Metadata-Assisted Spatial Audio (MASA) – An Overview

Jouni Paulus
Nokia Technologies
Munich, Germany
jouni.paulus@nokia.com

Lasse Laaksonen
Nokia Technologies
Tampere, Finland
lasse.j.laaksonen@nokia.com

Tapani Pihlajakuja
Nokia Technologies
Espoo, Finland
tapani.pihlajakuja@nokia.com

Mikko-Ville Laitinen
Nokia Technologies
Espoo, Finland
mikko-ville.laitinen@nokia.com

Juha Vilkamo
Nokia Technologies
Espoo, Finland
juha.vilkamo@nokia.com

Adriana Vasilache
Nokia Technologies
Tampere, Finland
adriana.vasilache@nokia.com

***Abstract*—This paper provides an overview of the novel metadata-assisted spatial audio (MASA) format which is one of the supported input formats in the 3GPP IVAS codec. MASA format spatial audio consists of a transport audio signal with one or two channels and parametric metadata describing the dominant directional sound and the diffuseness and coherence properties of the scene. While MASA is mainly intended for parametric spatial audio acquisition from mobile devices, this paper describes a way to determine the parameters from an Ambisonics signal as an example using a widely-available source format. Additionally, the file format for the MASA spatial metadata ingested by the IVAS codec is described. These descriptions are supported by the provided software tools, consisting of a C-language implementation of the described analysis, and a Python-language library for reading and writing the metadata files.**

***Index Terms*—spatial audio, metadata, audio format, MASA, IVAS, Ambisonics**

I. Introduction

Immersive audio provides the listener the feeling of being inside the sound scene. Providing this immersion has been a part of the various surround sound formats available in consumer electronic devices, e.g., [1]–[5]. Most of these focus on coding of channel-based content, i.e., the audio signals of individual loudspeaker channels are coded. In such a format, the scene is normally static which poses limitations to the immersion in the scene. To address this limitation, formats that allow interaction with the audio content and formats where the input, transport, and reproduction channel layouts are decoupled have been proposed, e.g., [4]–[6] and recently extended even to a format allowing defining entire virtual audio scenes [7].

The formats above already provide quite credible immersion in the audio scene. However, they have been developed in the view of the application with offline encoding and decoding without strict constraints on end-to-end delay, computational complexity, or bitrate. These are important aspects in communication audio codecs, such as the ones standardized in the Third Generation Partnership Project (3GPP), e.g., the widely-deployed Enhanced Voice Services (EVS) codec [8]–[10].

For this reason, the recently-completed 3GPP Release 18 (”5G-Advanced”) includes the standardization of a new codec for Immersive Voice and Audio Services (IVAS) [11]. For the interested reader, [12] provides an overview of the standardization work and the resulting codec. The IVAS codec is the first fully-immersive communications codec designed for 5G mobile systems. Building on top of the EVS mono codec, IVAS offers additional support for coding and rendering of stereo, multi-channel, Scene-Based Audio (SBA, Ambisonics), objects, and Metadata-Assisted Spatial Audio (MASA).

IVAS enables spatial calls where the call participants can share their surrounding audio scene with the other end, or where multiple teleconference participants can use a single mobile device while still obtaining spatial separation, for example. The important mobile applications in mind, MASA is a new parametric audio format designed for optimal spatial audio pick-up directly from smartphones, and it can be coupled with light-weight binaural rendering.

The IVAS codec facilitates the coding, transport, decoding, and rendering of immersive audio in various supported formats at a wide range of bitrates and with low algorithmic delay. The codec also provides jitter buffer management (JBM) and packet loss concealment (PLC) support for addressing sub-optimal network conditions. The IVAS standard [11] also defines the RTP (Real-time Transport Protocol) payload format and SDP (Session Description Protocol) parameters. In 3GPP Release 18, the MTSI (Multimedia Telephony Service for IMS) clients in terminals offering speech communication are recommended to support immersive audio communication based on IVAS [13].

IVAS supports binaural output, with and without room effect, with and without headtracking, and using custom head-related transfer functions (HRTF). This makes it a candidate for addressing the challenge of interoperable or standardized connection between devices [14], [15] in Internet-of-Sounds (IoS) field.

Considering IoS applications, e.g., in the context of cultural heritage [16], [17] where the binaural rendering is envisioned to be placed at the 5G edge, IVAS supports head-tracked binaural output with split rendering. In case of split rendering, the decoding and main rendering takes place on the edge device, and the light-weight end-device receives a prototype binaural signal with parametric information allowing it to perform head orientation adjustment on the output with low computational load.

The algorithmic delay of the IVAS codec is 32–38 ms depending on the input and output audio formats [11, Table 4.4-1]. The delay aspects of different parts of the system architecture in different IoS applications, mostly in the scope of NMP (networked music performance) have been evaluated, e.g., in [15], [18]–[20].

This paper provides an overview of the MASA format starting with a motivation for the new format and a technical description of the MASA metadata file format and the included parameters in Sec. II. Sec. III provides a description of an example audio front-end for estimating MASA metadata parameters from audio signals, and Sec. IV describes how the parameters can be used in rendering of the output signal, providing more concrete intuition on their meaning. Sec. V describes how MASA format content is coded in IVAS and provides an overview of other coding technologies in the IVAS codec that are based on the MASA format: McMASA for low-bitrate multi-channel coding and OMASA as a combined format of audio objects and MASA. Sec. VI provides selected listening test results demonstrating the performance of parametric spatial audio from mobile devices and IVAS MASA operation.

In short, this paper demonstrates that it is possible to obtain a parametric description of an immersive sound scene using only a small number of microphones in a device designed under constraints other than a dedicated microphone array, encode and decode this with a wide range of bitrates with a standardized codec, and to render this into a high-quality immersive sound scene, all with low computational complexity and algorithmic delay suitable for a communication codec.

The description in this paper is supported by software tools[1], consisting of a C-language implementation for computing the MASA metadata parameters, and a Python-language library for reading and writing the metadata files.

II. MASA SPATIAL AUDIO FORMAT

This section provides a motivation for the new parametric spatial audio format, and a brief technical description of it including the file format of the associated metadata files.

A. Background of the MASA format

While no communications codec nor service support has been available for immersive audio before the IVAS codec, multi-microphone mobile devices allowing spatial audio capture for user-generated content were first introduced to the market already about a decade ago. Due to their challenging form factors (i.e., flat devices with large screens), integration of irregular microphone arrays has been a necessity for these designs. Moreover, acoustic design has not been always a design priority.

TABLE I
MASA METADATA FRAME STRUCTURE

Descriptive metadata	Spatial metadata			
	Sub-frame 1 Spatial metadata	Sub-frame 2 Spatial metadata	Sub-frame 3 Spatial metadata	Sub-frame 4 Spatial metadata

TABLE II
MASA SPATIAL METADATA STRUCTURE FOR 1- AND 2-DIRECTIONAL METADATA

1-directional	Direction 1 Spatial metadata	Common spatial metadata	
2-directional	Direction 1 Spatial metadata	Direction 2 Spatial metadata	Common Spatial metadata

Thus, dedicated spatial audio capture processing has been required for multi-microphone mobile devices, based on an understanding of the various device properties including the microphone configuration, allowing for deriving the directional information of captured sounds as suitable parameters. Based on the parameterization of the spatial audio capture, it is possible to synthesize also other spatial audio representations. On the other hand, the highest audio quality, e.g., in binaural rendering, is achieved by directly rendering based on the parametric representation itself, thus avoiding any lossy format conversions. At the same time, the parametric spatial audio representations used on smartphones are based on a limited number of audio channels and a relatively small number of spatial parameters, which makes these representations attractive from coding perspective.

Such light-weight parametric format was soon understood to be ideal also for introducing immersive audio communications in 3GPP, as the program for standardization of IVAS was launched. Maintaining optimal quality of the original spatial audio capture based on smallest amount of data to be coded and transmitted while avoiding the complexity and potential quality degradation of format conversions allow for efficient and attractive immersive audio especially for smartphones. The metadata-assisted spatial audio (MASA) format was introduced to enable those features.

B. Technical description of the MASA format

The following description of the MASA parameters and the metadata file format is largely based on the standard specification [21, Annex A].

The IVAS codec, like EVS, operates using a 20-ms framesize. In order to enable the best possible integration, the MASA format similarly utilizes frame length of 20 ms with division into 4 sub-frames (each 5 ms in length). Furthermore, to allow for optimization of the associated computational complexity, a frequency resolution compatible with the complex-valued low-delay filterbank (CLDFB) [22] used in the EVS and IVAS codecs is utilized in the format.

[1] https://github.com/nokia/MASA_tools_for_IVAS

The MASA format consists of mono or stereo transport audio signals and the associated metadata. Each MASA metadata frame corresponding to 20 ms of audio signal is structured as illustrated in Table I: the *Descriptive metadata* field is for the whole frame followed by four *Spatial metadata* sub-frames. Each *Spatial metadata* sub-frame consists of one or two *Directional spatial metadata* fields and a *Common spatial metadata* field, depending on whether the data contains a description for one or two simultaneous directions, as illustrated in Table II. The MASA frames are self-contained and independent, so, e.g., the number of directions may change between frames.

The contents of the *Descriptive metadata* are shown in Table III. It consists of a 64-bit format identifier followed by a 16-bit *Channel audio format* field combining the information of the following fields: the number of simultaneous directions (i.e., *Directional spatial metadata* fields), the number of channels in the transport audio signal, the source format of the content from which the MASA metadata was created from, and a variable description field containing further parameters based on the source format. Further details are available in [21, Annex A.4]. A single *Descriptive metadata* block is defined for the entire frame.

The *Spatial metadata* is provided for a time-frequency resolution of 5-ms temporal sub-frames and 24 frequency bands. The frequency bands are non-uniformly spaced and cover the frequency range of 0–24 kHz. The lowest 20 bands are 400 Hz in width, corresponding to the resolution of the 60-bin CLDFB representation in IVAS, and the four highest bands are 2000 Hz, 2000 Hz, 4000 Hz, and 8000 Hz wide. Lower sampling rates (16 kHz and 32 kHz) and audio bandwidths (wideband, WB and super-wideband, SWB, in addition to the full-band, FB) are directly supported by ignoring values in the corresponding higher parameter bands. For each of the 24 sub-bands there are 4 sub-frames, forming 96 time-frequency (TF) tiles. The full set of spatial metadata parameters are defined for all 96 TF tiles.

The parameters contained in the spatial metadata are divided into *Common spatial metadata* and one or two *Directional spatial metadata*. The *Directional spatial metadata* consists of three parameters describing sound with a direction of arrival: direction index, direct-to-total energy ratio, and spread coherence. Direction index [23] encodes the direction of arrival (azimuth and elevation angles in the spherical coordinate system) using indices of approximately equidistantly distributed points on a unit sphere. Applying a 16-bit index, approximately 1-degree accuracy is achieved for any spatial direction. Direct-to-total energy ratio is a value in the range 0..1 describing how much of the energy in the specific TF-tile is directional energy coming from the direction defined by the direction index. Spread coherence is a parameter that defines if the sound in the given direction is representative of a point-like source, a spatially spread phantom source, or in between of the two extremes.

TABLE III
MASA DESCRIPTIVE METADATA FIELD STRUCTURE

	Field	#Bits	Content
	Format descriptor	64	IVASMASA 8 ASCII values in 8-bit unsigned integers
Channel audio format	Number of directions	1	(number of directions in this frame) - 1
	Number of channels	1	(number of channels in the transport audio signal) - 1
	Source format	2	source audio format
	Variable description	12	additional info on source audio format or zero-pad

The *Common spatial metadata* contains parameters that are independent of the direction of the dominant sound source(s): diffuse-to-total energy ratio, surround coherence, and remainder-to-total energy ratio. The diffuse-to-total energy ratio is related to the direct-to-total energy ratio and describes the ratio of the energy in the TF-tile that does not have a definite direction, i.e., that is non-directional (e.g., diffuse or ambient). In some cases, e.g., if there is a known amount of microphone noise or some other energy unrelated to the captured sound field, the level of this energy is described with the remainder-to-total energy ratio. The total sum of the direct-to-total ratios (one or two), diffuse-to-total ratio, and remainder-to-total ratio is required to be exactly 1. The last parameter, the surround coherence, describes the coherence of the non-directional sound over surrounding directions. In other words, whether the non-directional energy in the TF-tile (described by the diffuse-to-total energy ratio) should be rendered as coherent or as incoherent (e.g., using decorrelation).

The data is ordered in the metadata file such that each frame begins with the *Descriptive metadata* shown in Table III, which also contains the information of the number of directions in the *Directional spatial metadata*. This is followed by the data for each complete sub-frame as shown in Table IV. The field for each parameter contains the values for all 24 bands.

C. Reader / writer library

To enable easier interaction with MASA format metadata files, a Python-language library is provided[2]. At the time of the publication of this paper, the library supports the MASA metadata format used by the IVAS codec implementation [21].

III. EXAMPLE AUDIO ANALYSIS FRONT-END

The analysis front-end for acquiring MASA format is usually created specifically for a given device or a certain type input audio signals. The commonality is that there is a minimum requirement to be able to analyze the direction of arrival and the direct-to-total energy ratio for a time-frequency resolution compatible with the format specification. For mobile device capture, the MASA parameters are determined from the signals of the (usually irregular) array of microphones with device-specific optimized algorithms. The details on such product designs are typically not available in the public. However, for Ambisonics input, methods related to the Directional

[2] https://github.com/nokia/MASA_tools_for_IVAS

TABLE IV
MASA DIRECTIONAL SPATIAL METADATA FIELD STRUCTURE FOR EACH METADATA SUB-FRAME. THE VALUES OF EACH PARAMETER FOR THE 24 BANDS ARE GIVEN CONSECUTIVELY.

	Field	#Bits	Content
Direction 1 spatial metadata	Direction index	16 * 24	16-bit unsigned integer
	Direct-to-total energy ratio	8 * 24	8-bit unsigned integer, uniform 0..1
	Spread coherence	8 * 24	8-bit unsigned integer, uniform 0..1
Direction 2 spatial metadata (if present)	Direction index	16 * 24	16-bit unsigned integer
	Direct-to-total energy ratio	8 * 24	8-bit unsigned integer, uniform 0..1
	Spread coherence	8 * 24	8-bit unsigned integer, uniform 0..1
Common spatial metadata	Diffuse-to-total energy ratio	8 * 24	8-bit unsigned integer, uniform 0..1
	Surround coherence	8 * 24	8-bit unsigned integer, uniform 0..1
	Remainder-to-total energy ratio	8 * 24	8-bit unsigned integer, uniform 0..1

Audio Coding (DirAC) [24] can be used to obtain the MASA parameters.

This section describes how the MASA format is formed from the Ambisonics input. The direction and the ratio parameters are obtained based on the DirAC methodology, with slight modifications in order to obtain parameters compatible with the MASA format. The coherence parameters are obtained with dedicated methods developed for the MASA format.

As the MASA format contains the option for using one or two simultaneous directions, the following Sections III-A and III-B present parameter analysis for both.

A. One-directional analysis

The one-directional analysis determines one dominant sound source direction for each TF-tile. This begins by having the input signal in the 4-channel first-order Ambisonics (FOA) format. The time-domain signals are first transformed into the frequency domain using a filterbank, e.g., STFT or CLDFB. The resulting signals are denoted as

$$\mathbf{s}(n,k) = \begin{bmatrix} s_w(n,k) \\ s_x(n,k) \\ s_y(n,k) \\ s_z(n,k) \end{bmatrix}, \tag{1}$$

where n is the time index, k is the frequency bin index, and s_w, s_x, s_y, s_z are the four Ambisonics basis signals consisting of zeroth-order omnidirectional component and the three first-order components.

The overall signal energy in parameter TF-tile (t,b) is computed with

$$e(t,b) = \sum_{n\in t, k\in b} \frac{|\mathbf{s}(n,k)|^2}{2}, \tag{2}$$

where t is the temporal sub-frame index and b is the frequency band index. Then, a directional vector related to the intensity vector[3] is computed with

$$\begin{aligned} \mathbf{i}(t,b) &= \begin{bmatrix} i_x(t,b) \\ i_y(t,b) \\ i_z(t,b) \end{bmatrix} \\ &= \sum_{n\in t, k\in b} \Re\left\{ s_w(n,k) \begin{bmatrix} s_x^*(n,k) \\ s_y^*(n,k) \\ s_z^*(n,k) \end{bmatrix} \right\}, \end{aligned} \tag{3}$$

where $\Re\{x\}$ denotes taking the real part of the complex-valued number x, and x^* is the complex conjugate of x.

In the following description the time and frequency tile indices are dropped for clarity, as all parameters and equations are defined for each TF-tile identically.

The direction of arrival of the dominant sound is determined to have the azimuth angle of

$$\theta = \arctan \frac{i_y}{i_x}, \tag{4}$$

and the elevation angle of

$$\phi = \arctan \frac{i_z}{\sqrt{i_x^2 + i_y^2}}, \tag{5}$$

where $\arctan()$ resolves the correct quadrant. The azimuth and elevation angles are stored in a single spherical index [23] instead of two separate values.

The direct-to-total energy ratio is estimated with

$$r_{dir} = \frac{\mathbb{E}\{|\mathbf{i}|\}}{\mathbb{E}\{e\}}, \tag{6}$$

where $\mathbb{E}$ denotes the expectation operator, which can be approximated with temporal averaging.

Then, a measure of general coherence is computed from the FOA component signals with

$$C_{gen} = 1 - \frac{\mathbb{E}\{s_x^2\} + \mathbb{E}\{s_y^2\} + \mathbb{E}\{s_z^2\}}{\mathbb{E}\{s_w^2\}}. \tag{7}$$

Using this value, the spread coherence can be computed with

$$C_{spr} = r_{dir} C_{gen}, \tag{8}$$

[3] To be exact, this vector points to the direction-of-arrival, so it is pointing to the opposite direction as the intensity vector.

and the surround coherence with

$$C_{sur} = r_{diff} C_{gen}, \tag{9}$$

where the diffuse-to-total energy ratio is

$$r_{diff} = 1 - r_{dir}. \tag{10}$$

B. Two-directional analysis

The two-directional MASA metadata analysis in the provided tool is based on the principles described in [25]. The input must be higher-order Ambisonics (HOA) of the second order, or possible to convert into that representation. The second-order HOA signal $\mathbf{s}_{HOA2}$ is converted into two FOA signals representing the left and right halves of the sound scene using the filters $\mathbf{h}_L$ and $\mathbf{h}_R$ with

$$\mathbf{s}_L = \mathbf{h}_L \mathbf{s}_{HOA2} \tag{11}$$

$$\mathbf{s}_R = \mathbf{h}_R \mathbf{s}_{HOA2}. \tag{12}$$

The intensity-related direction vectors, signal energies, sound source directions, and energy ratios are computed for both sectors independently similar to the one-directional analysis. The signal energies of the sectors computed with (2) are e_L and e_R, and the per-sector direct-to-total energy ratios computed with (6) are $r_{dir,L}$ and $r_{dir,R}$. These are combined to overall direct-to-total energy ratio parameters with

$$r_{dir,1} = \frac{r_{dir,L} e_L}{e_L + e_R} \tag{13}$$

$$r_{dir,2} = \frac{r_{dir,R} e_R}{e_L + e_R}. \tag{14}$$

The overall diffuse-to-total energy ratio is then

$$r_{diff} = 1 - r_{dir,1} - r_{dir,2}. \tag{15}$$

In addition to the two sector-FOA signals, a third overall scene-FOA signal is computed and the general coherence value C_{gen} is computed from it with (7). The per-sector spread coherences are compute from this using the per-sector direct-to-total energy ratios

$$C_{spr,1} = r_{dir,1} C_{gen} \tag{16}$$

$$C_{spr,2} = r_{dir,2} C_{gen}. \tag{17}$$

The surround coherence is computed as

$$C_{sur} = \left(1 - \frac{r_{dir,1} + r_{dir,2}}{2}\right) C_{gen}. \tag{18}$$

C. Transport audio signal

The transport audio signal is an important source of information in the MASA format. For the mono transport case, the signal can be generated either by a mono downmix (passive or active) of the original audio channels or by selecting a single channel (e.g., the omnidirectional channel W from the FOA signals).

For the stereo transport channels, similar alternatives exist: selecting two specific microphones from the original microphone-array audio signals (e.g., in the case of Eigenmike input, selecting a microphone on the left side of the array and the right side of the array) and possibly applying post-filtering on them, or by forming a virtual stereo microphone pair signals from the original audio signals (e.g., in the case of Ambisonics signals, creating cardioid signals pointing to ±90 degrees).

D. Analysis software

To enable easier creation of MASA format files from raw Eigenmike microphone-array signals and from Ambisonics audio signals, a C-language implementation of the analysis described above is provided[4].

IV. Rendering

The rendering of the MASA format spatial audio stream with one direction to multi-channel loudspeaker, binaural, and stereo signals is concisely described in this section. For more details (including the two-direction MASA rendering), please see the algorithmic description of the IVAS standard [11]. For the rendering of MASA format files (created, e.g., with the analysis software of Sec. III-D), one can use the *IVAS renderer*, which is a part of the codec implementation [21].

A. Multi-channel loudspeaker rendering

The spatial synthesis is performed based on the transport audio signals and the MASA metadata to recreate the sound scene in a perceptual sense. The basic principles of the rendering follow the common principles of parametric spatial audio rendering methods (as described, e.g., for DirAC rendering [24]). Amplitude panning methods are used for steering the sound to the desired directions and decorrelation techniques are used for rendering the diffuse portion of the sound incoherently to surrounding directions. The results of the two processing paths are combined to form the output sound.

The rendering uses a prototype signal in both the direct, or panned, and the decorrelated paths. It is generated from the transport audio signals by associating the left transport channel to the left-side loudspeakers, and the right transport channel to the right-side loudspeakers. The sum of the channel signals normalized by $\sqrt{0.5}$ is used as the prototype signal for the median plane loudspeakers.

The rendering of the directional part is based on vector-base amplitude panning (VBAP) [26]. VBAP determines panning gains by triangulating the loudspeaker setup and using a triplet of loudspeakers to pan sounds to directions in between them. The triangulation requires that the physical locations of the loudspeakers should cover all directions around the listener. When this is not satisfied, the loudspeaker setup is appended with a virtual top, bottom, or rear loudspeakers, depending on the arrangement. The panning gains associated with any added virtual loudspeakers are distributed to the neighboring real loudspeakers.

[4] https://github.com/nokia/MASA_tools_for_IVAS

The processing therefore produces a set of gains values for the azimuth θ and elevation ϕ in each TF-tile

$$\mathbf{g}(\theta,\phi) = \begin{bmatrix} g_1(\theta,\phi) \\ g_2(\theta,\phi) \\ \vdots \\ g_N(\theta,\phi) \end{bmatrix}, \tag{19}$$

where N is the number of loudspeakers, not including the virtual ones. The gains are normalized so that $\mathbf{g}\mathbf{g}^T = 1$. Due to the usage of the virtual loudspeakers and the corresponding gain distribution, there may be more than three non-zero entries in the gain vector.

If the spread and surround coherence parameters are zero, $C_{spr} = C_{sur} = 0$, the rendering of the direct sound is simplified into applying the channel gains $\sqrt{r_{dir}}\mathbf{g}(\theta,\phi)$ to the prototype signals, applying gain $\sqrt{\frac{r_{diff}}{N}}$ to the decorrelated versions of the prototype signals, and adding these results together. However, to account for the spread and surround coherence parameters C_{spr} and C_{sur}, the processing parameters $\mathbf{g}(\theta,\phi)$, r_{diff}, and r_{dir} are modified as follows.

Firstly, omitting the indices (θ,ϕ), three panning gain vectors $\mathbf{g}_c$, $\mathbf{g}_+$ and $\mathbf{g}_-$ are determined. where $\mathbf{g}_c = \mathbf{g}$, and $\mathbf{g}_+$ and $\mathbf{g}_-$ contain panning gains formulated for $(\theta + 30°, \phi)$ and $(\theta - 30°, \phi)$, i.e., $\pm 30°$ on the sides of the sound direction. Then, gain modifiers g_c and g_s are formulated as

$$g_c = \begin{cases} 1 - 2C_{spr} + \frac{2C_{spr}}{\sqrt{3}}, & \text{if } C_{spr} < 0.5 \\ 2 - 2C_{spr}, & \text{otherwise} \end{cases} \tag{20}$$

and

$$g_s = \begin{cases} \frac{2C_{spr}}{\sqrt{3}}, & \text{if } C_{spr} < 0.5 \\ 1, & \text{otherwise} \end{cases}. \tag{21}$$

Then, spread coherent panning gains are formulated by

$$\mathbf{g}_{spr} = g_c\mathbf{g}_c + g_s(\mathbf{g}_+ + \mathbf{g}_-) \tag{22}$$

after which $\mathbf{g}_{spr}$ is normalized so that $\mathbf{g}_{spr}\mathbf{g}_{spr}^T = 1$.

The above formulation causes that when

- $C_{spr} = 0$, the panning is only to the direction (θ,ϕ), i.e., the source is rendered as point-like,
- $C_{spr} = 0.5$, the panning is equally to all three directions (θ,ϕ) and $(\theta \pm 30°, \phi)$, i.e., the source is rendered coherently from these spread directions,
- $C_{spr} = 1$, the panning uses the sides only, i.e., the source is rendered as a phantom source, and
- interpolation takes place between these values of C_{spr}.

The surround coherence parameter C_{sur} also affects the rendering. Firstly, the diffuse and direct ratio parameters are updated by

$$r'_{diff} = (1 - C_{sur})r_{diff} \tag{23}$$

and

$$r'_{dir} = r_{dir} + r_{diff}C_{sur}. \tag{24}$$

Then, the panning gains are modified by

$$r_{sur} = \frac{r_{diff}C_{sur}}{r_{dir} + r_{diff}C_{sur}} \tag{25}$$

$$\mathbf{g}' = \mathbf{g}_{spr}\sqrt{1 - r_{sur}} + \sqrt{\frac{r_{sur}}{N}} \tag{26}$$

The gains are then normalized so that $\mathbf{g}'\mathbf{g}'^T = 1$. Finally, the modified processing values $\mathbf{g}'$, r'_{diff}, and r'_{dir} are used in place of the original ones.

The above formulation causes that when C_{sur} becomes higher, the r'_{diff} becomes smaller and less sound is produced as surrounding incoherent ambience. At the same time, the amplitude panning gains and r'_{dir} are modified so that the spatial sound is reproduced with more surrounding coherence.

B. Binaural and stereo rendering

The rendering for binaural and stereo uses the covariance-matrix-based rendering technique described in [27]. The approach is to measure the input covariance matrix of the transport audio signal, determine a target covariance matrix for the output signal based on the spatial metadata, and formulating a mixing matrix to achieve the processing according to these matrices. The method also includes means to inject decorrelated sound when there is not enough independent signal energy in the transport signals to achieve the result by only mixing.

For the binaural output, the directional part of the target covariance matrix is built based on the spatial metadata using an HRTF database. For each frequency and direction to be rendered the complex-valued HRTF pair is denoted with

$$\mathbf{h}_{HRTF} = \begin{bmatrix} h_{HRTF,L} & h_{HRTF,R} \end{bmatrix}^T. \tag{27}$$

The corresponding covariance matrix is defined as

$$\mathbf{C}_{HRTF} = \mathbf{h}_{HRTF}\mathbf{h}_{HRTF}^*. \tag{28}$$

This matrix is weighted with the direct-to-total energy ratio value r_{dir} to get the direct part of the target covariance matrix. The diffuse part target covariance matrix is determined by weighting a diffuse field binaural covariance matrix with the diffuse-to-total energy ratio r_{diff}. If the spread and surround coherence parameters are zero, $C_{spr} = C_{sur} = 0$, the target covariance matrix is the sum of these two component matrices, further weighted by the overall signal energy.

The spread coherent sounds are handled similarly as in Sec. IV-A, such that the direct part of the target covariance matrix contains mutually coherent portions at directions (θ,ϕ), $(\theta + 30°, \phi)$ and $(\theta - 30°, \phi)$. The HRTF vectors $\mathbf{h}_{HRTF}$ at these three directions are weighted based on g_c and g_s, and the result is converted to the direct part covariance matrix as above.

The surround coherent sound is controlled by adjusting the inter-channel coherence of the diffuse part, so that when $C_{sur} = 0$, the inter-channel coherence is according to the binaural diffuse-field inter-aural coherence, and when $C_{sur} = 1$ the diffuse part covariance matrix indicates full coherence.

Headtracking can be accounted for in the rendering by rotating the direction parameters when formulating the direct part covariance matrix. In addition, if the listener is facing a rear direction, the transport audio signals are swapped prior to the rendering for obtaining better prototype signals.

The stereo rendering is performed with the same system with the difference that the diffuse part covariance matrix is a diagonal matrix and that the HRTF data is replaced with an amplitude-panning function.

V. MASA in IVAS

MASA format spatial audio is supported by the IVAS codec natively. The codec contains also coding techniques for low-bitrate multichannel coding based on MASA, and a combined format for joint coding of MASA and audio objects. These are described in the following.

A. Coding of MASA format content in IVAS

The IVAS codec supports MASA format input at the full bitrate range from 13.2 kbps to 512 kbps. The MASA transport audio channels are coded using the one audio channel IVAS encoder unit or with audio channel pair IVAS encoder unit depending on whether one or two transport audio channels are used. The total bitrate is split between the transport audio codec and the metadata codec, and the transport audio codec receives any signal-dependent excess from the metadata codec. Approximately 12% to 19% of the total bitrate is given as a maximum for the metadata and rest is given for the transport signals to maintain the best possible overall quality.

The uncompressed MASA spatial metadata is 273.8 kbps or 422.4 kbps, depending on number of directions present. Since this would be prohibitively large, even in a compressed form, in most of the target IVAS bitrates, the codec applies several metadata reduction techniques in order to keep the perceptually most relevant data while achieving the required target bitrate. This includes, most importantly, reducing the raw number of parameter sets by merging values over time (sub-frames) and/or frequency bands, and merging of the directions inside individual TF-tiles, all depending on the content and the bitrate constraints. After the reduction of the raw number of metadata parameter sets, the metadata is quantized and coded. Notably, most of the metadata reductions steps are lossy, while the perceptual quality is optimized to minimize the effect of the metadata coding to the overall audio quality.

For the details on MASA metadata reduction, encoding, and decoding in IVAS, the interested reader is referred to [11] and [21].

B. Other MASA-family technologies in IVAS

The underlying signal model of MASA can be applied also for multi-channel content (Multi-channel MASA, McMASA) and extended for a combined coding of audio objects and MASA, known as Objects with MASA (OMASA).

McMASA mode, used for lower bitrates in IVAS, introduces adaptations to MASA metadata coding, as the source of the MASA format is known, and the metadata is formed within the encoder. This allows analyzing the metadata directly using the appropriate time/frequency-resolution based on the bitrate, and the direction metadata compression can take advantage of the most likely directions of arrival based on the known input content channel layout. Furthermore, very low-bitrate parametric LFE-channel coding is introduced, and, when the bitrate allows it, the center channel is separated into its own transport audio stream, removing it from the parametric model.

Likewise, OMASA introduces techniques for combined encoding of audio objects and MASA. It is envisioned to be used, e.g., in teleconferencing applications where each (up to 4) talker has a dedicated microphone and they can be represented as audio objects while the rest of the audio scene is captured by a mobile device or a microphone array into a MASA format bed. OMASA employs four different coding operating modes, depending on the total bitrate available and the number of objects. These are in the order of increasing bitrate and output quality:

- Encoder-side mixing of the audio objects into the MASA format bed and delivering everything in the MASA format.
- The dominant audio object is coded as a separate object, while the rest are mixed into the MASA format bed.
- Similar to the earlier mode, except that parametric representation of the objects mixed into the MASA format bed is formed. This allows parametric reconstruction of the objects at the decoder.
- The audio objects are encoded independently of the MASA format bed, including dynamic bitrate allocation between the objects and the MASA format bed.

The details of the McMASA and OMASA operation can be seen in the algorithmic description of the IVAS standard [11].

VI. Performance

In the above, the description of generating MASA format spatial audio signal is exemplified using the Ambisonics input. This is also provided in the example analysis software. However, the main motivation for a parametric format, such as MASA, is to allow direct pick-up from a generic microphone array, e.g., as implemented as a part of a smartphone audio capture design, for uncompromised quality by avoiding lossy format conversions. Sec.VI-A below describes an evaluation of such smartphone capture. Then, an excerpt of the selection test results [28] is shown in Sec.VI-B demonstrating the performance of using MASA format input compared to naïve FOA transport with multiple EVS-coded signals at various bitrates.

A. Parametric spatial audio capture

In the experiment, the perceptual quality of a smartphone spatial audio pickup was evaluated, based both on a parametric model and on an Ambisonics representation, using two audio capture devices: one with three microphones (allowing horizontal planar spatial audio capture) and one with four microphones (allowing full 3D spatial audio capture).

The recordings carried out for the experiment cover several signal types related to rich immersive communications and user-generated content. The recordings were performed in a controlled environment in three separate recording sessions. The recorded sound sources include both loudspeaker playback (of well-known immersive audio test samples, background music, nature sound clips, individual instruments, etc.) as well as live sound sources (male and female talkers, a violinist, etc.). Thus, the sound scenes correspond to, e.g., multi-channel music playback, conference-room discussions with or without background sounds, live music performances, outdoor scenes, and so on.

The position of the capture device was different in each of the recording sessions and its relative position to sound sources also varied during some sessions (due to placement and movement of the sources). The capture device was typically mounted on a tripod or similar support either in the middle of the recording space or on a conference room table. Sound sources provided either full or partial coverage of the horizontal plane (depending on the sample). The elevation component was generally fairly small, ranging between a maximum of about 45 degrees of elevation above and less than that below the horizontal plane.

The target of the comparison was to evaluate the quality achievable in an uncompressed audio transmission. The parametric format was similar to the MASA format: two uncompressed transport audio channels with a spatial metadata stream that was using approximately 25 kbps. This format was rendered using a parametric rendering scheme, similar to one discussed in IV. The first-order Ambisonics (FOA) signal was rendered using two alternative solutions, a linear binauralization method and a parametric binauralization method.

In the *Linear* rendering, the left and right ear signals were computed by applying Ambisonics-to-binaural processing filters that were designed as processing matrices in the STFT domain, in each frequency bin independently. At frequencies below 1300 Hz, the processing matrices were complex-valued and designed to approximate the real and imaginary parts of the target HRTF dataset in a least-square sense. Above this frequency, the processing matrices were real-valued and were designed to approximate the absolute values of the HRTF dataset.

The ITU-T P.800 Comparison Category Rating (CCR) test methodology [29] was used with 24 listeners. The listening test was conducted using Sennheiser HD650 headphones (with no headtracking) in quiet listening booths. Each of the 24 listeners listened to 120 sample pairs, where each individual sample pair was listened to in both directions (i.e., AB and BA) by each listener with one out of six randomized listening orders. Thus, a total of 2880 votes were given. Out of the 24 listeners 14 were audio experts (not all in the field of spatial audio), while 10 listeners were naïve. Mono signal was used as a low reference, and it was compared against the parametric format. In addition, a stereo signal (without spatial processing) was compared to itself as a control condition.

Fig. 1 presents the subjective test results. The results indicate that the parametric format provides a significantly higher perceived spatial audio quality than FOA, with both binauralization approaches. No significant differences were observed on sample level or sample group level attributable to differences between the capture devices (3-microphone and 4-microphone devices) or in terms of sound scenes. Thus, in order to achieve uncompromised spatial audio quality, it is highly beneficial to avoid any unnecessary format conversions.

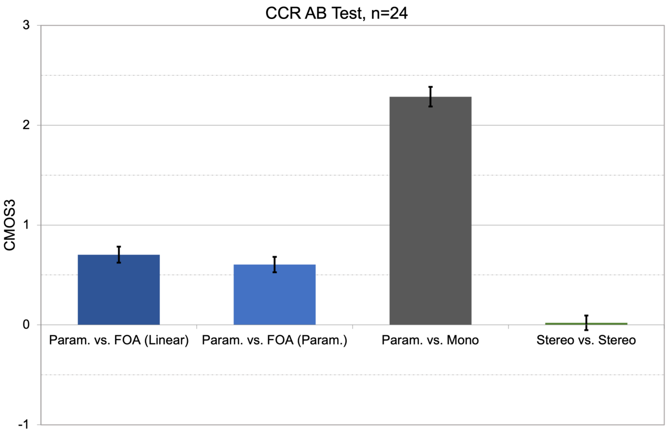

Fig. 1. Subjective listening test results from N=24 listeners comparing mobile device MASA-like parametric capture (*Param*) against first-order Ambisonics (*FOA*) with *Linear* and *Param*etric binauralization and *Mono* low anchor. *Stereo vs. Stereo* without spatial processing acts as a control for the listener reliability. The test methodology is Comparison Category Rating (CCR) [29] where the Comparison Mean Opinion Score (CMOS) scale is $-3 \ldots +3$ (a positive value means a preference towards the first condition and negative towards the second condition). The wide bars show the mean ratings and the black narrow bars the 95% confidence intervals using Student's t-distribution.

B. Parametric representation vs. discrete-channel FOA

Additionally, the performance of IVAS MASA format input operation has been evaluated as a part of the 3GPP IVAS Selection tests that were conducted in 2023 by external test laboratories. The full documentation of the results is available in [28]. Here, one relevant test is described as an example. In this P.800 DCR (ITU-T P Suppl. 29) [30] test, the IVAS performance was evaluated for MASA format inputs under clean speech conditions (i.e., with no background noise). The source material is based on Danish sentences converted into an Ambisonics representation and then into MASA format using methods similar to the ones discussed in Sec. III-D.

In the experiment, MASA format input performance of IVAS is evaluated against a multi-instance EVS coding of the corresponding (4 first-order) Ambisonics channels at the closest corresponding bitrates. According to the results, shown in Fig. 2, IVAS with MASA format input at 13.2 kbps achieves a quality level of the EVS-based approach operating at 38.4 kbps. IVAS with MASA format input at 24.4 and 32 kbps provide similar quality as the EVS-based approach at 65.6 kbps. These results show the high quality achievable using the IVAS codec with MASA format inputs and the benefit of a parametric spatial audio format over discrete coding of the Ambisonics channels.

C. Complexity

The IVAS codec is currently expected to operate at three capability levels that correspond to maximum complexities of 3xEVS, 6xEVS, and 10xEVS comparing to the complexity

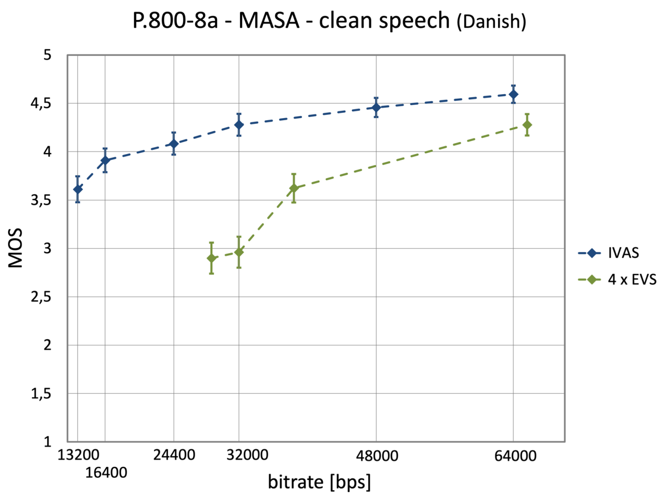

Fig. 2. Subjective evaluation results reproduced from [28], using P.800 methodology [30] on Danish language speech without background noise. The test compares MASA format coding to 4 x EVS FOA at similar total bitrates. The used scale is the Degradation Category Rating (DCR) rating scale ranging from 1 (”very annoying”) to 5 (”inaudible”) and expressed as Mean Opinion Score (MOS). Modified listener instructions according to [30] were used.

of the EVS codec [8], i.e., ca. 390, 780, and 1300 WMOPS (weighted millions of operations per second), respectively [31]. Based on TS 26.258 (floating point C code) [21], the IVAS codec with MASA format input is estimated to have a maximum complexity of ca. 297 WMOPS for the total of encoding, decoding, and binaural rendering for head-tracked presentation, being clearly below the lowest complexity limit. The full complexity estimation of the IVAS codec is available when TS 26.251 (fixed point C code) [32] is completed.

VII. Conclusions

This paper has provided an overview of the new format of Metadata-Assisted Spatial Audio (MASA), which is primarily targeted for parametric spatial audio acquisition on mobile devices in the context of 3GPP communication services utilizing the new IVAS codec. The MASA format consists of a mono or stereo transport audio signal and parametric metadata describing the main perceptual factors of a spatial sound scene. The paper has described the spatial metadata parameters, their meaning, how they can be computed, how they are stored in the metadata files, and how they are used in the reproduction of the sound scene.

The MASA format is defined as a part of the 3GPP IVAS standard and is supported at all immersive operation bitrates between 13.2 kbps and 512 kbps with an algorithmic delay of 37 ms. The MASA format is extended in IVAS also for multi-channel signal encoding and for joint encoding of a MASA format with audio objects. Listening tests show the benefit of utilizing the original parametric representation (e.g., captured on a smartphone) for transmission and the quality achievable with IVAS and the MASA format compared to using multiple EVS streams for delivering Ambisonics at similar total bitrates.

Complexity measurements indicate that the MASA content encoding, decoding, and rendering to head-tracked binaural output is well below the maximum complexity of the lowest envisioned complexity level corresponding to three times the complexity of the EVS codec. Thus, using parametric MASA format sound scene capture front-end with low-bitrate IVAS coding for the transport is an option for having a standard-compliant immersive audio available from and for IoS devices with low computational complexity.

Acknowledgment

The authors would like to thank Lauros Pajunen for providing the complexity measurements and Anssi Rämö for the listening test results.

References

[1] S. M. F. Smyth *et al.*, “DTS coherent acoustics delivering high-quality multichannel sound to the consumer,” *Journal of the Audio Engineering Society*, May 1996.

[2] J. Couling, “Dolby Digital surround systems,” *Journal of the Audio Engineering Society*, vol. ASC-19, Jun. 1999.

[3] J. Herre *et al.*, “MPEG Surround - the ISO/MPEG standard for efficient and compatible multichannel audio coding,” *Journal of the Audio Engineering Society*, vol. 56, no. 11, pp. 932–955, Nov. 2008.

[4] 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.

[5] J. Riedmiller *et al.*, “Delivering scalable audio experiences using AC-4,” *IEEE Transactions on Broadcasting*, vol. 63, no. 1, pp. 179–201, 2017.

[6] 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.

[7] 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.

[8] 3GPP, “Universal mobile telecommunications system (UMTS); LTE; 5G; Codec for Enhanced Voice Services (EVS); Detailed algorithmic description,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 26.445, 2024, v18.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1467

[9] S. Bruhn *et al.*, “Standardization of the new 3GPP EVS codec,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Australia, Apr. 2015, pp. 5703–5707.

[10] M. Dietz *et al.*, “Overview of the EVS codec architecture,” in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Australia, Apr. 2015, pp. 5698–5702.

[11] 3GPP, “Codec for Immersive Voice and Audio Services; Detailed algorithmic description incl. RTP payload format and SDP parameter definition,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 26.253, 2024, v18.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3319

[12] 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, New York, Oct. 2024.

[13] 3GPP, “IP multimedia subsystem (IMS); multimedia telephony; media handling and interaction (release 18),” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 26.114, 2024, v18.7.0.

[14] L. Turchet *et al.*, “The Internet of Audio Things: State of the art, vision, and challenges,” *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 10 233–10 249, 2020.

[15] L. Turchet *et al.*, “The internet of sounds: Convergent trends, insights, and future directions,” *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 264–11 292, 2023.

[16] C. Rinaldi *et al.*, “On the exploitation of 5G multi-access edge computing for spatial audio in cultural heritage applications,” *IEEE Access*, vol. 9, pp. 155 197–155 206, 2021.

[17] F. Martusciello *et al.*, "Edge-enabled spatial audio service: Implementation and performance analysis on a MEC 5G infrastructure," in *2023 4th International Symposium on the Internet of Sounds*, 2023.

[18] P. Cairns, H. Daffern, and G. Kearney, "A DAW-based approach to immersive audio system evaluation in network music performance contexts," in *Proc. of AES International Conference on Spatial and Immersive Audio*, Huddersfield, UK, Aug. 2023.

[19] L. Turchet *et al.*, "5G-enabled internet of musical things architectures for remote immersive musical practices," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 4691–4709, 2024.

[20] L. Turchet and P. Casari, "On the impact of 5G slicing on an internet of musical things system," *IEEE Internet of Things Journal*, 2024, to appear.

[21] 3GPP, "Codec for Immersive Voice and Audio Services; C code (floating-point)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 26.258, 2024, v18.1.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3324

[22] 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.

[23] S. Ragot and A. Vasilache, "Spherical vector quantization for spatial direction coding," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023.

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

[25] A. Politis and V. Pulkki, "Higher-order direction audio coding," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds. John Wiley & Sonds, Ltd, 2017, pp. 141–159.

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

[27] J. Vilkamo, T. Bäckström, and K. Achim, "Optimized covariance domain framework for time–frequency processing of spatial audio," *Journal of the Audio Engineering Society*, vol. 61, pp. 403–411, Jul. 2013.

[28] 3GPP, "IVAS codec performance characterization," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 26.997, 2024, v18.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3323

[29] International Telecommunication Union, *ITU-T Recommendation P.800: Methods for subjective determination of transmission quality*, Geneva, Switzerland, 1996.

[30] International Telecommunication Union, *ITU-T P Suppl. 29 (01/2023): ITU-T P.800 – Use cases*, Geneva, Switzerland, 2023.

[31] "IVAS design constraints (IVAS-4)," IVAS Permanent Document S4-231031, accessed 2024-08-28. [Online]. Available: https://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/IVAS_Permanent_Documents/IVAS-4_S4-231031.zip

[32] 3GPP, "Codec for Immersive Voice and Audio Services; C code (fixed-point)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 26.251, draft. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3319