|  |  |
| --- | --- |
|  | Moving Picture, Audio and Data Coding  by Artificial Intelligence  www.mpai.community |

|  |  |
| --- | --- |
| **M762** | 2021/01/10 |
| **Source** | Leonardo Chiariglione |
| **Title** | The MPAI Metaverse Project |
| **Target** | MPAI Members |

Disclaimer: This document is work in progress. A previous version of this document has been discussed at MMC-DC and CAE-DC meetings.

# Introduction

This document proposes the MPAI Metaverse Project, a tool to 1) develop a unified view of an important portion of MPAI activities, to 2) promote synergies among them, and 3) identify new areas of work.

Components of standards and projects should be allocated to Areas of endeavour when they are considered useful to develop Metaverses. A new proposed idea should be checked for consistency with the MMP and may be found to fit in, need revision or cause a revision of the MMP.

This document:

1. Proposes the MPAI Metaverse Model and related terms and definitions.
2. Identifies and populates Project Areas with components of MPAI activities – standards or projects – to Project Areas.

# The MPAI Metaverse Model

*Metaverse* is a processing environment containing data structures that are acquired and processed, generated, animated, and possibly distributed.

## Data acquisition and processing

The portion of the physical world relevant to a Metaverse is called a *Real Space*. It contains *Real Objects* having *Real Features*. Important Real Objects are *Real Audio Objects*, in particular *Real Speech Objects*, and *Real Visual Objects*, in particular *Real Human Objects*.

Real Objects are sensed, converted to either *Virtual Audio Objects* and *Virtual Visual Objects* (collectively called *Virtual Objects*), possibly processed, and made available to a Metaverse, possibly via the network.

## Animation

The Metaverse may:

1. Further process the Virtual Objects received to make them usable for purposes dictated by the internal logic of the Metaverse or by interactions sensed from the Real Space.
2. Preserve the existing dynamics of the Virtual Objects or animate them according to the Metaverse internal logic or by interactions sensed from the Real Space.
3. Create and animate synthetic Virtual Objects.
4. Blend and animate received and internally generated Virtual Objects.

The animated set of Virtual Objects is called a *Virtual Space.* Artificial Intelligence may play a significant role in the animation of Virtual Spaces.

Real Spaces may have different access rights to a common Virtual Space and may have different presentations of the common Virtual Space.

A Metaverse may run different independent or connected Virtual Spaces.

## Distribution

The Metaverse can make available a defined subset of the Virtual Space to a Real Space in a form that allows Actuators, possibly connected with the Metaverse via a network, to process and present the received Virtual Space and/or its Virtual Objects.

Three cases are possible:

1. A Real Space may be connected to
   1. One Metaverse as in the case of a user playing a video game.
   2. More than one Metaverses as in the case of a user playing a video game while listening to a live concert.

Some of the Metaverses can be connected or independent.

1. A Metaverse may be connected to
   1. One Real Space as in the case of a user playing a video game.
   2. More than one Real Space as in the case of a videoconference. Some of the Real Spaces may be the same as in the case of a videoconference where some participants are in the same room.
2. A plurality of Real Spaces can be connected to a plurality of Metaverses as in the case of a swarm of Connected Autonomous Vehicles (CAV).

# Definitions of Metaverse term

The MPAI Metaverse Model (MMM) is depicted in Figure 1. For simplicity, only the case of multiple Real Space – one Metaverse – One Virtual Space is represented Note that VAO = Virtual Audio Object and VVO = Virtual Video Object.

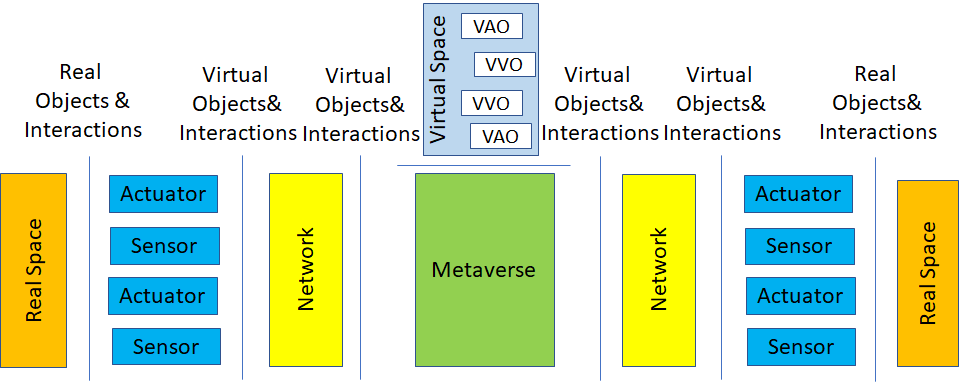


Figure 1 – The MPAI Metaverse Model

The MPAI Metaverse Terminology is given by *Table 1*.

*Table 1 –* *MPAI Metaverse Terminology*

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Actuator | A device capable to act on the Physical World, e.g., by generating audio and video, providing tactile experiences, activating machines, etc. |
| Audio Features | Elements embedded in an Audio Object whose extraction enables its identification as an object of a specific type. |
| Audio Object | A Real Audio Object or a Virtual Audio Object. |
| Audio Space | The Audio Objects in a Space. |
| Audio-Visual Object | A combination of Audio Objects and Visual Objects. |
| Human Visual Features | Characteristics describing the visual aspects of a dynamic human, e.g., face, head, arms, and hands. |
| Interaction | The ability of Real Objects and Virtual Objects to influence each other by exchanging data. |
| Metaverse | A computing environment equipped with sensors and actuators possibly connected to them via the network and able to:   1. Receive and process Virtual Objects from the Real Space. 2. Generate Virtual Objects that are entirely synthetic or contain elements derived from Virtual Objects received from the Real Space. 3. Animate Objects based on internal logic or interaction with the Real Space. |
| Real Audio Object | A sound field sensed or generated by the Virtual Space in the audible range. |
| Real Human | A Real Visual Object corresponding to a human. |
| Real Object | A Real Audio Object or a Real Visual Object or an Audio-Visual Real Object |
| Real Space | The set of Real Objects before conversion to Virtual Objects. |
| Real Visual Object | An electromagnetic field sensed or generated by the Virtual Space in the visible range. |
| Sensor | A device capable to convert information generated from the Physical World into digital data, e.g., converting audio and visual information into a stream of digital data, converting pressure on a mouse button, etc. |
| Space | A Real or Virtual Space. |
| Speech Features | The information embedded in a Speech Object, such as: text, emotion, colour. |
| Speech Object | An Audio Object generated by a speaking human or containing one or more Audio Objects generated by speaking human(s). |
| Virtual Audio Object | A digital representation of a Real Audio Object or a synthetically generated data structure that can be perceived at presentation time as capable to generate or reflect sound or a combination of both. |
| Virtual Human Objects | Also called Avatars, are Virtual Object that can be perceived at presentation time as either having a human-like appearance or having the ability to generate sound or both. |
| Virtual Object | A Virtual Audio Object or a Virtual Visual Object or a Virtual Audio-Visual Object. |
| Virtual Space | The set of Virtual Objects and Interactions handled by the Metaverse |
| Virtual Visual Object | A digital representation of a Real Visual Object or a synthetically generated data structure that can be perceived at presentation time as being visible and possibly being felt as solid or a combination of both. |
| Visual Features | Elements embedded in a Visual Object enabling it to be identified as an object of a specific type. |
| Visual Object | A Real Visual Object or its corresponding Virtual Visual Object. |
| Visual Space | The Visual Objects in a Space. |

# MPAI standards and projects relevant to the Metaverse

The following MPAI use cases illustrates the intended use of the MPAI Metaverse Project:

## CAE-EES – Emotion Enhanced Speech

A Speech Object has no particular emotion embedded. It can be changed by embedding in it an emotion selected from a list of standard emotions or by imitating the emotion embedded in a model Speech Object.

## CAE-AOG – Audio On the Go

An Audio Object is separated into its component Audio Objects, which are classified and made independently available.

## MMC-UST – Unidirectional Speech Translations

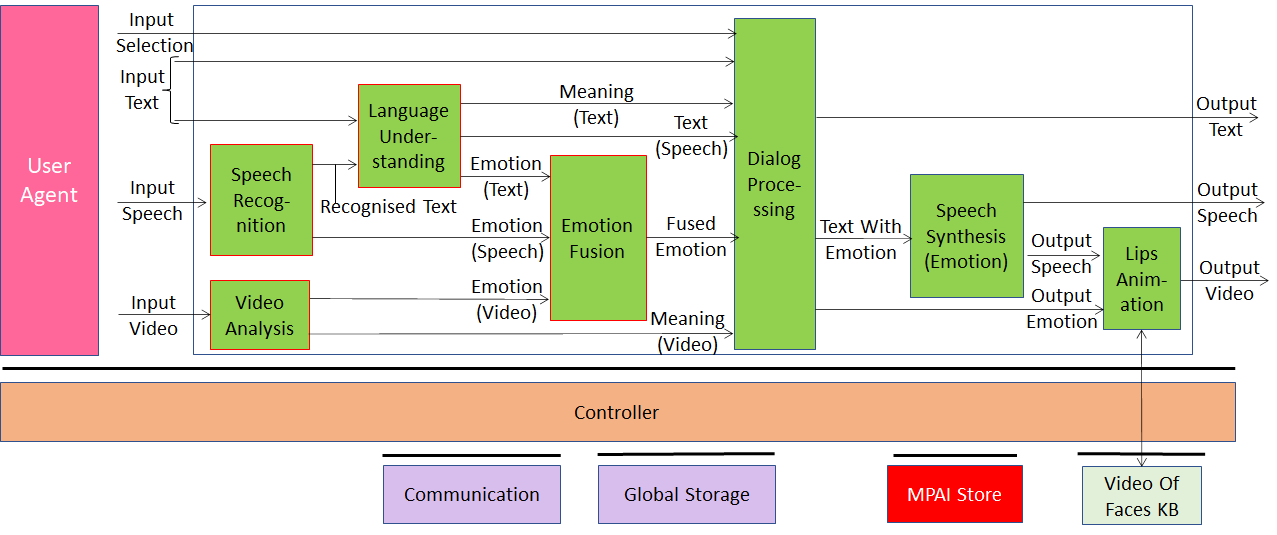
A Speech Object composed of speech in a given language with any emotion or colour is automatically translated into another language preserving the original emotion or colour. Bidirectional and One-to-Many Speech Translation are variations of the same topic.

## CAE-EAE – Enhanced Audioconference Experience

The Audio Object of an audio/videoconference room captured by a microphone array is processed to separate its Speech Objects from the other Audio Objects, separate the individual Speech Objects and create a description of the Audio Space. The processing yields an Audio Object containing the Audio Space description and the separated Speech Objects for transmission to a Metaverse (conference server).

## MMC-CWE – Conversation with Emotion

A Text, a Speech Object and a Visual Object of the face of the speaker (text and speech are alternative) are processed to extract text, emotion and meaning embedded in them. The Metaverse processes the 3 inputs and produces text and text with emotion which are appropriate responses to the 3 inputs. Further, the Metaverse converts text with emotion to a Speech Object, and uses the latter Speech Object and the output emotion to animate the lips of an Avatar.

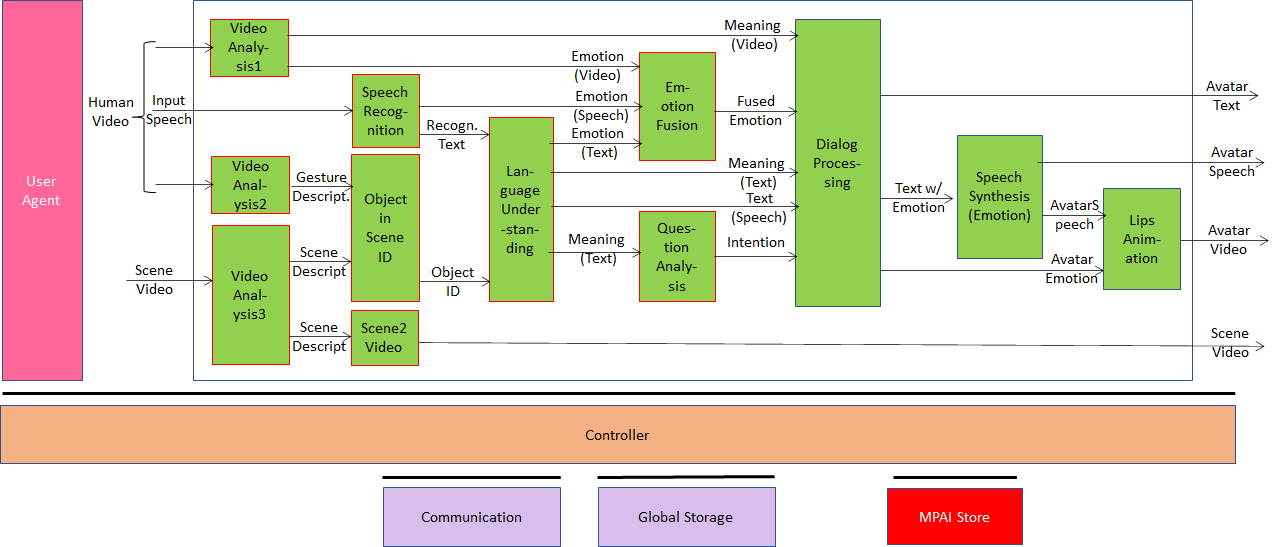


## MMC-MQA – Multimodal Question Answering

A Text, a Speech Object and a Visual Object of the speaker holding an object (text and speech are alternative) are processed to extract text, meaning and intention embedded in them. The Metaverse processes the 3 inputs and produces a Text or a Speech Object which are appropriate responses to the 3 inputs.

## MMC-CAS – Conversation about a scene

A Real Visual Object (a human in a scene with objects) and a Speech Object (the human’s speech) are processed to separate and identify the static Visual Objects and the dynamic Human Object, and to create a Visual Space description. The Metaverse locates and identifies the object the Human Object points to by following the movement of arms, hands and fingers; extracts the text, meaning and intention embedded in the Speech Features; and produces text and a Speech Object which are appropriate responses.



## CAV-HCI – Human-CAV Interaction

### Authenticate a user of a CAV

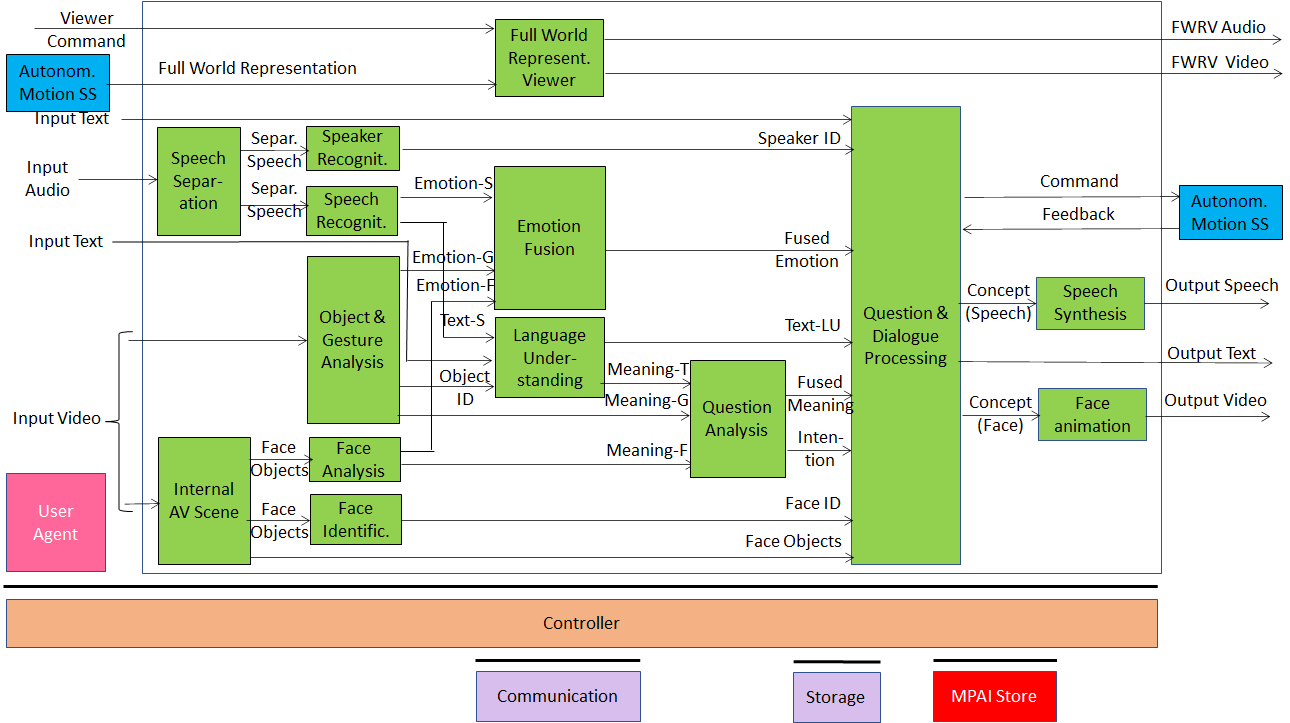
1. Separate the Speech Object generated by a Human Object captured by the CAV’s microphone array from the rest of the Audio Object (sound surrounding the CAV).
2. Extract the Speech Features of the separated Speech Object.
3. Identify the Human Object using the Speech Features.
4. Separate the Human Object from the rest of the Visual Object captured by the CAV’s camera array.
5. Extract the Human Features of the separated Human Object.
6. Identify the Human Object using the Human Features.

### Human-CAV dialogue

1. Capture the Visual Object containing a group of Human Objects in a Real Space (CAV cabin).
2. Separate the Human Objects from the other Visual Objects.
3. Create a description of the Human Objects in the Real Space.
4. Extract the Human Features and the Speech Features of the Human Objects.
5. Use all features to derive text, emotion, meaning and intention.
6. Produce appropriate CAV text and emotion using text, emotion, meaning and intention of 5.
7. Produce and animate a Human Object of which
   1. The Speech Object embeds the text and emotion of 6.
   2. The Visual Human Object embeds text and emotion in lips, head, and arms/hands/ fingers and whose eyes gaze at the Human Objects in a way consistent with the output text and emotion.

### Humans consume Full World Representation

Cabin passengers may wish to view and hear the outside environment of the CAV by getting a presentation of the Virtual World (called Full World Representation) created by the Autonomous Motion Subsystem Metaverse.

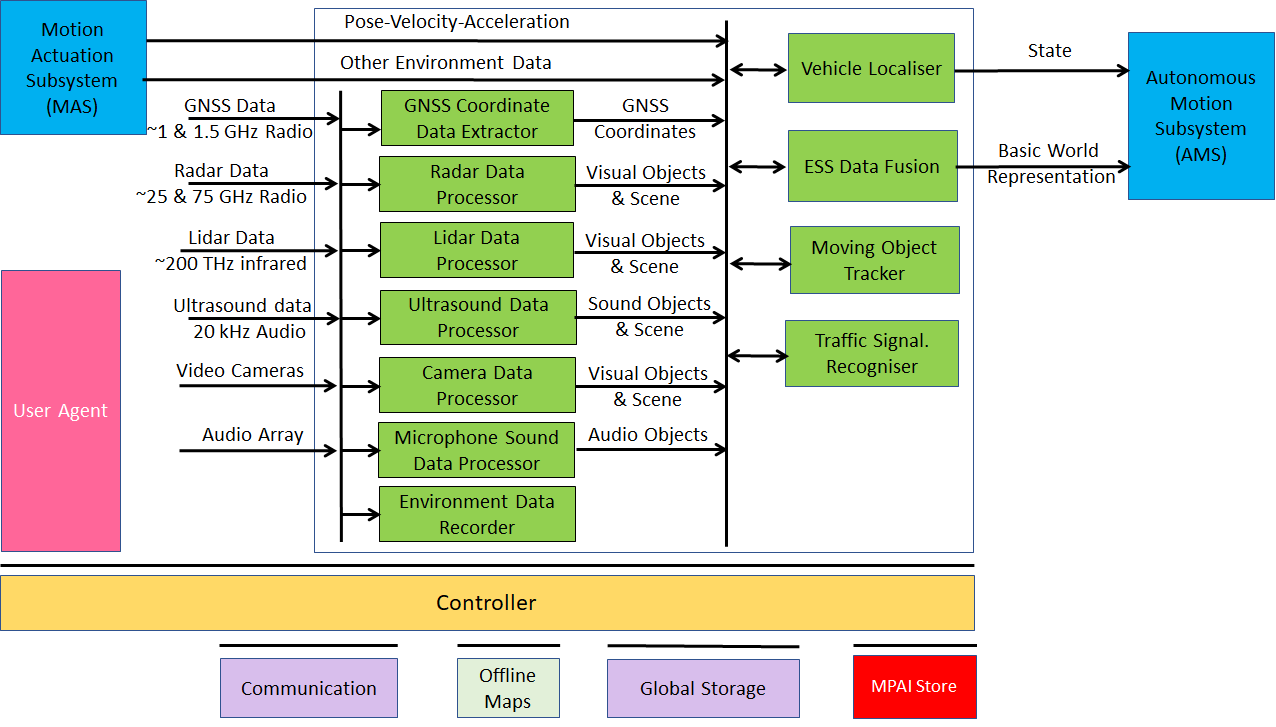


## CAV-ESS – Environment Sensing Subsystem

CAV-ESS performs the following functions

1. Captures different instances of the Real Space (called Environment in MPAI-CAV) using a variety of “visual” technologies, such as camera, lidar, radar, and ultrasound, and also other sensors, such as for position, velocity, and acceleration.
2. Separates and identifies the Visual Objects of each captured instance.
3. Creates a Visual Space description for each captured instance.
4. Merges the Visual Space descriptions.
5. Adds data from other sensors, e.g., temperature, humidity.
6. Creates a complete Visual Space description, called Basic World Representation (BWR).

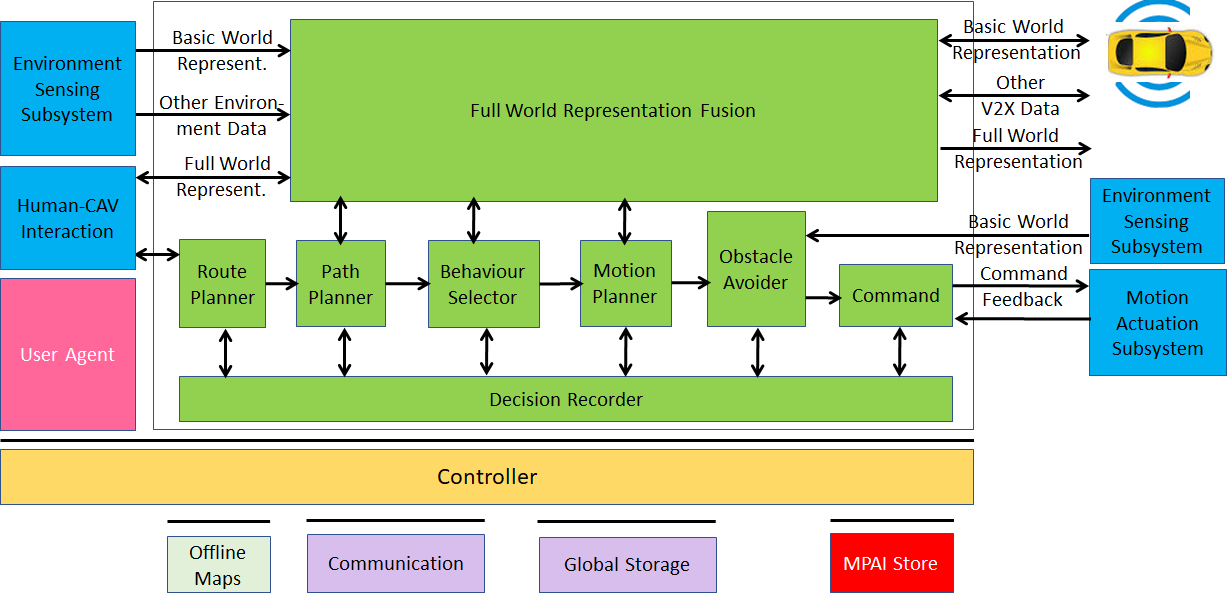
The BWR is shared with other CAV subsystems and CAVs in range, but is not consumed by a passenger.



## CAV-AMS – Autonomous Motion Subsystem

The AMS of a CAV is a Metaverse sharing its Virtual Space with the ESS in the same CAV and the ESSs of all CAVs in range. It integrates the different views of the Real Space to create the best estimate of the Real World in its Virtual Space, called Full World Representation (FWR).

The FWR may be shared with other CAVs in range and may be consumed by a passenger.



## CAV-MAS – Motion Activation Subsystem

As a result of the Metaverse processing performed by the Autonomous Motion Subsystem (AMS), the Motion Activation Subsystem (MAS) receives commands to actuate a movement to an assigned Pose in e.g., 5 seconds with given velocity, acceleration and angle. MAS provides feedback to AMS thus allowing it to update its Virtual Space.

## MPAI-MCS – Mixed-reality Collaborative Spaces

Remote participants are represented in a Mixed-reality Collaborative Space (MCS) by Virtual Human Objects animated in sync with the movements of remote participants’ torsos: the same speech of the participant, and face/head/arms/hands/fingers.

Participants may add Virtual Audio-Visual Objects to the Virtual Space (playing a role equivalent to today’s slide presentation).

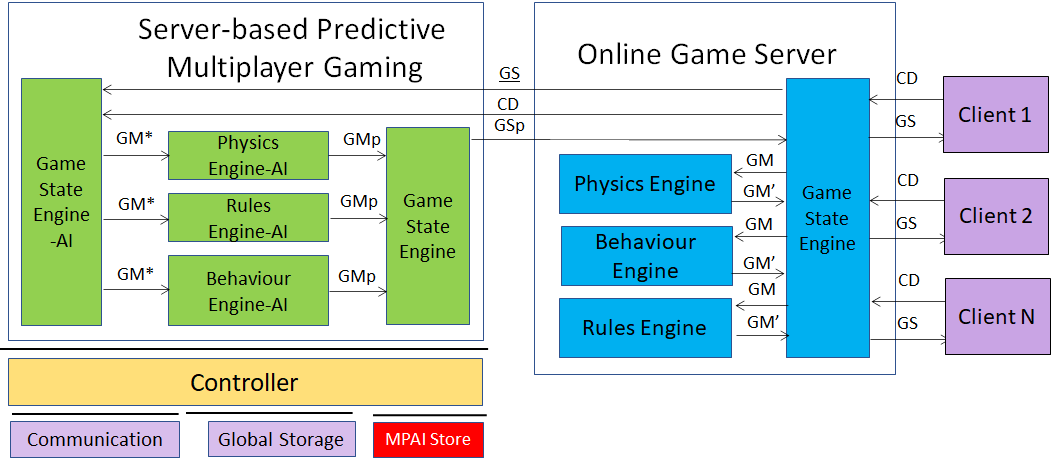
Participants may create their *own* version of the Virtual Space by moving objects (e.g., by placing a Virtual Human Object at a particular position).

## Server-based Predictive Multiplayer Gaming

Each online game Client is a Metaverse that creates a Virtual Space (the Game State) and is connected to an Actuator (the video display) that converts the Virtual Space into a Real Audio and Visual Object for human player’s use. The Client’s Game State is constantly updated by the Game Server.

The Game Server uses the controller data received from the Client to update its Game State, the Virtual Space of the Game Server. The new Game State is notified to all Clients in the game. The Virtual Space of the Game Server never gets converted into any Audio and Visual Objects.

If there are missing controller data, the Game Server predicts the Virtual Space (Game State)with a neural network.



# The Metaverse Project

The three Areas used to map technologies into the Metaverse project are:

1. *From Real to Virtual Space*: digitisation of Real Objects and processing performed before a Virtual Object is made part of the Virtual Space.
2. *In the Virtual Space*: creation and animation of s Virtual Audio Objects and Virtual Video Objects that are synthetically created or derive from Real Objects.
3. *From Virtual to Real Space*: conversion of a Virtual Space to a Real Space, including presentation and actuation.

## From Real to Virtual Space

This section addresses the data generated by sensors and processed to make the data usable in the Virtual World.

### Audio Object processing

Audio Object processing includes:

1. Separation of the individual Audio Objects contained in an Audio Object to enable the Metaverse to handle them individually.
2. Extraction of the Audio Features.
3. Identification of the spatial coordinates of the individual Audio Objects.

In MPAI:

|  |  |
| --- | --- |
| CAE-EAE | 1. Remove non-Speech Audio Objects from the sensed Real Audio Object. 2. Separate the remaining Speech Object its component Speech Objects. 3. Provide a description of the Audio Space. |
| CAE-AOG | Separate, identify and make independently available the individual Audio Objects of an Audio Space. |

### Speech Object processing

Speech Object processing seeks to achieve the goals of a human brain extracting information from speech and more results than a human brain can possibly achieve.

In MPAI:

|  |  |
| --- | --- |
| CAE-EES | Add an emotion to a Speech Object that has no embedded emotion by either:   1. Providing the standard identifier of the emotion. 2. Giving a model Speech Object and asking the Metaverse to add an emotion to the Speech Object that imitates the emotion embedded in the model Speech Object. |
| MMC-CWE | Extract text, emotion and meaning from a Speech Object. |
| MMC-MQA | Extract text, meaning and intention from a Speech Object. |
| CAV-HCI | Authenticate a user of a CAV talking to the CAV by   1. Separating the Speech Object of the Human Object captured by the CAV’s microphone array from the rest of the surrounding sound 2. Extracting and processing the Speech Features of the Speech Object to determine the identity of the Human Object. |

### Visual Object processing

Visual Object processing includes

1. Separation of the individual Visual Objects contained in a Visual Object to enable the Metaverse to handle them individually.
2. Extraction of Visual Features.
3. Identification of the possibly varying spatial coordinates of the individual Visual Objects.

In MPAI:

|  |  |
| --- | --- |
| MPAI-MQA | Separate and give an identifier to a Visual Object held by a human. |
| MPAI-CAS | (Conversation about a scene)   1. Separate and identify the Visual Objects including the Human Object of a Visual Space. 2. Create a Visual Space description. 3. Follow the movement of the arms, hands and fingers of the Human Object. 4. Recognise the Visual Object the Human Object is pointing at. |
| CAV-HCI | (Human-CAV Interaction)  Authenticate a user of a CAV talking to the CAV by   1. Separating the Human Object captured by the CAV’s cameras from the rest of the Visual Space. 2. Extracting and processing the Human Features of the Human Object to determine its the identity. |
| CAV-ESS | (Environment Sensing Subsystem)   1. Separate and identify the Visual Objects captured in the sensed area of the CAV environment using a variety of “visual” technologies, such as camera, lidar, radar, ultrasound, and other sensors, such as position, velocity, acceleration, temperature, humidity 2. Create a Visual Space description called Basic World Representation. |
| CAV-AMS | (Autonomous Motion Subsystem)   1. Combine the CAV’s own Basic World Representation with those coming from other CAVs in range, considered as Metaverses interacting with the same Real Space. 2. Create the Full World Representation, the best estimate of the CAV’s Real Space. |

### Human Object processing

Human Object processing includes:

1. Identification of the possibly time-dependent Human Features in a Human Object.
2. Extraction of the possibly culture dependent semantics of Human Features, e.g., language, winking.

In MPAI

|  |  |
| --- | --- |
| CAV-HCI | Separate and locate Human Objects in the Visual Space (i.e., the CAV’s cabin). |
| MPAI-CAS | (Conversation about a scene)   1. Separate and identify the Visual Objects of the Visual Space. 2. Recognise arm-hand-finger of a Visual Human Object pointing to a Visual Object. 3. Identify the Visual Object that the finger is pointing at. |

### Interaction

Interaction of a Real Space with a Virtual Space may use a variety of devices.

|  |  |
| --- | --- |
| MPAI-SPG | (Server-based Predictive multiplayer Gaming)  Each online game Client is a Metaverse sharing a Virtual Space (the Game State) with the Game Server and connected to an actuator (the video display)   1. The Client    1. Presents the Virtual Space (Game State) to the video display    2. Receives data from the controller    3. Sends the data to the game server 2. The Game Server    1. Receives controller data from the clients.    2. Updates the Virtual Space (Game State)    3. Sends the Game State to the clients    4. If there are missing controller data, it predicts the Virtual Space (Game State)with a neural network. |
|  |  |

## In the Virtual Space

### Introduction

This section deals with creation and animation of Virtual Audio Objects and Virtual Video Objects that may:

1. Derive from Real Audio Objects and Real Visual Objects.
2. Be synthetically created.

Objects have formats suitable to be handled by the Metaverse.

### Object conversion

The Metaverse may need a suitably processed Real Object or may need to convert it to a different Virtual Object.

In MPAI:

|  |  |
| --- | --- |
| MMC-UST | A Speech Object is processed to extract text and other Speech Features.  A Speech Object is synthesised that is a translation of the text into another language, pronounced with the Speech Features of the original speech and the Speech Features of the target language. |
| MMC-BST | Ditto |
| MMC-MST | Ditto |

### Object synthesis

A new Object may be synthesised in response to Virtual Objects of external or internal provenance.

In MPAI:

|  |  |
| --- | --- |
| MMC-MQA | A Speech Object is synthesised from plain text. |
| MMC-CWE | A Speech Object is synthesised from text and a separate emotion. |
| MMC-CWE | An Avatar made of a face whose lips are animated in accordance with an input Speech Object and a separate emotion equivalent to emotion embedded in the Speech Object. |
| MMC-UST | A Speech Object in a language is synthesised from text and Speech Features. |
| MPAI-MCS | An Avatar is synthesised with head, face, arms, hands and fingers copying the movements of a Human Object. |
| CAV-HCI | 1. An Avatar made of a head is synthesised from text, emotion, meaning and intention generated by the “Reaction to stimuli” component of the Metaverse. 2. The lips move expressing the emotion generated by the “Reaction to stimuli” component of the Metaverse. 3. The eyes gaze at a specific person when the Speech Object is intended to a specific Human Object or gaze at different Human Objects when no specific Human Object is intended. |

### Reaction to stimuli

The Metaverse may give an object the ability to generate reactions to stimuli coming from Objects in the Virtual Space. A typical reaction is the generation of Speech Objects or Avatars animation of whose lips, face, head, arms, hands, fingers move in sync with the intended reaction.

In MPAI:

|  |  |
| --- | --- |
| MMC-CWE | Text, emotion, and meaning are processed to generate text and Speech Objects with an embedded emotion, and an Avatar whose lip movements are congruent with the input text, emotion and meaning. |
| MMC-MQA | Text, meaning, and intention are processed to generate text and Speech Objects that are congruent with the input text, emotion and meaning. |
| CAV-HCI | Text, emotion, meaning and intention are processed to generate Speech Objects with an embedded emotion, and an Avatar whose head, face, eye, lip movements are congruent with the input text, emotion, meaning and intention. |

## From the Virtual to the Real Space

### Introduction

This section concerns the conversion of a Virtual Space to a Real Space. This includes presentation and actuation.

A Virtual Space may be provided for consumption in final form, or a level of actions may be permitted, e.g., a Human’s Avatar may be allowed to move in the Virtual Space or the Human can simply consume the Virtual Space looking at/listening to it “from the outside”.

|  |  |
| --- | --- |
| CAV-HCI | A CAV passenger may wish to obtain a presentation of the Virtual Space (called Full World Representation in MPAI-CAV) to be aware of the audio and visual surroundings of the CAV. |
| CAV-MAS | (Autonomous Motion Subsystem)  As a result of the Metaverse (CAV-AMS) processing, the Autonomous Motion Subsystem (MAS) receives a command to actuate a movement from “Pose A now” to “Pose B” in e.g., 5 seconds with given velocity, acceleration and angle. MAS provides feedback to AMS thus allowing it to update its Virtual Space. |

# Conclusion

This is an initial effort at allocating functionalities of MPAI standards and projects to the identified Metaverse Areas.