**Source: Nokia Corporation, Orange**

**Title:** **Proposal for Metadata-assisted spatial audio**

**Document for: Agreement**

**Agenda Item: 7.5 – IVAS\_Codec**

# 1. Introduction

IVAS support for spatial audio based on smartphone capture has been actively discussed since the Fukuoka meeting (SA4#97) leading to proposal of the Metadata-assisted spatial audio (MASA) as one IVAS encoder format. The MASA spatial metadata parameter set was later agreed at the Kochi meeting (SA4#100). Specifically, a proposal on MASA definition of spatial metadata (Table 1 of S4-181216) was agreed in SA4#100 to be later inserted as annex to IVAS-4 (once MASA is accepted) and documented in [1].

In addition to spatial metadata, the MASA format proposal includes descriptive metadata required for correct data ingest of variable size metadata chunks and for providing additional information potentially useful for encoding, decoding, or rendering. Example creation and rendering of the MASA format is possible using the IVAS MASA C Reference Software, which was originally made available at SA4#106 Busan meeting [2].

The most recent version of the software package is available in [3]. In [4], updates to MASA format and the software are proposed.

# 2. Proposal

Following the updates proposed in [4], the MASA format is more concise, and the open points from previous discussions have been addressed. The IVAS MASA C Reference Software is also improved in terms of computational complexity and by ensuring full coverage of the 216 points on the spherical grid.

We propose to agree on updating the Encoder Input Formats in IVAS-4 to include the MASA format as follows:

|  |  |
| --- | --- |
| **Encoder Input Formats** | Metadata-assisted spatial audio according to definition in Annex A.  [Spatial audio, [N] channels and spatial metadata defined by [TBD].] |

The MASA format in Annex A is proposed to be agreed as a basis for further work.

The relevant Output Formats for Spatial audio are captured by:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Output Formats** | The IVAS codec shall support the following output formats for the corresponding input format:   |  |  | | --- | --- | | **Encoder Input Format** | **Output Format** | | Metadata-assisted spatial audio | Metadata-assisted spatial audio, Binaural audio, Stereo, Mono  Multi-channel on arbitrary loudspeaker configurations of up to [K] speakers. | |

Here the Metadata-assisted spatial audio Output Format corresponds to Pass-Through Operation, where, e.g., use of external rendering can be assumed.

# References

[1] S4-181223: Draft report from SA4#100 EVS SWG

[2] S4-191167: Description of the IVAS MASA C Reference Software, Nokia Corporation

[3] S4-210840: Updates to IVAS MASA C Reference Software, Nokia Corporation

[4] S4-220443: MASA format updates, Nokia Corporation, Orange *(input to this meeting)*

# Annex A:

# Metadata-assisted spatial audio (MASA) format

This Annex describes the Metadata-assisted spatial audio (MASA) format. The MASA format consists of audio signals and metadata. The audio signals for MASA can be mono or stereo. The metadata is provided according to a structure defined here, and it comprises descriptive metadata and spatial metadata, as defined below.

Editor’s Note: Audio signal description for mono and stereo audio will be part of IVAS-7.

# MASA format metadata structure

MASA format input to IVAS encoder follows the 20-ms frame size. For each 20-ms audio frame, one corresponding metadata frame is provided. Each metadata frame is structured as illustrated in Figure A.1. The descriptive metadata common for the whole frame is written first. This is followed by the spatial metadata, which consists of four spatial metadata subframes, each corresponding to 5 ms of audio. The structure of the spatial metadata subframes depends on the number of direction parameters in the frame. There are two options for the structure, illustrated in Figure A.2 and Figure A.3 for one direction and two directions, respectively.

Descriptive common metadata

(Table A.1)

Subframe 1

Spatial metadata

(Figure A.2 & A.3)

Subframe 2

Spatial metadata

(Figure A.2 & A.3)

Subframe 3

Spatial metadata

(Figure A.2 & A.3)

Subframe 4

Spatial metadata

(Figure A.2 & A.3)

Spatial metadata

MASA metadata frame

Figure A.1: Metadata structure for one MASA input signal frame

Direction 1

Spatial metadata

(Table A.2a)

Common

Spatial metadata

(Table A.2b)

Figure A.2: MASA spatial metadata structure for one subframe with one direction

Direction 1

Spatial metadata

(Table A.2a)

Direction 2

Spatial metadata

(Table A.2a)

Common

Spatial metadata

(Table A.2b)

Figure A.3: MASA spatial metadata structure for one subframe with two directions

Table A.1 presents the MASA descriptive common metadata parameters in order of writing. The definitions and use of the descriptive metadata parameters are described in clause A.3.

Table A.2a and Table A.2b present the MASA spatial metadata parameters dependent and independent of the number of directions, respectively. The definitions and use of the spatial metadata parameters are described in clause A.4.

Table A.1: MASA format descriptive common metadata parameters

|  |  |  |
| --- | --- | --- |
| **Field** | **Bits** | **Description** |
| **Format descriptor** | 64 | Defines the MASA format for IVAS. Eight 8-bit ASCII characters:  01001001, 01010110, 01000001, 01010011,  01001101, 01000001, 01010011, 01000001  Values stored as 8 consecutive 8-bit unsigned integers. |
| **Channel audio format** | 16 | Combined following fields stored in two bytes.  Value stored as a single 16-bit unsigned integer. |
| **Number of directions** | (1) | Number of directions described by the spatial metadata.  Each direction is associated with a set of direction dependent spatial metadata.  Range of values: [1, 2] |
| **Number of channels** | (1) | Number of transport channels in the format.  Range of values: [1, 2] |
| **Source format** | (2) | Describes the original format from which MASA was created. |
| **(Variable description)** | (12) | Further description fields based on the values of ‘Number of channels’ and ‘Source format’ fields.  When all bits are not used, zero padding is applied. |

Table A.2a: MASA format spatial metadata parameters (dependent of number of directions)

|  |  |  |
| --- | --- | --- |
| **Field** | **Bits** | **Description** |
| **Direction index** | 16 | Direction of arrival of the sound at a time-frequency parameter interval. Spherical representation at about 1-degree accuracy.  Range of values: “covers all directions at about 1° accuracy”  Values stored as 16-bit unsigned integers. |
| **Direct-to-total energy ratio** | 8 | Energy ratio for the direction index (i.e., time-frequency subframe).  Calculated as energy in direction / total energy.  Range of values: [0.0, 1.0]  Values stored as 8-bit unsigned integers with uniform spacing of mapped values. |
| **Spread coherence** | 8 | Spread of energy for the direction index (i.e., time-frequency subframe).  Defines the direction to be reproduced as a point source or coherently around the direction.  Range of values: [0.0, 1.0]  Values stored as 8-bit unsigned integers with uniform spacing of mapped values. |

Table A.2b: MASA format spatial metadata parameters (independent of number of directions)

|  |  |  |
| --- | --- | --- |
| **Field** | **Bits** | **Description** |
| **Diffuse-to-total energy ratio** | 8 | Energy ratio of non-directional sound over surrounding directions.  Calculated as energy of non-directional sound / total energy.  Range of values: [0.0, 1.0]  (Parameter is independent of number of directions provided.)  Values stored as 8-bit unsigned integers with uniform spacing of mapped values. |
| **Surround coherence** | 8 | Coherence of the non-directional sound over the surrounding directions.  Range of values: [0.0, 1.0]  (Parameter is independent of number of directions provided.)  Values stored as 8-bit unsigned integers with uniform spacing of mapped values. |
| **Remainder-to-total** **energy ratio** | 8 | Energy ratio of the remainder (such as microphone noise) sound energy to fulfil requirement that sum of energy ratios is 1.  Calculated as energy of remainder sound / total energy.  Range of values: [0.0, 1.0]  (Parameter is independent of number of directions provided.)  Values stored as 8-bit unsigned integers with uniform spacing of mapped values. |

# MASA format time-frequency resolution

The MASA spatial metadata parameters describe the spatial characteristics of the captured spatial sound scene. This parametric representation is based on frequency bands. A certain spatial characteristic thus relates to a frequency band, and a neighbouring frequency band can exhibit a different characteristic. For MASA format, 24 frequency bands are used. Table A.3 presents these frequency bands.

The metadata frame corresponding to 20-ms frame of audio is divided into four subframes of 5 ms each, which allows for higher temporal resolution of the spatial characteristics than offered by the frame size. The parametric representation in each frame therefore consists of 24 frequency bands in 4 time slots giving a total of 96 time-frequency tiles.

When a frame describes the scene using one spatial direction, there are 96 instances of each of the spatial metadata parameters corresponding with the 96 time-frequency tiles. When a frame describes the scene using two spatial directions, there are two values per time-frequency tile for some of the spatial metadata parameters. In this case, there are 192 instances of those spatial metadata parameters in one metadata frame.

Table A.3. MASA spatial metadata frequency bands

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Band | LF (Hz) | HF (Hz) | BW (Hz) | Band | LF (Hz) | HF (Hz) | BW (Hz) |
| 1 | 0 | 400 | 400 | 13 | 4800 | 5200 | 400 |
| 2 | 400 | 800 | 400 | 14 | 5200 | 5600 | 400 |
| 3 | 800 | 1200 | 400 | 15 | 5600 | 6000 | 400 |
| 4 | 1200 | 1600 | 400 | 16 | 6000 | 6400 | 400 |
| 5 | 1600 | 2000 | 400 | 17 | 6400 | 6800 | 400 |
| 6 | 2000 | 2400 | 400 | 18 | 6800 | 7200 | 400 |
| 7 | 2400 | 2800 | 400 | 19 | 7200 | 7600 | 400 |
| 8 | 2800 | 3200 | 400 | 20 | 7600 | 8000 | 400 |
| 9 | 3200 | 3600 | 400 | 21 | 8000 | 10000 | 2000 |
| 10 | 3600 | 4000 | 400 | 22 | 10000 | 12000 | 2000 |
| 11 | 4000 | 4400 | 400 | 23 | 12000 | 16000 | 4000 |
| 12 | 4400 | 4800 | 400 | 24 | 16000 | 24000 | 8000 |

# MASA descriptive metadata parameters

The MASA descriptive metadata is provided once per frame. It includes information for correctly reading the metadata frame and information relating to creation of the current MASA format signal and its transport audio signals that can be used to assist encoding or rendering of the spatial audio.

Format descriptor (64 bits)

The unique format descriptor code is provided at the beginning of every MASA format metadata frame. It specifies MASA format for the IVAS codec.

|  |  |  |
| --- | --- | --- |
| Required bit value | Decoded value | Additional description |
| 01001001, 01010110, 01000001, 01010011,  01001101, 01000001, 01010011, 01000001 | “IVASMASA” | Unique format descriptor |

Channel audio format (16 bits as specified below)

Two bytes providing the following individual fields:

* Number of directions
* Number of channels
* Source format

and a variable 12-bit description configured based on ‘Number of channels’ and ‘Source format’.

Number of directions (1 bit)

This parameter field indicates how many directions are described in current MASA format frame. Size of the metadata chunk associated with the current frame depends on the number of directions.

|  |  |  |
| --- | --- | --- |
| Bit value | Decoded value | Additional description |
| 0 | 1 direction |  |
| 1 | 2 directions |  |

Number of channels (1 bit)

This parameter field indicates how many transport channels are used for the MASA format. This parameter is required in some form to read the correct number of channels. Some additional channel format descriptors depend on the number of channels.

|  |  |  |
| --- | --- | --- |
| Bit value | Decoded value | Additional description |
| 0 | 1 channel |  |
| 1 | 2 channels |  |

Source format (2 bits)

This parameter field describes the format of source signals that were used to form the MASA format input file/stream. This parameter provides additional information that can benefit encoding, decoding, and/or rendering. First bit value (00) is the default value.

|  |  |  |
| --- | --- | --- |
| Bit value | Decoded value | Additional description |
| 00 | Default/Other | Audio originates from unknown format(s) including mixed sources |
| 01 | Microphone grid | Audio originates from various (irregular) microphone grids (e.g., smartphones or other UEs) |
| 10 | Channel-based | Audio originates from premixed channel-based audio (e.g., 5.1) |
| 11 | Ambisonics | Audio originates from Ambisonics format |

Variable description (12 bits including zero padding)

Based on the values of the ‘Number of channels’ bit and ‘Source format’ bits, the variable description is configured to provide up to three additional fields to further describe the source format or transport channels. This information can guide, e.g., metadata encoding and rendering. The following presents the possible field combinations and their definitions.

**Source format == 00 (Default/Other)**

If number of channels is 1 (bit value 0), no additional metadata is specified. Instead, 12-bit zero padding is applied.

If number of channels is 2 (bit value 1), following additional fields are configured in order:

* Transport definition field (3 bits). This field describes the configuration of the two transport channels. The possible bit values and corresponding configurations are provided in Table 3.
* Channel angle field (3 bits). This field describes symmetric angle positions for transport signals with directivity patterns. In this notation, 0° corresponds to the front. The bit values and corresponding configuration are defined in Table 4.
* Channel distance field (6 bits). The bit values and corresponding configuration are defined in Table 5.

Table 3. Transport definition field for Source formats: Default/Other and Microphone grid

|  |  |  |
| --- | --- | --- |
| Bit value | Decoded value | Additional description |
| 000 | Unknown/Other |  |
| 001 | Omni |  |
| 010 | Subcardioid |  |
| 011 | Cardioid |  |
| 100 | Supercardioid |  |
| 101 | Hypercardioid |  |
| 110 | Dipole |  |
| 111 | Binaural |  |

Table 4. Channel angles for directive patterns for Source formats: Default/Other and Microphone grid

|  |  |  |
| --- | --- | --- |
| Bit value | Decoded value | Additional description |
| 000 | Unspecified |  |
| 001 | ±90 deg. |  |
| 010 | ±70 deg. | XY |
| 011 | ±55 deg. | XY, ORTF |
| 100 | ±45 deg. | NOS, XY, Blumlein |
| 101 | ±30 deg. |  |
| 110 | ±0 deg. | AB. Needs spacing for any stereo image |
| 111 | Reserved |  |

Note: If Transport definition value is “Unknown”, “Omni”, or “Binaural”, value 000 is assumed.

The channel distance parameter is defined with a few predefined values and the distance values between 0.01 m and 1 m are calculated as an equal multiplicative interval such that there are 60 values from 0.01 m to 1 m. The equation for this is given as:

where is the decoded distance value and is the bit value as an integer value, i.e., . The result is in meters.

Table 5. Channel distance for Source formats: Default/Other and Microphone grid

|  |  |  |
| --- | --- | --- |
| Bit value | Decoded value | Additional description |
| 000000 | Unspecified | Distance is not specified, or it is unknown |
| 000001 | 0 m / coincident | No distance between microphones, i.e., they are coincident |
| 000010 | < 0.01 m | Distances smaller than 0.01 m |
| 000011 | 0.01 m | (Distances formed with equation above) |
| … | … | (Distances formed with equation above) |
| 111110 | 1 m | (Distances formed with equation above) |
| 111111 | > 1 m | Distances larger than 1 m |

**Source format == 01 (Microphone grid)**

If number of channels is 1 (bit value 0), no additional metadata is specified. Instead, 12-bit zero padding is applied.

If number of channels is 2 (bit value 1), following additional fields are configured in order:

* Transport definition field (3 bits). This field describes the configuration of the two transport channels. The possible bit values and corresponding configurations are provided in Table 3.
* Channel angle field (3 bits). This field describes symmetric angle positions for transport signals with directivity patterns. In this notation, 0° corresponds to the front. The bit values and corresponding configuration are defined in Table 4.
* Channel distance field (6 bits). The bit values and corresponding configuration are defined in Table 5.

The field definitions used for Microphone grid source format and Default/Other source format are the same. Differentiation is based on Source format parameter itself.

**Source format == 10 (Channel-based)**

For premixed content, the original layout can be provided. In addition to common CICP layouts relevant for IVAS, two generic options (3D and 2D) are available. The description of the bit values is provided in Table 6. The transport signals with this source format are assumed to be a mono (1 channels) or left-right stereo (2 channels) downmix of the multi-channel signals, and thus the number of channels can be 1 or 2 (bit values 0 or 1).

In addition to the 3-bit Channel layout field, 9 bits of zero padding is applied to complete the 12-bit variable description.

Table 6. Channel layout field for the channel-based source format

|  |  |  |
| --- | --- | --- |
| Bit value | Decoded value | Additional description |
| 000 | Unknown/Other | Unknown layout or other (3D) layout. Default option. |
| 001 | Other planar | Other 2D layout |
| 010 | 2.0 | CICP2 positions, ITU order |
| 011 | 5.1 | CICP6 positions, ITU order |
| 100 | 5.1+2 | CICP14 positions azimuth, 35° elevation, ITU order |
| 101 | 5.1+4 | CICP16 positions azimuth, 35° elevation, ITU order |
| 110 | 7.1 | CICP12 positions, ITU order |
| 111 | 7.1+4 | CICP19 positions azimuth, 35° elevation, ITU order |

Note 1: ITU channel order is given in ISO/IEC 23008-3:2015, Table 95.

Note 2: Azimuth positions are given in ISO/IEC 23091-3:2018, Table 3.

**Source format == 11 (Ambisonics)**

If number of channels is 1 (bit value 0), no additional metadata is specified. Instead, 12-bit zero padding is applied.

If number of channels is 2 (bit value 1), following two additional fields are configured in order:

* Transport definition field (3 bits). This describes the configuration of the two transport channels. The possible bit values and corresponding configurations are provided in Table 3. However, bit values 001 (omni) and 111 (binaural) are not allowed and should be interpreted as bit value 000.
* Channel angle field (3 bits). Describes symmetric angle positions for transports signals with directive patterns. 0° is assumed to point directly to the front. This is defined in Table 4.
* In addition, 6 bits of zero padding is applied to complete the 12-bit variable description.

For Ambisonics-based transport signals, transport channels are assumed to be coincident, and there is therefore no ‘Channel distance’ field specified.

# MASA spatial metadata parameters

The MASA spatial metadata describes the spatial audio characteristics corresponding to the one or two transport audio signals. Thus, the spatial audio scene can be rendered for listening based on the combination of the transport audio signals and the spatial metadata.

The MASA spatial metadata is provided once per subframe in each frame following the time-frequency resolution presented in clause A.2. Spatial metadata for each subframe contains one or two first sets of parameters depending on the number of directions (as defined by the corresponding metadata field in descriptive metadata, clause A.2) and one second set of parameters that does not depend on the number of directions. As shown in Figure A.2 and Figure A.3, the parameters corresponding to Table A.2a are written first in the stream, followed by the parameters corresponding to Table A.2b.

The definitions and use of the MASA spatial metadata parameters are described in order in the following.

The IVAS MASA C Reference Software provides implementation examples of the analysis and synthesis methods for these parameters using established methods.

## A.4.1 Direction index: Spatial direction(s)

Spatial directions represent the directional energy flows in the sound scene. Each spatial direction together with corresponding direct-to-total energy ratio describes how much of the total energy for each time-frequency tile is coming from that specific direction. In general, this parameter can also be thought of as the direction of arrival (DOA).

There can be one or two spatial directions for each time-frequency tile in the input metadata. Each spatial direction is represented using a 16-bit direction index. This is an efficient representation of directions as points of a spherical grid with an accuracy of about 1 degree in any arbitrary direction.

The direction indexing corresponds to the function for transforming the audio direction angular values (azimuth ϕ and elevation θ) into an index, and the inverse function for transforming the index into the audio direction angular values.

Each pair of values containing the elevation and the azimuth is first quantized on a spatial spherical grid of points and the index of the corresponding point is constructed. The structure of the spherical grid is defined first, followed by the quantization function and lastly the index formation followed by the corresponding de-indexing function.

The spherical grid is defined as a succession of horizontal circles of points. The circles are distributed on the sphere, and they correspond to several elevation values. The indexing functions make the connection between the angles (elevation and azimuth) corresponding to each of these points on the grid and a 16-bit index.

The spherical grid is on a sphere of unitary radius that is defined by the following elements:

* The elevation values are equidistant between -90 and +90 degrees; the value 0 is represented and corresponds to the circle situated on the equator. The values are symmetrical with respect to the origin. The number of positive elevation values is
* For each elevation value there are several equally spaced azimuth values. One point on the grid is given by the elevation and the azimuth value. The number *n*(*i*) of azimuth values is calculated as follows:
  + on the equator of the spherical grid () it is set to
  + there is one point at each of the poles ( degrees)
  + the function calculating the number of points on the grid for other elevation indices, uses the following definition:

with and

where is the uniform quantization step for , is a rounding function to the nearest even integer (above for , closest for ) The term gives the cumulative cardinality (i.e., cumulative number of points in the spherical grid) in a spherical zone going from the first non-zero elevation value to the -th elevation value. This cumulative cardinality is derived from the relative area on the spherical surface, assuming a (near) uniform point distribution of the remaining number of points (let alone the equator and poles).

* The azimuth values start from the front direction and are in trigonometrical order from 0 to .
* The quantized azimuth values for odd values of are equally spaced and start at 0.
* The quantized azimuth values for even values of are equally spaced and start at .
* There is a same number of quantized azimuth values for same absolute value elevation codewords.

The quantization in the spherical grid is done as follows:

* The elevation value is quantized in the uniform scalar quantizer to the two closest values
* The azimuth value is quantized in the azimuth scalar quantizers corresponding to the elevation values
* The distance on the sphere is calculated between the input elevation azimuth pair and each of the quantized pairs
* The pair with lower distance is chosen as the quantized direction.

The resulting quantized direction index is obtained by enumerating the points on the spherical grid by starting with the points for null elevation first, then the points corresponding to the smallest positive elevation codeword, the points corresponding to the first negative elevation codeword, followed by the points on the following positive elevation codeword and so on.

Further details of the direction indexing functions can be found in [4].

## A.4.2 Direct-to-total energy ratio(s)

Direct-to-total energy ratios work together with spatial directions as described above. Each direct-to-total energy ratio corresponds to a specific spatial direction and describes how much of the energy comes from that specific spatial direction compared to the total energy.

## A.4.3 Spread coherence

Spread coherence is a parameter that describes the directional energy flow further. It represents situations where coherent directional sound energy is coming from multiple directions at the same time. This is represented with a single spread coherence parameter that describes how the sound should be synthesized.

In synthesis, this parameter should be used such that value 0 means that the sound is synthesized to single direction as directed by the spatial direction, value 0.5 means that the sound is synthesized to the spatial direction and two surrounding directions as coherent, and 1 means that the sound is synthesized to two surrounding directions around the spatial direction.

## A.4.4 Diffuse-to-total energy ratio

Diffuse-to-total energy ratio represents non-directional energy flow in the sound scene. This is a complement to the direct-to-total energy ratios and in an ideal capture with no undesired signal (or synthesized sound scene), the diffuse-to-total ratio value is always

## A.4.5 Surround coherence

Surround coherence is a parameter that describes the non-directional energy flow. It represents how much of the non-directional energy should be presented as coherent reproduction instead of decorrelated reproduction.

## A.4.6 Remainder-to-total energy ratio

Remainder-to-total represents all the energy that does not “belong” to the captured sound scene based on the used model. This includes possible microphone noise and other capture artefacts that have not been removed from the signal in pre-processing. This means that by considering the direct-to-total energy ratio, the diffuse-to-total energy ratio, and the remainder-to-total energy we end up with a complete energy ratio model of

when there is any remainder energy present. Otherwise, the energy ratio equation in subclause A.4.4 can be followed.