Combined object-based audio and MASA format for enhanced spatial mobile communication

Mikko-Ville Laitinen*, Adriana Vasilache†, Anssi Rämö†, Jouni Paulus‡
Nokia Technologies
*Espoo, Finland, †Tampere, Finland, ‡Munich, Germany,
{mikko-ville.laitinen, adriana.vasilache, anssi.ramo, jouni.paulus}@nokia.com

Václav Eksler
Consultant for VoiceAge Corporation
Montreal, QC, Canada
vaclav@eksler.cz

***Abstract*—The new 3GPP IVAS codec for immersive voice and audio communication services in 5G networks supports a variety of spatial audio formats beyond mono and stereo voice and audio, and even combinations of them. This paper presents the OMASA combined audio format which provides a joint coding of object-based audio and metadata-assisted spatial audio (MASA). It enables immersive communication experiences, e.g., in teleconferencing scenarios where a mobile end-device captures the environmental audio in MASA format while individual talkers' voices are captured with their own microphones as separate audio objects. This paper describes coding methods newly developed and optimized depending on the available bitrate and the number of audio objects to efficiently deliver a 3D audio experience at a wide OMASA bitrate range from 13.2 kbps to 512 kbps. The benefits of the presented joint coding of audio formats over separate coding at the same bitrates are presented in terms of subjective evaluation and computational resources.**

***Index Terms*—Spatial Audio, MASA, Object-based audio, IVAS**

I. Introduction

Immersive audio coding has been present, until recently, mainly in the cinema and broadcasting industry. For example, the Motion Picture Experts Group (MPEG) developed the MPEG-H 3D audio codec [1] and MPEG-I immersive audio codec [2] for coding channel-based audio, object-based audio [3], and scene-based audio (SBA). Another example is ETSI Next Generation Audio, known as AC-4 [4], [5].

The new Immersive Voice and Audio Services (IVAS) codec [6] has been recently standardized within the 3rd Generation Partnership Project (3GPP). It delivers a 3D audio and voice experience for communication optimized for 5G mobile networks. To ensure backward compatibility for mono operation, IVAS is built upon EVS [7], the voice and audio coding standard for 4G LTE networks. IVAS opens new experiences by supporting stereo, traditional spatial audio formats (channel-based, object-based, and scene-based audio (SBA)), and a new spatial audio format called metadata-assisted spatial audio (MASA) [8] providing mobile spatial audio capture and coding.

Furthermore, the combinations of object-based audio and either MASA or SBA are supported by the IVAS codec. The feature of joint coding of two formats is specifically useful in teleconferencing scenarios, where the talkers may be captured as individual audio objects while the environment is captured in the spatial audio format. This paper focuses on the joint coding of object-based audio and MASA as a combined format called OMASA (objects with MASA). Consequently, while built on MASA and audio object coding methods described in [9], [10], the current work addresses the challenges present in the combined OMASA format while overcoming constraints of communication codecs compared to MPEG codecs.

The paper is structured as follows. A brief description of the format is followed by a description of four coding modes designed to cover the whole bitrate range of the IVAS codec and the different number of input audio objects. The content-dependent bitrate distribution between audio objects and MASA is described together with the corresponding coding. The evaluation section shows the advantages of joint coding of these two input formats in terms of output signal quality as well as memory and complexity requirements.

II. OMASA format

A. OMASA format overview

A detailed technological flow of OMASA is described in the IVAS codec specification [11]. OMASA consists of object-based audio represented by audio signals and metadata of up to four independent audio objects and MASA represented by two audio signals accompanied by MASA metadata [8]. The audio signals are processed in 20 ms long frames and coded at a constant bitrate (ranging from 13.2 kbps to 512 kbps).

A block diagram of the OMASA encoder is shown in Fig. 1. The encoder receives $N + 2$ input audio signals (N is the number of input audio objects and '2' of MASA input audio signals), and metadata of N input objects and MASA input metadata. A mono input signal for MASA is also supported in which case it is duplicated to dual mono signals before the processing. Next, one of four OMASA coding modes (Sec. II-B) is chosen in the configuration module based on the available bitrate and the number of input objects. Based on the coding mode, the input audio signals are subject to analysis, metadata coding, and audio signals mixing forming M objects transport audio signals and 2 MASA transport audio signals.

The transport audio signals corresponding to M objects are further subject to pre-processing, configuration, and coding using M mono core-coders (see [11]). The two MASA transport signals are similarly pre-processed, configured, and coded using a stereo core-coder (see [11]). The objects and MASA

 DOI: 10.1109/ICASSP49660.2025.10889750

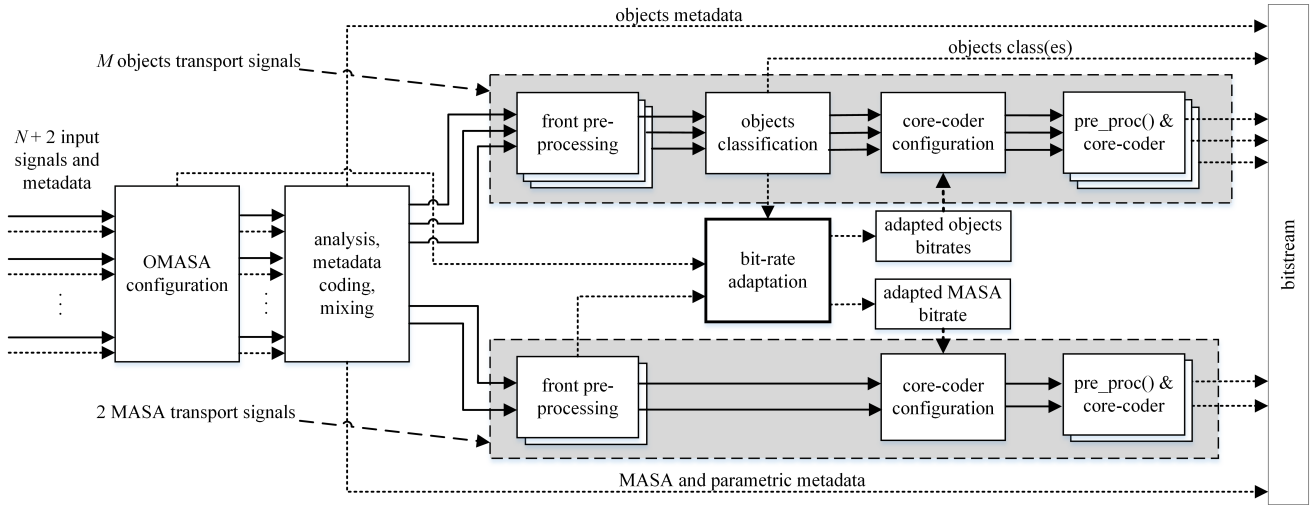

Fig. 1. Block diagram of the OMASA encoder. Audio signals are denoted by solid lines, and metadata and parameters by dotted lines.

TABLE I
OMASA CODING MODES IN IVAS.

Bitrate [kbps]	Number of objects 1	2	3	4
13.2–16.4	PreRend	PreRend	PreRend	PreRend
24.4	Disc	PreRend	PreRend	PreRend
32	Disc	OneParam	OneMasa	OneMasa
48	Disc	Disc	OneMasa	OneMasa
64–80	Disc	Disc	OneParam	OneParam
96	Disc	Disc	Disc	OneParam
128–512	Disc	Disc	Disc	Disc

metadata are then quantized and coded using native object-based audio and MASA formats' tools (see [11]). Finally, a bitrate adaptation logic (Sec. II-G) is introduced to efficiently distribute the available bit-budget between MASA and objects core-coders.

B. OMASA coding modes

The most straightforward approach to coding the OMASA format is a separate coding of MASA and all individual objects' audio signals and their metadata. However, this approach is suitable for coding only at medium and high bitrates, while there is not enough available bit-budget to use it at lower bitrates. Thus, pre-rendering and parametric approaches are used to enable the coding of OMASA at bitrates as low as 13.2 kbps. Consequently, four coding modes were designed and tuned to provide the best subjective experience for a particular available bitrate and number of audio objects. These coding modes are pre-rendering (*PreRend*), one object with MASA representation (*OneMasa*), one object with parametric representation (*OneParam*), and discrete (*Disc*) coding mode. Their use in IVAS is summarized in Table I.

C. Pre-rendering coding mode

In the pre-rendering (*PreRend*) coding mode, the objects are converted to object-based MASA representation (i.e., MASA spatial metadata and two transport audio signals computed from the objects). This representation is combined with the original MASA part, and the resulting MASA representation is coded as ordinary MASA format (see [9] for details on MASA coding), i.e., $M = 0$.

The object-based MASA spatial metadata is determined as follows. First, the object input audio signals $s(t, i)$ (where t is time and i is object index) are converted to the time-frequency (TF) domain using the complex low-delay filterbanks [12], yielding $S(k, n, i)$ (where k is the frequency bin and n the temporal slot). The TF-domain object audio signals are converted to First-Order Ambisonic (FOA) signals based on the object directions from the objects metadata. Using these FOA signals, direction and direct-to-total energy ratio TF-domain MASA parameters are determined using methods similar to directional audio coding (DirAC) [13]. The rest of the MASA parameters are set to zero. The resulting object-based MASA metadata is then merged with the original MASA metadata using the method described in [11].

The object-based MASA transport audio signals are determined from the object input signals as follows. First, the stereo amplitude panning gains $g(t, i, j)$ are determined for each object i and mixed channel j, based on the object directions, using vector base amplitude panning (VBAP) [14]. Then, the two transport audio signals are computed with

$$s_{obj,dm}(t, j) = \sum_{i=1}^{N} g(t, i, j) s(t, i). \quad (1)$$

These object-based MASA transport audio signals and the transport audio signals of the original MASA part are mixed to obtain two merged MASA transport audio signals.

Thus, in this mode, only two combined MASA transport audio signals and the combined MASA metadata are coded and transmitted.

D. One object with MASA coding mode

In the one object with MASA representation (*OneMasa*) coding mode, one audio object is adaptively selected in each

frame to be separated from the other objects. This object is separately coded using one mono core-coder, i.e. $M = 1$. The remaining objects are converted to MASA transport audio signals and metadata, mixed with the original MASA audio signals and metadata, and coded together as described in the *PreRend* mode above. In this mode, an improved coding and more accurate rendering of the selected object is thus achieved compared to the *PreRend* mode.

The object to separate from the other objects is determined as follows. First, the object with the largest energy ($E(i) = \sum_{t=t_1}^{t_2} s(t,i)^2$) is selected as the object to be separately coded. However, to avoid frequent switching of the selected object back and forth (which could cause subjective artifacts), the selected object is changed only when the energy of a new object is larger than the energy of the previously selected object by an adaptively selected threshold. The aim is to change the selected object when the MASA part can mask the change and disfavor the change of the object when there is no masking effect present. This is achieved by having a smaller threshold when the MASA signals are masking the object signals and a larger threshold when they are not. More details about the object separation can be found in [11].

Thus, in this mode, the two combined MASA transport audio signals and one separated object audio signal, combined MASA metadata and the separated object metadata are coded and transmitted.

E. One object with parametric representation coding mode

In the one object with parametric representation (*OneParam*) coding mode, one audio object is adaptively selected and coded, i.e., $M = 1$, similarly as in the *OneMasa* mode. However, in this coding mode, the remaining objects are not converted to the MASA representation but parametric TF-domain object metadata is computed as follows. First, the TF-domain object energies are computed as

$$E_{obj}(b,m,i) = \sum_{k=k_1(b)}^{k_2(b)} \sum_{n=n_1(m)}^{n_2(m)} |S(k,n,i)|^2, \qquad (2)$$

where k_1, k_2, n_1, and n_2 define the first and last indices of the frequency band b and subframe m. The energy is similarly computed also for the MASA transport audio signals ($E_{MASA}(b,m)$). Object energy ratios $r_{obj}(b,m,i)$ and MASA-to-total energy ratios $r_{M2t}(b,m)$ are computed as

$$r_{obj}(b,m,i) = \frac{E_{obj}(b,m,i)}{\sum_{i=1}^{N-1} E_{obj}(b,m,i)}, \qquad (3)$$

$$r_{M2t}(b,m) = \frac{E_{MASA}(b,m)}{E_{MASA}(b,m) + \sum_{i=1}^{N-1} E_{obj}(b,m,i)}. \qquad (4)$$

Using the object and MASA-to-total energy ratios, the decoded objects can be separated to some degree and thus can be edited on the receiving side (e.g., level and direction).

The audio signals of the remaining objects are converted into two object-based MASA transport audio signals and combined with the original MASA transport audio signals, as was done in the *OneMasa* mode.

Thus, in this mode, the two combined MASA transport audio signals, the separated object audio signal, the MASA metadata, the determined parametric TF-domain object metadata, and the original object metadata are coded and transmitted.

F. Discrete coding mode

In the discrete (*Disc*) coding mode, audio signals and metadata of all input audio objects are coded separately, i.e., $M = N$. This mode provides the highest subjective quality and enables the most flexibility in rendering and adjustments of the audio scene on the receiving side.

Thus, in this mode, the two MASA transport audio signals, N object transport audio signals, MASA metadata, and N objects metadata are coded and transmitted.

G. Bitrate adaptation

The OMASA format coding contains a bitrate adaptation logic that adaptively distributes the constant IVAS total bitrate, $brate_{IVAS}$, to variable MASA and audio objects bitrate parts. The audio objects part is further a sum of variable bitrates assigned to individually coded objects, i.e.,

$$brate_{IVAS} = brate_{MASA} + \sum_{i=1}^{M} brate_{obj}(i), \qquad (5)$$

where $brate_{MASA}$ is the adapted MASA bitrate and $brate_{obj}(i)$ is the adapted bitrate for coding object i, where $i = 1...M$.

The bitrate adaptation in OMASA extends the bitrate adaptation logic proposed for object-based audio in [10]. It is performed in every frame and consists of the following parts: 1) analyzing objects and MASA input audio signals in a front pre-processing, 2) objects classification, and 3) adaptation of objects and MASA nominal bitrates to adapted objects and MASA bitrates. The nominal bitrates are constant values and represent an average bitrate for coding the respective objects and MASA parts.

The objects classification classifies each audio object based on signal characteristics and Sound Activity Detection. It defines four importance classes: inactive, low importance, medium importance, and high importance class (details can be found in [10]). Moreover, long-term noise energy values of objects and MASA are employed to decide when a low bitrate is assigned to coding a specific object (e.g., if its background noise is masked by the MASA scene audio).

The bitrate adaptation logic then assigns a higher bitrate to objects with a higher importance class and a lower bitrate to objects with a lower importance class, i.e.,

$$brate_{obj}(i) = \gamma_{class}(i) \cdot brate_{obj,nom}(i), \qquad (6)$$

where $brate_{obj,nom}(i)$ is the nominal bitrate for coding object i. The value of the weighting factor $\gamma_{class}(i)$ depends on the importance class, and is in the range from 0.8 to 1.4. Based on the IVAS total bitrate, the coding mode, and the number of objects, five sets of weighting factors were experimentally found and defined for OMASA coding in IVAS, see [11] for details. Once the adapted bitrates for coding all objects are computed, the adapted MASA bitrate is determined using (5).

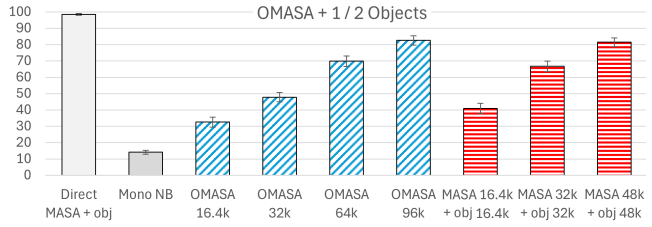

Fig. 2. OMASA listening test results for MASA and 1 or 2 objects.

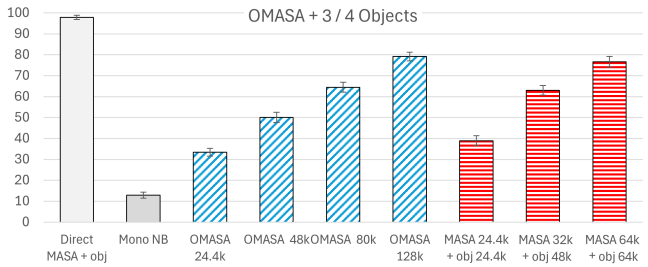

Fig. 3. OMASA listening test results for MASA and 3 or 4 objects.

H. Decoder and renderer

On the receiving side, the MASA and objects metadata and MASA and M objects transport audio signals are decoded from the OMASA bitstream and supplied to the renderer [11]. In IVAS, the rendering can be done to several output setups including mono, stereo, multichannel (from 5.1 to 7.1.4), Ambisonics (order 1 - 3), and binaural (with or without reverberation, and with and without head-tracking).

III. Evaluation

A. Subjective evaluation

Two MUSHRA listening tests were conducted, both with six expert listeners and a binaural headphone presentation. The audio scenes in both tests were complex: in addition to speech objects, about half of the samples contained musical objects while a majority of them contained overlapping objects. The tests compared the subjective quality of OMASA to the separate coding of MASA and audio objects at approximately the same total bitrates. The first test consisted of 13 different OMASA samples with MASA and 1 (five samples) or 2 (eight samples) objects. The results are shown in Fig. 2. As can be seen from the results, OMASA at 32 kbps is significantly better than separate coding of MASA at 16.4 kbps and objects at 16.4 kbps. OMASA shows quality benefits also at the total bitrates of 64 and 96 kbps. Additionally, OMASA enables combined coding down to 13.2 kbps, which is not possible with separate coding.

The second test consisted of 18 different OMASA samples with MASA and 3 (six samples) or 4 (twelve samples) objects. The results are shown in Fig. 3. It can be seen that OMASA has significantly better quality at 48 kbps than a separate coding of MASA at 24.4 kbps and objects at 24.4 kbps. Also, OMASA at 80 and 128 kbps show a quality improvement over the separate coding. Even the lowest tested OMASA bitrate 24.4 kbps provides adequate quality for practical applications.

B. Computational resources

The OMASA coding as part of the floating-point IVAS codec implementation was evaluated in terms of computational resources. The computational complexity (in worst-case Weighted Millions Operations Per Second, WMOPS) and static RAM requirements were measured using the WMC tool [15]. A comparison between OMASA and a separate coding of MASA and audio objects at the same total bitrates is provided in Table II for the input of MASA with 4 objects decoded to binaural output using an IVAS internal renderer and sampled at 48 kHz. As IVAS cannot provide coding of objects and MASA separately at the total bitrate of 24.4 kbps, those numbers are missing in the table.

TABLE II
Complexity and memory requirements comparing OMASA coding vs. separate coding of MASA and audio objects.

Variant	Bitrate [kbps]	Encoder		Decoder	
		WMOPS	RAM [kB]	WMOPS	RAM [kB]
OMASA	24.4	146	286	126	341
separate		-	-	-	-
OMASA	48	229	390	167	536
separate		275	465	249	777
OMASA	80	225	359	181	568
separate		377	657	331	1076
OMASA	160	403	631	305	1120
separate		395	629	326	1108

Table II shows that the complexity and memory requirements highly depend on the OMASA coding mode. The requirements are the least for the *PreRend* mode (24.4 kbps), higher for *OneMasa* and *OneParam* modes (48 kbps and 80 kbps), and the highest in the *Disc* mode (160 kbps). When comparing the requirements for OMASA and separate coding of MASA and objects, significantly lower computational resources are needed for OMASA in all modes, except for *Disc* mode, for which they are very similar. The complexity of the combined format is reduced by up to 40% (encoder) and 45% (decoder). Similarly, the static RAM consumption is reduced by up to 45% (encoder) and 47% (decoder) when compared to the separate coding.

IV. Conclusions

This paper investigates a newly developed OMASA combined format that extends the possibilities of immersive voice and audio experience delivered by the recently standardized 3GPP IVAS communication codec. Listening tests show that OMASA scores better than the separate coding of MASA and audio objects at the tested operating points, some of which have statistical significance. Moreover, OMASA allows coding down to low bitrates at which the separate coding is not possible. The proposed approach also significantly reduces the requirements for computational resources compared to the separate coding.

REFERENCES

[1] J. Herre *et al.*, "MPEG-H Audio - The new standard for universal spatial / 3D audio coding," *Journal of the Audio Engineering Society*, vol. 62, no. 12, pp. 821–830, Dec. 2014.

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

[3] J. Herre *et al.*, "MPEG Spatial Audio Object Coding - the ISO/MPEG standard for efficient coding of interactive audio scenes," *Journal of the Audio Engineering Society*, vol. 60, no. 9, pp. 655–673, Sep. 2012.

[4] K. Kjörling *et al.*, "AC-4 — The next generation audio codec," in *Proc. of AES 140th Convention*, Paris, France, Jun. 2016.

[5] H. Purnhagen *et al.*, "Immersive audio delivety using joint object coding," in *Proc. of AES 140th Convention*, Paris, France, 2016.

[6] M. Multrus *et al.*, "Immersive Voice and Audio Services (IVAS) codec – The new 3GPP standard for immersive communication," in *Proc. of AES 157th Convention*, New York, USA, 2024.

[7] 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.

[8] J. Paulus *et al.*, "Metadata-assisted spatial audio (MASA) - an overview," in *Proc. of the IEEE Int. Symposium on the Internet of Sounds*, Erlangen, Germany, 2024.

[9] 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.

[10] V. Eksler, "Bitrate adaptation in object-based audio coding in communication immersive voice and audio systems," in *Proc. of AES 157th Convention*, New York, USA, 2024.

[11] 3GPP Technical Specification 26.253, "Codec for Immersive Voice and Audio Services; Detailed algorithmic description incl. RTP payload format and SDP parameter definition," 3GPP, TS, 2024.

[12] M. Schnell *et al.*, "Low delay filterbanks for enhanced low delay audio coding," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 235–238.

[13] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, pp. 503–516, Jun. 2007.

[14] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, pp. 456–466, Jun. 1997.

[15] Recommendation ITU-T G.191 (2024), "Software tools for speech and audio coding standardization," ITU-T, Recommendation, May 2024.