**Source: China Mobile Com. Corporation**

**Title: 3D model loading with MRF assistance**

**Agenda Item: 10.6**

**Document for: Discussion and Agreement**

1. **Introduction**

As a 3D dynamic representation, 3D model is animated with expression or motion signals captured from a camera. In order to reduce session establishment time and computational procedure, the generation of 3D models is not necessary for every session. UEs can load previously generated models, and the generation steps only need to be performed once.

This document will discuss requirements on preparing non-real-time 3D model with MRF assistance for avatar based AR call.

1. **Proposal**

It is proposed to revise the following two changes in the IBACS permanent document.

|  |
| --- |
| **1st Change** |

### **2.1.1 AR two-party calls**

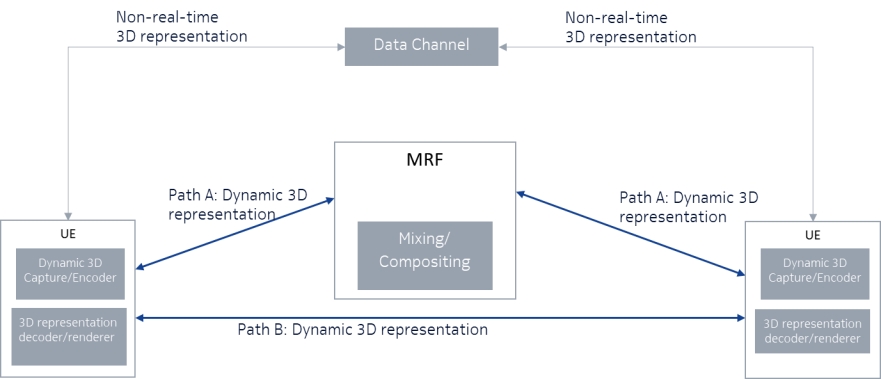


Figure 2.1.1.1: AR two-party call

The use case in Figure xx establishes a bidirectional AR two-party call, which may or may not use the MCU (MRF for IMS). It should be possible to combine other functions, e.g., 2D video, 360-degree video, images, etc., which are not shown in the figure for simplicity.

The dynamic 3D representations delivery in this case is over RTP. It is bidirectional with conversational latency requirements. The dynamic 3D representation can be delivered over one of the two paths shown in the figure. Path A goes through the MCU (MRF for IMS) and Path B is point-to-point.

Non-real-time 3D representations can be delivered via the data channel as shown in the figure.; the term 3D representation here includes both dynamic and static 3D representations, which are not captured and delivered under conversational latency requirements. The data channel is a WebRTC data channel or an IMS data channel. The IMS data channel used is as defined by TS 26.114. Further requirements, if needed, can be defined for using the data channel to transport 3D representations.

The following set of requirements relate to the bidirectional conversational dynamic 3D representation:

Call setup and control: this building block covers

signalling to setup a call or a conference – basic functions already provided by MTSI and will be covered also in IRTCW.

fetching of the entry point for the AR experience. The protocol needs to support upgrading and downgrading to/from an AR experience. Dependency on MeCAR to define device types.

Formats: The media and metadata types and formats include in addition to the ones already covered by MTSI, volumetric media. Format properties and codecs need to be defined for dynamic 3D representations along with appropriate RTP payload formats and functions. Appropriate codecs need to be defined by MeCAR for encoding, decoding and rendering dynamic 3D representations.

Real-time encoding and decoding with latency requirements for conversational media.

Enhancements to SDP, scene description to support AR telephony and content server (e.g.EMRF

or DCMF) to serves non-real-time 3D representation e.g. through a data channel.

For AR telephony media types(e.g., dynamic 3D representation), the necessary QoS characteristics need to be defined.

Support for AR media processing in the MCU (MRF for IMS).

5G system integration: offering the appropriate support by the 5G system to AR telephony includes:

discovery and setup of MCU (MRF) resources to process AR telephony media types.

defining the necessary QoS characteristics for AR telephony media types.

data collection and reporting.

### **2.1.2 AR multi-party calls**

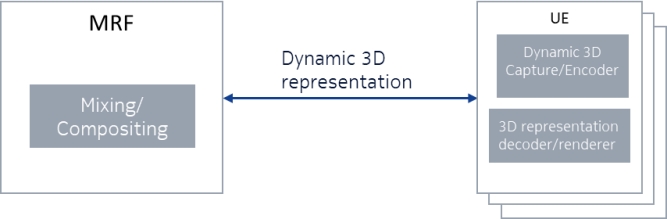


Figure 2.1.2.1: AR multi-party call

The use case in Figure 2.1.2.1 establishes a multiparty call. The call may be unidirectional (one dynamic 3D representation sender and multiple receivers, 1:N) or bidirectional (multiple dynamic 3D representation senders and multiple dynamic 3D representation receivers, N:N). The 1:N case can be addressed first as part of this work as it is simpler. The N:N case can be addressed later.

In addition to the requirements of use case 2.1, the following requirements need to be considered for AR multi-party calls:

Signalling for establishing a multiparty call. This may be done in a similar way as for traditional MTSI/WebRTC calls.

Expanding scene description to address the case of multiple senders and multiple receivers; and pre-loading non-real-time 3D representation e.g. through a data channel.

defining appropriate procedures to maintain position of all participants in the rendered space for each participant.

Mixing/transcoding in the MCU (MRF for IMS) to combine content from multiple participants. This may include e.g., scaling and placement of 3D representations in a virtual room. Other requirements can also be studied.

More advanced requirements may also be considered based on the existing use cases in FS\_5GXR and FS\_5GSTAR

Integration with other 5G services such as 5GMS for DASH delivery of AR media (that is not used for delivering conversational AR media but possibly video streams for a shared experience) along with conversational media.

Maintaining consistent head motion and eye-contact in a multiparty call with 3D avatars.

|  |
| --- |
| **Change End** |

|  |
| --- |
| **2nd Change** |

## 4.2 AR call with real-time Animated Avatar

## **4.2.1 MRF-assisted 3D modelling and skinning**

**4.2.2 MRF-assisted 3D model loading and animating**

3D model loading and animating needs to prepare non-real-time 3D model with MRF assistance for avatar based AR call. In order to reduce session establishment time and computational procedure, the generation of 3D models is not necessary for every session, the avatar data is generated offline. UEs can preload previously generated models, and the generation steps only need to be performed once.The avatar data (e.g. 3D mesh) can be stored on a UE or cloud server. the Enhanced MRF or Data Channel Media Function can assist in avatar animation.

NOTE1: The security and privacy issues of storage or management of 3D models on EMRF/DCMF or other network elements need to be considered in SA2 or SA3 architecture.

NOTE2: The step of delivering 3D model may include a 3D mesh, which is an animated model with fixed mesh point and texture, and it can be rendered with user view.

Figure 4.2.1.3 shows a call flow where an EMRF or DCMF preloads a 3D model and fetch facial expression or motion signals from images received from UE1 and deliver it to UE2. The call flow is shown as unidirectional for clarity but can work also as bi-directional. MRF assistance alleviates the need for UE1 to process the images itself.



Figure 4.2.1.3: Call flow for avatar based AR call with MRF assistance for model loading and animation

Steps 1-12 under A. Call Setup, B. Scene description retrieval and C. scene description update will be the same as for the basic AR call flow. The remaining steps are defined below.

1. The Enhanced MRF or DCMF will load a 3D model for UE1 for preparing the avatar call.
2. UE1 will send source images (e.g., RGB or RGB-D streams) to the Enhanced MRF as an image or video stream. The media description of the streams contains the camera configuration as well.
3. The Enhanced MRF processes the received images to create expression or motion signals during the session.
4. The Enhanced MRF animate the 3D model based on expression or motion signals .
5. The Enhanced MRF delivers the 3D model to UE2.
6. UE2 renders the 3D model with user view.

Figure 4.2.1.4 shows a call flow where an EMRF or DCMF prepares a 3D model and UE1 fetch facial expression or motion signals and deliver it to EMRF. The call flow is shown as unidirectional for clarity but can work also as bi-directional.



Figure 4.2.1.5: Call flow for avatar based AR call with MRF assistance for model animation

The related steps are defined below.

1. The Enhanced MRF or DCMF will load a 3D model to UE1 for preparing the avatar call and send this model to UE1.
2. UE1 processes the captured images to create expression or motion signals during the session.
3. UE1 delivers the expression or motion signals .
4. The Enhanced MRF uses the 3D model and the signals to animate 3D model.
5. The Enhanced MRF delivers the 3D model to UE2.
6. UE2 renders the 3D model with user view.

If avatar can be animated with expression or motion singals in the client, EMRF/DCMF will send the non-real-time 3D model to UE1. UE1 can directly create volumetric video streaming for the 3D model.



Figure 4.2.1.4: Call flow for avatar based AR call with MRF assistance for model loading

1. EMRF/DCMF will send the non-real-time 3D model to UE1. The model is rigged already to a human skeleton.
2. UE1 processes the images to create expression or motion signals during the session.
3. UE1 renders the 3D model with animation based on the signals.
4. UE1 delivers the 3D model to UE2 via Enhanced MRF.
5. UE2 renders the 3D model with user view.

**Requirements:**

Based on the above call flows, an MRF-assisted 3D avatar conversational AR call has the following requirements:

1. An SDP indication for generation or pre-loading of a 3D model by the MRF, including an indication that the MRF can further create motion signals. The SDP negotiation would include the media properties of the image stream provided by the source (sender UE).
2. A media format for the source images. Any existing encoded video/image streams can be used for this purpose. Furthermore, a volumetric video stream may be used as well.
3. Transport and format for the expression or motion signals.

|  |
| --- |
| **Change End** |