation, and radius), linear gain, spread values (spread in width, height, and depth dimension, or a uniform one), and dynamic object priority [15] of every audio object. In interactive communications codecs, the number of transported metadata parameters is usually further limited due to the available lower transmission bitrate. In addition, their coding needs to consider the presence of frame losses, in which case the differential coding approach is vulnerable to accumulate errors.

III. Object-based audio communications codec

We describe the OBA coding and reproduction system as part of the real-time Immersive Voice and Audio Services (IVAS) codec, but the principles of the presented methods are generally applicable to other codecs as well. IVAS [7] was standardized by 3GPP in September 2023. It is an extension of the 3GPP Enhanced Voice Services (EVS) codec [13] and is optimized for 5G mobile communications networks. IVAS operates on audio frames of 20 ms at several constant bitrates with the lowest bitrate of 13.2 kbps. It supports interactive stereo and several immersive audio formats: CBA, SBA, Metadata-Assisted Spatial Audio (MASA) [16], and OBA. OBA then supports discrete coding of up to four independent audio objects.

Fig. 1 shows a block diagram of the presented OBA codec in a mobile communications system employing, e.g., IVAS. At the transmitter side, N audio objects composed of N input audio channels with associated N metadata are buffered, analyzed, processed, and encoded for each frame at the encoder. The encoded parameters are multiplexed into a bitstream, which is transmitted over a communications network. At the receiver side, the coded parameters are demultiplexed from the bitstream, and N decoded audio channels along with their associated metadata are obtained from the decoder. The metadata can be optionally further adjusted or edited. The obtained audio channels and metadata are finally subject to rendering to get the output audio for loudspeaker or binaural presentation.

A. Audio channels coding

In the presented OBA system with discretely coded objects, the input audio channels are analyzed, preprocessed, and then sequentially encoded using N core encoders. On the decoder side, the audio channels are similarly sequentially decoded using N core decoders and then passed to the renderer. An example of the core encoder and core decoder from Fig. 1 is a mono communications codec (e.g., EVS in the case of IVAS) in which the bitrate is adaptively selected in each frame based on the objects' importance. E.g., the adaptation algorithm described in [17] can be employed which classifies each audio object based on a metric of how critical the coding of the particular object is for obtaining a given quality in the whole object-based audio system. The bitrate adaptation block (see Fig. 1) of this algorithm then assigns a higher bitrate to objects with a higher importance class and a lower bitrate to objects with a lower importance class to achieve an overall improved performance.

TABLE I. Overview of OBA Metadata in the IVAS Codec

parameter	range	quantizer	resolution
azimuth	-180º to +180º	7-bit	2.5º/5.0º
elevation	-90º to +90º	6-bit	2.5º/5.0º
yaw	-180º to +180º	7-bit	2.5º/5.0º
pitch	-90º to +90º	6-bit	2.5º/5.0º
radius	0 to 15.75 m	6-bit	0.25 m
panning gain	-90º to +90º	7-bit	2.5º

B. Metadata quantization and coding

In parallel to the audio channels coding, the input metadata for all N objects is similarly analyzed, quantized, and encoded at the encoder side, and then decoded at the decoder side. As discussed above, due to the usually low available bitrate in communications codecs, several constraints limit the choice of the number of coded metadata parameters, their resolution (i.e. the quantization step) as well as their coding approach.

At IVAS lower bitrates (48 kbps and lower), only directional metadata, i.e., azimuth θ and elevation φ are considered. The metadata is quantized with a certain resolution to keep the number of bits used for quantization reasonable. The azimuth θ is quantized using a 7-bit uniform scalar quantizer with quantization step of 2.5º between –140º and 135º and 5º otherwise, and the elevation φ is quantized using a 6-bit uniform scalar quantizer with a quantization step of 2.5º between –70º and 65º and 5º otherwise [18]. At IVAS higher bitrates (64 kbps and higher), in addition to directional metadata, also orientation (yaw and pitch) and radius metadata are optionally supported, while the orientation metadata is quantized and coded in the same way as the directional metadata. Finally, when a non-diegetic audio object is encoded, the panning gain is coded and quantized. The metadata parameters, their values, quantizer size, and quantization resolution used in OBA in IVAS are summarized in TABLE I.

The quantization indexes of azimuth and elevation are further coded. As the trajectory of a moving audio object can be generally considered smooth and continuous, differential coding is a natural default choice for a low-bitrate coding approach. However, as communications networks are subject to frame losses and jitter, the loss of a frame with differentially coded metadata would result in incorrectly decoded metadata not only in the lost frame but also in the following frames. To limit the accumulation of errors, absolute coding of metadata is used after a certain interval, although it usually consumes more coding bits. In IVAS, this interval is set to 10 frames [18]. Similarly, absolute coding is used in frames in which the differential coding approach would consume a higher or equal number of bits compared to the absolute coding approach, or when no metadata is coded in the previous frame.

The coding approach (absolute vs. differential) used in a particular frame is signaled by a 1-bit flag, one flag for each metadata parameter. The choice of coding approach can be further extended by two rules:

- *Intra-object metadata coding:* The use of absolute coding is limited to one metadata parameter in a given frame, which limits the large fluctuation of bits spent for metadata coding between frames and thus avoiding too reduced bitrate left for the core coder. Moreover, this rule implies that when the signaling flag is set to the absolute coding

for one metadata parameter, the flags do not need to be transmitted for the other metadata parameters (their coding approach is set automatically to differential coding), which saves a few additional coding bits.

- *Inter-object metadata coding logic*: A similar rule that limits the fluctuation of metadata bits between frames can be extended over all audio objects, such that the number of metadata parameters of different audio objects coded absolutely is minimized in a given frame. This rule is implemented using a set of parameters controlling the number of consecutive frames with differentially coded metadata parameters. Consequently, only the first metadata parameter of first object (e.g. azimuth) is coded absolutely in frame *F*, only another parameter of first object (e.g. elevation) is coded absolutely in frame *F*+1, only the first metadata parameter of second object (e.g. azimuth) is coded absolutely in frame *F*+2, only another parameter of second object (e.g. elevation) is coded absolutely in frame *F*+3, etc.

Another advantage of these two rules is the increased robustness of the codec against frame losses and jitter. When a frame is lost, only a minimal number of absolutely coded metadata parameters among all objects is lost, and the accumulation and propagation of errors due to differential coding is thus limited to several frames of a small number of objects. Thus, the effect of the lost frame on the whole decoded and rendered audio scene is minimized compared to a scenario in which a critical frame containing several absolutely coded metadata parameters is lost.

By adopting the described rules, the bitrate of the metadata is kept low, while it is variable depending on the number of metadata parameters and coding approach in a particular frame. In IVAS, it fluctuates between *N* times 0.2 kbps and *N* times 1.85 kbps depending on the number of audio objects *N*.

C. Rendering

Rendering is a receiver functionality that converts *N* decoded audio channels and *N* decoded metadata for binaural reproduction on headphones or for different loudspeaker configurations. In the case of loudspeaker reproduction, panning gains are computed to position a sound source at a requested point in space by activating a given set of loudspeakers using, e.g., Vector Base Amplitude Panning (VBAP) [19] or Edge Fading Amplitude Panning (EFAP) [20] algorithm. In the case of binaural rendering of audio objects, the decoded audio channels are convolved with the impulse response of a Head-Related (HR) filter for the listener's relative Direction of Arrival (DOA) of each object. HR filters can be generated using various approaches [21], e.g., a compact HR filter model based on B-spline basis functions proved to be an efficient approach for rendering OBA [22].

A realistic immersive audio experience in binaural rendering is then achieved by supporting features like head-tracking, scene rotation, orientation tracking, acoustic-parameters-based or model-based reverberation effect, or support of personalized Head-Related Transfer Function (HRTF) set [18]. Interactivity and the addition of artistic or listener's personal preferences to the rendered audio scene are achieved by supporting a metadata editing feature (see Fig. 1) in which one or more metadata parameters can be adapted, replaced, or added before the rendering process. In the context of IoS, this feature provides users with the possibility of direct interaction with content or Sound Things.

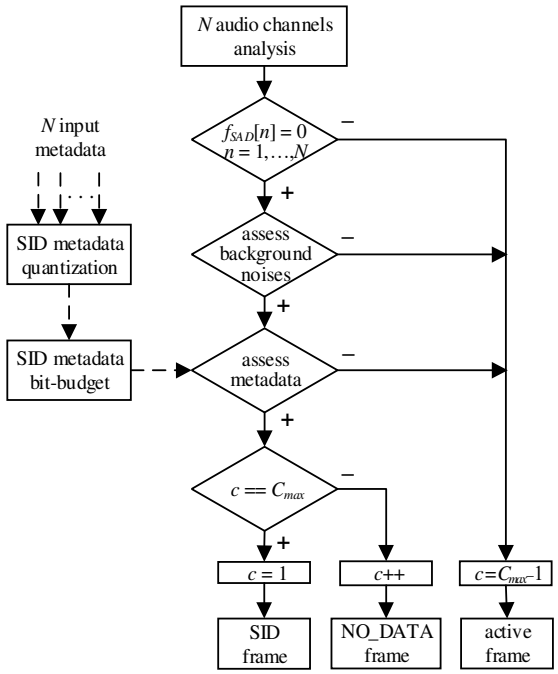

Fig. 2. Flow chart of the OBA DTX control mechanism.

Moreover, IVAS supports split rendering functionality where the decoding and head-tracked binaural pre-rendering are placed at a 5G edge device. The output sound is then post-rendered and reproduced at a lightweight, low-complexity user device. This functionality can be particularly interesting for IoS applications, e.g., in the scope of real-time NMP systems.

IV. Discontinuous Transmission in Object-Based Audio

Discontinuous Transmission (DTX) and Comfort Noise Generation (CNG) are used to reduce the transmission bitrate in communications systems by interrupting transmission during inactive signal portions. When a mono codec operates in the DTX/CNG, a Sound Activity Detector (SAD), generating an SAD flag f_{SAD}, is used to analyze the input audio signal to distinguish whether a frame is active ($f_{SAD} = 1$) or inactive ($f_{SAD} = 0$). When $f_{SAD} = 1$, regular encoding and decoding are performed. When $f_{SAD} = 0$, i.e., the audio frames contain merely background noise, DTX functions are run at the encoder that returns either an SID frame at a very low bitrate or a NO_DATA frame at zero bitrate, and CNG is applied at the decoder.

On the other hand, the DTX/CNG operation in a multichannel codec with *N* audio objects needs to address a trade-off between (a) keeping the SID bitrate low and (b) representing a higher number of channels (objects). Coding of a multichannel SID would result in *N* times the mono SID bitrate plus the metadata bitrate which would be too high. To reduce the overall SID bitrate, an efficient coding of SID information has been developed and is described below. The CNG then aims to recreate both the spatial and spectral characteristics of the background noises of all objects to be as close as possible to the audio objects present at the encoder input.

In the IVAS codec, the overall SID bitrate is set to 5.2 kbps and the SID transmission interval is set to 8 frames [7] for all

multichannel formats, including OBA. The following sections describe the DTX/CNG methods tailored for interactive OBA systems, which made IVAS to be the first communications codec supporting discontinuous transmission for OBA.

A. DTX control mechanism in OBA

The flow chart of the OBA DTX control mechanism is shown in Fig. 2. At the beginning, the mechanism receives the SAD flags f_{SAD}, one flag per audio object, from the audio channels analysis block from Fig. 1 and controls the active/SID/NO_DATA frame decision. Naturally, the time instances of active and inactive frames differ between different audio objects. In general, an inactive frame in the OBA system is declared when frames of all objects are simultaneously declared as inactive, i.e., $f_{SAD}[n] = 0$ for all objects $n = 1, \dots, N$.

In the next step, the DTX control mechanism assesses background noises (see Fig. 2) and it selects an active frame when background noises energies or variations do not allow an efficient coding of CN parameters (see more in section IV.B). For this assessment, the long-term background noise energy values [23] of all audio objects, $e_{noise}[n]$, from the audio channels analysis block in Fig. 1 are used and the mean value $\bar{e}_{noise}$ of $e_{noise}[n]$ energy values over all objects is computed. Similarly, the background noise variation v_{noise} over all objects is computed to check for background noise similarities over all objects. Then, the active frame is selected when $\bar{e}_{noise} > \delta_1$. Similarly, the active frame is selected when $v_{noise} > \delta_2$ and $\bar{e}_{noise} > \delta_3$. The values of the thresholds were determined experimentally and set in IVAS to $\delta_1 = 50$, $\delta_2 = 32$, and $\delta_3 = 25$ [18]. Finally, the metadata bit budget is assessed (see Fig. 2) which impacts the active/inactive frame selection as described in section IV.C.

When the DTX control mechanism does not detect an active frame, it continues with the next steps as part of an inactive frames coding. The segment of inactive frames is controlled by a global SID counter, c, which counts the number of consecutive inactive frames since the last SID frame. When the counter $c = C_{max}$, the DTX control mechanism selects the SID frame; otherwise, it selects the NO_DATA frame, while C_{max} is the SID transmission interval (in IVAS, default $C_{max} = 8$). The counter c is reset (initialized) to $c = C_{max} - 1$ in an active frame, set to $c = 1$ in a SID frame, and incremented by one in every NO_DATA frame. The counter thus guarantees the SID transmission interval and synchronizes the SID/NO_DATA frame decisions over all objects.

B. Comfort noise coding

The CN parameters that are coded and transmitted in the OBA SID frame represent core-coder CN parameters and spatial CN parameters. Due to a limited SID bitrate, CN core-coder parameters of only one dominant audio object are considered, where the object with the highest long-term energy is chosen as the dominant one. The assessment of the background noises from the DTX control mechanism (see section IV.A) ensures that there is always only one dominant object present when an inactive frame is selected. An example of the CN core coder, which is also employed in IVAS, is the EVS frequency domain CNG with a bitrate of 2.4 kbps [18]. The spatial CN parameters then represent the identification of the dominant object and the coherence between the dominant object and all other objects ($N - 1$ values).

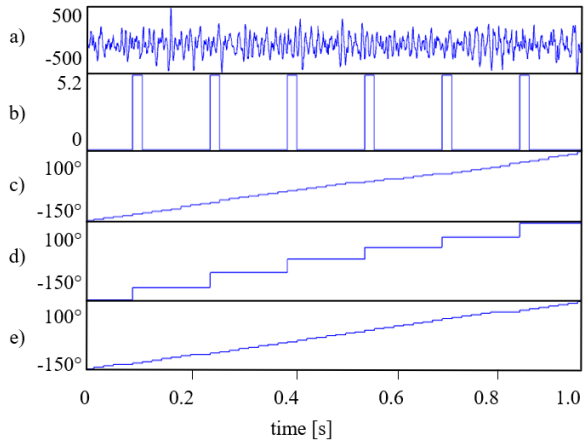

Fig. 3. Metadata coding in DTX inactive frames: a) background noise audio signal, b) codec bitrate [kbps], c) input azimuth, d) decoded azimuth, e) decoded azimuth with adjustment.

At the decoder, the CNG mechanism as part of the core decoder is based on noise tracking [18] in which the background noise is separately estimated for each object in active frames. In inactive frames, the background noise of the dominant object is then recreated from the received CN core-coder parameters and the noise tracking characteristic. The background noise of the other (non-dominant) objects is recreated by shaping the tracked noise with the received spatial CN parameters [18].

C. Metadata quantization and coding in DTX

In inactive frames, the same principles are followed as in the coding of active frames (section III.B), but there are also a few differences for efficient coding of metadata at the limited SID bitrate and SID transmission interval. First, metadata in SID frames is coded using solely the absolute coding approach to minimize potential degradations in case of an SID frame loss in a long segment of inactive frames. Then, compromises in the metadata bit budget need to be made. E.g., in IVAS, only about one-half of the SID bitrate (5.2 kbps) can be used by metadata (the rest of the bit budget is reserved for signaling and CN parameters). This is sufficient when only one or two audio objects are coded, but several additional trade-offs are needed when coding three or more objects.

First, the resolution of the metadata depends on the number of coded objects where a lower resolution, consuming less bits, is used with a higher number of objects. In IVAS, the following strategy is employed: (a) when the number of coded objects is one or two, the azimuth θ is quantized with 8 bits and the elevation φ is quantized with 6 bits; (b) when the number of coded objects is three or four, the azimuth θ is quantized with 6 bits and the elevation φ is quantized with 4 bits; the quantization strategy is known at the decoder from the number of coded objects which is by default signaled in the bitstream.

Second, the metadata is not coded and transmitted for a specific object when its values in a present SID frame do not differ from their last transmitted values by more than certain thresholds. For this purpose, a 1-bit metadata presence flag is computed and transmitted for each object. The flag set to 0 can also indicate the absence of input metadata for the associated audio object. For example, for directional

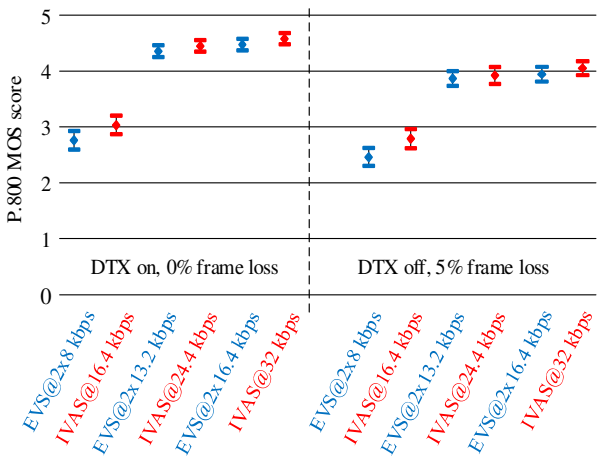

Fig. 4. IVAS codec Selection test results for two clean speech objects (P800-7 DCR, lab a, 30 listeners, 95% CI).

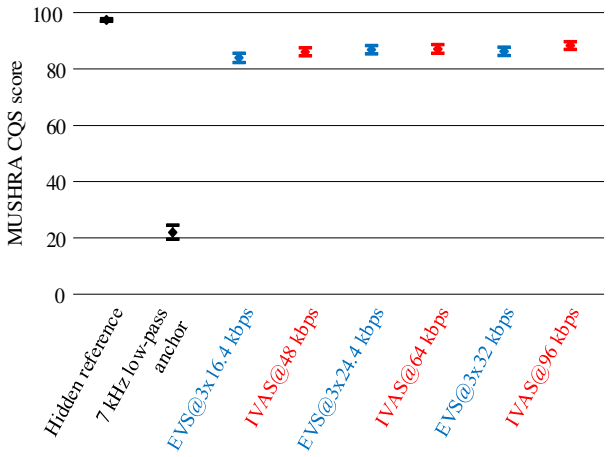

Fig. 5. IVAS codec Selection test results for three generic audio objects (BS1534-6a, labs b+d, 2x14 listeners, 95% CI).

metadata, the metadata presence flag, $f_{MD}[n]$, one for each object, is computed as follows (the index $[n]$ is avoided for clarity):

$$f_{MD} = \begin{cases} 1 & \text{if } \left(|\theta - \theta_{last}| > \delta_\theta \text{ or } |\varphi - \varphi_{last}| > \delta_\varphi\right) \\ 0 & \text{otherwise, or metadata is not present} \end{cases} \quad (1)$$

where θ_{last} and φ_{last} are the last coded and transmitted azimuth and elevation of an audio object, and δ_θ and δ_φ are the thresholds for azimuth and elevation difference. In IVAS, the thresholds are set to $\delta_\theta = 10°$ and $\delta_\varphi = 10°$ [18]. When $f_{MD}[n] = 1$, the metadata is transmitted for object n; otherwise, it is not.

Third, the bit budget for encoding and quantizing metadata of all objects is estimated and compared to the available SID bit budget (i.e., the SID bit budget minus the bit budget of CN parameters and signaling). When the estimated metadata bit budget is lower than the available bit budget, or equal to it, the metadata are multiplexed into the SID frame bitstream. Otherwise the active frame coding is selected (see the decision to asses metadata as part of the DTX control mechanism in Fig. 2). This strategy considers that metadata is essential for the correct representation of immersive audio, so it needs to be transmitted in the interactive communications codec bitstream whenever it changes significantly.

D. Metadata adjustment in DTX

On the decoder side, the SID frames are received at the SID transmission interval C_{max}. It can thus happen that a metadata parameter changes between two subsequent SID frames with a large step, which can cause a subjective artifact. For example, the position of an audio object can change suddenly from one place to another. To avoid these spatial degradations, metadata parameter values are adjusted such that the differences of parameter values are reduced between adjacent frames. Specifically, an interpolation between the currently received and the previously (last) received metadata parameter values is applied in several frames following an SID frame, or following the first active frame. The adjustment is applied to every metadata parameter of every object such that the maximum difference of any metadata parameter between two adjacent frames is lower than a certain threshold. In IVAS, the threshold for azimuth and elevation is set to 5° [18].

The adjustment thus leads to a smooth evolution of metadata parameters. This effect is illustrated by an example in Fig. 3 for an azimuth metadata parameter which follows a circle trajectory. From top to bottom, there is a) 1 second of background noise audio signal of an object, b) codec bitrate indicating SID frames (corresponding to 5.2 kbps) and NO_DATA frames (corresponding to 0 kbps), c) input azimuth, d) decoded azimuth in DTX inactive frames without adjustment, g) decoded azimuth in DTX inactive frames with adjustment.

V. PERFORMANCE

The design trade-offs and the OBA methods described above have been thoroughly evaluated as part of the IVAS codec performance assessment. IVAS supports discrete OBA coding at bitrates suitable for conversational communications systems, i.e., from 13.2 kbps for one object, 16.4 kbps for two objects, and 48 kbps for three and four objects. We note that IVAS also supports a parametric OBA coding mode [24] for coding three or four objects at 24.4 and 32 kbps, not described in this paper. Then, the IVAS codec supports OBA rendering to mono, stereo, different loudspeaker layouts, ambisonics, and binaural outputs. The binaural rendering is able to take into account head-rotation and reverberation effects.

The algorithmic delay of the discrete IVAS OBA coding is 32 ms (excluding the HRTF-related delay), tailored to 5G communications services. We note that this delay is identical to that of the EVS codec, which was primarily developed for audio services in 4G communications networks. The computational and memory requirements of one coded object in IVAS are, in general, slightly lower than they are for coding a mono channel using EVS. For multiple objects, these requirements are then roughly proportional to the number of discretely coded objects so that the IVAS OBA coding can be efficiently implemented in mobile communications devices. Similarly, the low algorithmic delay, optimized resource requirements, wide range of bitrates and rendering capabilities, as well as support for error concealment and JBM make IVAS OBA also an option for various IoS communications applications and devices.

The subjective performance of the IVAS OBA coding was evaluated during the Selection phase of the 3GPP IVAS codec

standardization, and the results are summarized in a detailed report [25]. We show an excerpt from the results from one P.800 [26] experiment that evaluated the OBA system with two clean speech objects coded at low bitrates. We also present results from a MUSHRA [27] experiment with three generic audio objects coded at medium bitrates, where results from two independent laboratories are aggregated and shown. The plots include Mean Opinion Scores (MOS) or Continuous Quality Scale (CQS) scores, along with 95% confidence intervals based on Student's t-distribution. As IVAS is the first immersive codec that specifically targets conversational communications systems, a multi-mono EVS codec with uncoded metadata was chosen as a reference for comparison purposes. The coded samples were rendered for binaural headphones presentation, while the IVAS internal renderer with the default HRTF dataset [18] was used. Additional details about the test material, listeners, statistical analysis, etc., can be found in [25].

The P.800 test results in Fig. 4 show performance in DTX conditions without frame losses and in non-DTX conditions with 5% of lost frames at three bitrates. At all bitrates, IVAS performs better than the multi-mono EVS, where the difference is statistically significant at 16.4 kbps both in the DTX and the frame loss conditions. Next, the MUSHRA test results in Fig. 5 show that all tested conditions score in the *excellent* quality range (80% to 100%), and IVAS performs statistically better than the multi-mono EVS at 48 and 96 kbps. It is important to note that in these tests IVAS quantizes and transmits metadata as part of the bitstream, whereas the multi-mono EVS utilizes unquantized metadata. Moreover, the multi-mono EVS operates at multiples of its native bitrates, which, for several conditions, results in a higher overall bitrate than the corresponding IVAS bitrate. Although this gives the multi-mono EVS an advantage over IVAS, IVAS achieves higher subjective scores than the multi-mono EVS in all conditions.

VI. OBA IN COMBINATION WITH OTHER AUDIO FORMATS

OBA can be employed as a standalone audio scene representation or in a combination with another audio format, typically CBA or SBA, to represent complex audio scenes. A straightforward solution is a separate coding of individual formats. However, this results in a complex solution with high memory and computation complexity, as well as high transmission bitrates.

A novel approach was introduced in the IVAS codec which provides support for joint coding of OBA and either MASA (i.e., OMASA audio format [28]) or SBA (i.e., OSBA audio format [18]). These combined formats are particularly useful in teleconferencing scenarios, where individual talkers' voices are captured as audio objects while the environment is captured as a spatial bed using MASA or SBA format. The joint coding of two formats is supported in IVAS down to 13.2 kbps, where it employs a pre-rendering of objects into the MASA or SBA audio scene using the objects' metadata. With increasing bitrate, more audio objects are discretely coded, while at higher bitrates, all audio objects are coded discretely in addition to coding the MASA or SBA audio scene. These formats exploit redundancies between the objects and the spatial bed signal, achieving significant advantages over separate coding at similar overall bitrates in terms of decoded and rendered audio quality, as well as memory and computational complexity requirements [28].

VII. CONCLUSION

This paper describes newly developed methods for an efficient object-based audio coding system capable of delivering an immersive audio experience in modern mobile communications systems. These methods were adopted by the recently standardized IVAS codec, which for the first time enables new interactive communications services like immersive telephony, immersive experience sharing, multi-party immersive teleconferencing, or extended reality (XR) and metaverse real-time conversational services. Furthermore, IVAS supports generic API and common protocols, including RTP and SDP. It can be deployed in VoIP services, for example in IP Multimedia Subsystem (IMS) framework or over-the-top (OTT) communications applications.

IVAS's extensive features and capabilities also represent new possibilities that might be attractive in the IoS ecosystem. IVAS can become an interesting option to take advantage of a rigorously described immersive audio system in future IoS services and to bring new advancements to networked immersive audio experiences and applications.

REFERENCES

[1] J. Herre et al., "MPEG Spatial Audio Object Coding — The ISO/MPEG Standard for Efficient Coding of Interactive Audio Scenes," Journal of Audio Eng. Soc., vol. 60, no. 9, pp. 655—673, Sept. 2012.

[2] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H Audio - The new standard for universal spatial / 3D audio coding," Journal of Audio Eng. Soc., vol. 62, no. 12, pp. 821–830, Dec. 2014.

[3] F. Schuh et al., "Efficient Multichannel Audio Transform Coding with Low Delay and Complexity," in Proc. of Audio Eng. Soc. 141st Convention, paper #9660, Los Angeles, CA, USA, Sept. 2016.

[4] J. Herre and S. Disch, "MPEG-I Immersive audio – Reference model for the virtual / augmented reality audio standard," Journal of Audio Eng. Soc., vol. 71, no. 5, pp. 229–240, May 2023.

[5] K. Kjorling et al., "AC-4 — The next generation audio codec," in Proc. of Audio Eng. Soc. 140th Convention, paper #9491, Paris, France, Jun. 2016.

[6] H. Purnhagen, T. Hirvonen, L. Villemoes, J. Samuelsson, and J. Klejsa, "Immersive audio delivery using joint object coding," in Proc. of Audio Eng. Soc. 140th Convention, paper #9587, Paris, France, May 2016.

[7] M. Multrus et al., "Immersive Voice and Audio Services (IVAS) codec – The new 3GPP standard for immersive communication," in Proc. of Audio Eng. Soc. 157th Convention, paper #10188, New York, USA, Oct. 2024.

[8] L. Turchet et al., "The Internet of Sounds: Convergent Trends, Insights, and Future Directions," In Proc. of IEEE Internet of Things Journal, vol. 10, no. 13, pp. 11264-11292, July 2023.

[9] E. Fotopoulou et al., "Use-Cases of the new 3GPP Immersive Voice and Audio Services (IVAS) Codec and a Web Demo Implementation," in Proc. of the IEEE 5th Int. Symp. on the Internet of Sounds (IS2), pp. 1–6., Sept. 2024.

[10] L. Turchet, C. Rinaldi, C. Centofanti, L. Vignati, and C. Rottondi, "5G-Enabled Internet of Musical Things Architectures for Remote Immersive Musical Practices," IEEE Open Journal of the Communications Society, vol. 5, pp. 4691–4709, Aug. 2024.

[11] L. Vignati et al., "Is Music in the Air? Evaluating 4G and 5G Support for the Internet of Musical Things," in IEEE Access, vol. 12, pp. 38081–38101, March 2024.

[12] B. Bessette et al., "The adaptive multi-rate wideband speech codec (AMR-WB)," IEEE Trans. on Speech and Audio Processing, vol. 10, no. 8, pp. 620-636, Nov. 2002.

[13] M. Dietz et al., "Overview of the EVS codec architecture," in Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, Apr. 2015, pp. 5698–5702.

[14] Recommendation ITU-R BS.2076, "Audio Definition Model," International Telecommunication Union, 2019.

[15] S. Fug, et al., "Design, Coding and Processing of Metadata for Object-Based Interactive Audio," in Proc. of Audio Eng. Soc. 137th Convention, paper #9097, Los Angeles, USA, Oct. 2014.

[16] A. Vasilache, T. Pihlajakuja, and M.-V. Laitinen, "Metadata-assisted spatial audio coding in IVAS codec," in Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1–5.

[17] V. Eksler, "Bitrate adaptation in object-based audio coding in communication immersive voice and audio systems," in Proc. of Audio Eng. Soc.157th Convention, paper #10200, New York, USA, Oct. 2024.

[18] 3GPP TS 26.253, "Codec for Immersive Voice and Audio Services (IVAS); Detailed Algorithmic Description including RTP payload format and SDP parameter definitions," v 18.5.0, 2025.

[19] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," Journal of the Audio Eng. Soc., vol. 45, pp. 456–466, June 1997.

[20] C. Borß, "A polygon-based panning method for 3D loudspeaker setups," in Proc. of Audio Eng. Soc. 137th Convention, paper #9106, Oct. 2014.

[21] S. Li and J. Peissig, "Measurement of head-related transfer functions: A review," Applied Sciences, 2020.

[22] S. D. Kanik, E. Karlsson, T. Toftgard and E. Norvell, "Modeling of HR Filters for Audio Object Rendering," in Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1–5.

[23] 3GPP TS 26.445, "EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification," 2014.

[24] A. Eichenseer, S. Korse, G. Fuchs and M. Multrus, "Parametric Object Coding in IVAS: Efficient Coding of Multiple Audio Objects at Low Bit Rates," in Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1–5.

[25] 3GPP TR 26.997, "Codec for Immersive Voice and Audio Services (IVAS); Performance characterization," 2024.

[26] Recommendation ITU-T P Suppl 29, "ITU-T P.800 – Use Cases," International Telecommunication Union, 2023.

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

[28] M.-V. Laitinen, A. Vasilache, A. Rämö, J. Paulus and V. Eksler, "Combined object-based audio and MASA format for enhanced spatial mobile communication," in Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1–5.