**3GPP TSG-SA WG4 Meeting #119-e** ***S4-220698***

**Online, 11th – 20th May 2022**

**Source: Fraunhofer IIS**

**Title: Proposed Update of IVAS-4 Design Constraints: Ambisonics Order, max. Number of Loudspeaker Channels**

**Document for: Agreement**

**Agenda Item: 7.5**

## 1. Introduction

This contribution aims to complete the IVAS-4 Design Constraints [1] with respect to the following open points:

* Scene Based Audio: Supported Ambisonics Order
* Immersive Audio Formats: Number of Speakers for Multi-channel reproduction on arbitrary loudspeaker configurations

This contribution proposes to replace “[N]” and “[K]” by concrete numbers.

## 2. Discussion

### Ambisonics Order

Ambisonics is a scalable immersive audio format representing the captured sound scene/field in one spot. In short, the quality of the reproduced sound scene scales with the Ambisonics order. Dependent on the Ambisonics order is the number of signal components, i.e. Ambisonics coefficients. For a three-dimensional Ambisonics system of order *N*, the number of Ambisonics coefficients is given by *(N+1)2*:

|  |  |
| --- | --- |
| Order N | Number of coefficients |
| 0 | 1 |
| 1 | 4 |
| 2 | 9 |
| 3 | 16 |
| 4 | 25 |
| 5 | 36 |

In terms of quality, the Ambisonics order affects the following characteristics directly:

* Sweet spot: The area in the center of an Ambisonics system, in which a sound field can be recreated accurately, is called sweet spot, and is particularly important for loudspeaker reproduction. The physical sweet spot increases by Ambisonics order and decreases by frequency [3]. The physical sweet spot scales linearly with the Ambisonics order. However, based on practical experience, loudspeaker setup and Ambisonics rendering scheme (in the literature also referred to as “Ambisonics decoding”), the so called perceptual sweet spot might be significantly larger. However, also this sweet spot typically scales with Ambisonics order. Practical experiences at the source listening facilities have shown good experience with Ambisonics order 3 and above [6]. For headphone listening, typically a single point in the center of the sound field is considered, and thus the width of the sweet spot doesn’t play a role.
* Ideal minimum number of loudspeakers for playback: The required minimum number of loudspeakers to accurately reproduce a sound field within the sweet spot is again dependent on the Ambisonics order and corresponds to the number of Ambisonics coefficients [4]. For practical considerations when using loudspeaker playback, typically well-known layouts from industry standards are chosen, e.g. 7.1 + 4. As a rule of thumb, the number of speakers should at least approximately match the number of coefficients. However, facilities which can provide loudspeaker setups exceeded 7.1+4 loudspeaker setups are currently very rare, which in practice limits the potential quality gain from Ambisonics orders > 3.
* Localization: The localization of sound sources, especially objects, is also dependent on the Ambisonics order, as is shown in research [5]. Leaving the impact of different rendering algorithms aside, localization generally improves by increasing Ambisonics order, with First Order Ambisonics being the most problematic scenario. This applies to both, loudspeaker and headphone playback.
* Source Width: Differences in the perceived source width of sound sources can be observed [3]. These differences depend on the placement of source relative to the loudspeaker positions and get reduced with higher Ambisonics orders. This applies to both, loudspeaker and headphone playback.

In addition to these general quality characteristics of the Ambisonics format, also codec related considerations need to be made: The chosen Ambisonics order directly affects the number of Ambisonics coefficients (corresponding to the number of input/output signals to the codec), which has an impact on the following design constraints:

* Memory (RAM) demand: Increasing the number of Ambisonics coefficients increases the RAM demand of encoder, decoder and renderer. Directly affected are input and output buffers holding these coefficients. In addition to this, other extra RAM demand is dependent on the applied coding scheme: For discrete coding schemes the RAM demand will scale approx. linearly with the number of Ambisonics coefficients, for parametric coding scheme the increment might be in general somewhat lower. In addition, extra memory might be needed for handling of correlations or extra processing on the Ambisonics coefficients. Since high Ambisonics orders target high quality reproduction, a discrete coding scheme is expected to be used in combination with higher order Ambisonics at high bitrates.
* Computational complexity: Increasing the number of Ambisonics coefficients increases the computational complexity of encoder, decoder and renderer. Similar as for RAM, the increase typically depends on the implemented coding scheme: For discrete coding scheme the complexity is expected to scale with the number of Ambisonics coefficients, for parametric coding schemes it might be somewhat lower. In addition, extra processing might be needed for handling of correlations or extra signal processing on the Ambisonics coefficients. Since high Ambisonics orders target high quality reproduction, a discrete coding scheme is expected to be used in combination with higher order Ambisonics at high bitrates.
* Bitrate: The available bitrate is primarily dependent on channel capacity and application scenario. Increasing the Ambisonics order will also increase the number of Ambisonics coefficients and thus also the input signals, the available bitrate needs to be split. This leaves in general less bitrate per Ambisonics coefficient.

As can be seen from the discussion as outlined above:

* From a format-centric quality perspective, an increasing Ambisonics order seems beneficial
* From a codec perspective, an increasing Ambisonics order however is expected to impact negatively other design constraints such as RAM, computational complexity, total bitrate

In the context of IVAS, also other aspects need to be considered: The predominant listening device is expected to be headphones, where considerations such as sweet spot and number of loudspeaker do not really apply. Although for headphones the advantages from higher orders with respect to localization and source width would in general apply. However, it also needs to be understood that Ambisonics is not the only supported audio format: Issues such as localization and source width can be much better treated by dedicated audio formats, e.g. by using object-based audio or multi-channel audio.

Taking all these considerations and many years of practical experience into account, the source arrives at the following proposal: It is proposed to support Ambisonics up to 3rd order. This provides in general a good sound quality, without impacting the practical codec restrictions too much. In terms of input/output signals, the 16 Ambisonics coefficients for 3rd order Ambisonics are in the same order of magnitude as the already agreed 7.1+4 input format for multi-channel configurations. This would also correspond to the tested configurations during the VR Stream standardization [2].

### Number of Speakers for Multi-channel reproduction on arbitrary loudspeaker configurations

Along with the Ambisonics order, also the number of speakers for multi-channel reproduction on arbitrary loudspeaker configurations is currently still open. Purpose of the arbitrary loudspeaker configuration is to allow for optimal loudspeaker playback, even if the speakers installed in the room don’t match the standardized layouts.

With respect to the number of speakers, similar considerations as above should be made: The supported regular input/output configs are currently limited to twelve speakers (7.1+4). Extending the number of speakers over this limit has an impact on potentially both, memory (RAM) and computational complexity demand, with the benefit of additional flexibility.

We would propose to consider a joint optimization connected to the Ambisonics order: Supporting Ambisonics of order 3 would extend decoder/renderer buffers for Ambisonics to hold 16 signal components. In this light it is further proposed to align the number of speakers for arbitrary loudspeaker reproduction to this value, too, i.e. support at least 16 speakers for arbitrary loudspeaker configurations. Without need for additional memory, this would then allow for speaker re-arrangement for the supported configurations up to 7.1+4 and further optimization with respect to Ambisonics reproduction for up to 16 speakers, which then corresponds exactly to the number of Ambisonics coefficients for order 3.

## 3. Proposal

The following changes are proposed to be implemented in IVAS-4:

|  |  |
| --- | --- |
| **Encoder Input Formats** | The encoder shall support the following input formats:  Channel-based audio, including mono (1.0), stereo (2.0), surround (5.1 and 7.1), surround + height (5.1+4 and 7.1+4), TBD  Binaural audio  Scene-based audio (Ambisonics): FOA, HOA2 and HOA3.  Note: ACN component ordering and SN3D normalization.  Metadata-assisted spatial audio according to definition in Annex A.  [Spatial audio, [N] channels and spatial metadata defined by [TBD].]  Object-based audio, with support for at least [TBD] individual [mono] object streams. Each audio object shall be defined by [TBD metadata parameters].  [In addition, the IVAS codec shall support combinations of the above, totalling to no more than [TBD] audio streams.  ] |
| **Output Formats** | The IVAS codec shall support the following output formats for the corresponding input format:   |  |  | | --- | --- | | **Encoder Input Format** | **Output Format** | | Multi-channel 7.1+4 | Multi-channel 7.1+4, Binaural Audio, Stereo, Mono.  Multi-channel on arbitrary loudspeaker configurations of up to 16 speakers. | | Multi-channel 5.1+4 | Multi-channel 5.1+4, Binaural Audio, Stereo, Mono.  Multi-channel on arbitrary loudspeaker configurations of up to 16 speakers. | | Multi-channel 7.1 | Multi-channel 7.1, Binaural Audio, Stereo, Mono.  Multi-channel on arbitrary loudspeaker configurations of up to 16 speakers. | | Multi-channel 5.1 | Multi-channel 5.1, Binaural Audio, Stereo, Mono.  Multi-channel on arbitrary loudspeaker configurations of up to 16 speakers. | | Binaural Audio | Binaural Audio, [Stereo, Mono]  [Binaural Audio output assumes listening over headphones while Stereo output assumes listening over two channel Stereo loudspeaker configuration.  Editor’s note: Mono and Stereo output will not be tested in the selection phase] | | Stereo | Stereo, Mono | | Mono | Mono | | Scene-based audio | Scene-based audio of the same and lower orders than the input format, Binaural audio, Stereo, Mono  Multi-channel on arbitrary loudspeaker configurations of up to 16 speakers.  Editor’s note: at least one multi-channel configuration will be tested in the TBD phase | | Object-based audio | Object-based audio, Binaural audio, Stereo, Mono  Multi-channel on arbitrary loudspeaker configurations of up to 16 speakers.  Editor’s note: at least one multi-channel configuration will be tested in the TBD phase | | Metadata-assisted spatial audio | Metadata-assisted spatial audio, Binaural audio, Stereo, Mono  Multi-channel on arbitrary loudspeaker configurations of up to 16 speakers. | |  |  |   Editor’s note: Specification of output formats for the remaining input formats is needed.  Editor’s note: the term “arbitrary loudspeaker configuration” needs to be defined. One proposed definition is: rendered up to 16 loudspeaker positions on a 3D sphere. Potential further definition of minimum number of loudspeakers in an arbitrary configuration could be considered. More input is invited.  Editor’s Note: The exact codec configurations (bitrates etc.) for which particular output format is required is TBD, e.g., to be specified in IVAS-3 (Performance Requirements). |

## 4. Conclusion

The following updates to the IVAS design constraints are proposed:

* For Scene Based Audio: Support up to 3rd order Ambisonics
* Immersive Audio Formats: Support configurations with up to 16 speakers for reproduction on arbitrary loudspeaker configurations

It is proposed to agree on the outlined changes contained in Section 3 of this document and include them into IVAS-4.

## 5. References

[1] 3GPP S4-220551, IVAS Design Constraints (IVAS-4) v0.4.0   
[2] 3GPP TR 26.918, Virtual Reality (VR) media services over 3GPP  
[3] Matthias Frank, “How to make Ambisonics sound good”, Forum Acusticum, Krakow, 2014  
[4] Darren B. Ward, Thushara D. Abhayapala, “Reproduction of a Plane-Wave Sound Field Using  
an Array of Loudspeakers”, IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 9, NO. 6, SEPTEMBER 2001   
[5] Thirsa Huisman, Axel Ahrens, Ewen MacDonald, Ambisonics Sound Source Localization With Varying Amount of Visual Information in Virtual Reality, Frontiers in Virtual Reality, 2021   
[6] Franz Zotter, Matthias Frank, “Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality”, Springer Nature, 2019