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

**Public document**

|  |  |
| --- | --- |
| **N780** | 2022/07/19 |
| **Source** | MMC-DC |
| **Title** | MPAI-MMC V2 Use Cases and Functional Requirements |
| **Target** | MPAI Members |

Contents

[1 Introduction 3](#_Toc109166602)

[2 Scope of the Use Cases 4](#_Toc109166603)

[2.1 Use Case description 4](#_Toc109166604)

[2.2 Key technologies 6](#_Toc109166605)

[2.2.1 Audio and Visual Scene Description 6](#_Toc109166606)

[2.2.2 User identification 6](#_Toc109166607)

[2.2.3 Personal Status Extraction 6](#_Toc109166608)

[2.2.4 Personal Status Display 7](#_Toc109166609)

[2.2.5 Collaborative Virtual Environments 7](#_Toc109166610)

[2.3 Summary of functionalities 7](#_Toc109166611)

[3 Terms and Definitions 8](#_Toc109166612)

[4 References 10](#_Toc109166613)

[4.1 Normative References 10](#_Toc109166614)

[4.2 Informative References 11](#_Toc109166615)

[5 Composite AIM Architectures 12](#_Toc109166616)

[5.1 Personal Status Extraction 12](#_Toc109166617)

[5.1.1 Scope of Composite AIM 12](#_Toc109166618)

[5.1.2 Reference architecture 12](#_Toc109166619)

[5.1.3 Input/output data 12](#_Toc109166620)

[5.1.4 AI Modules 13](#_Toc109166621)

[5.2 Personal Status Display 13](#_Toc109166622)

[5.2.1 Scope of Composite AIM 13](#_Toc109166623)

[5.2.2 Reference Architecture 13](#_Toc109166624)

[5.2.3 Input/output data 14](#_Toc109166625)

[5.2.4 AI Modules 14](#_Toc109166626)

[6 Use Case Architectures 14](#_Toc109166627)

[6.1 Conversation About a Scene 14](#_Toc109166628)

[6.1.1 Scope of Use Case 14](#_Toc109166629)

[6.1.2 Reference architecture 15](#_Toc109166630)

[6.1.3 Input/output data 16](#_Toc109166631)

[6.1.4 AI Modules 16](#_Toc109166632)

[6.2 Human-CAV Interaction AIW (HCI) 17](#_Toc109166633)

[6.2.1 Scope of Use Case 17](#_Toc109166634)

[6.2.2 Reference architecture 18](#_Toc109166635)

[6.2.3 Input and output data 19](#_Toc109166636)

[6.2.4 AI Modules 20](#_Toc109166637)

[6.3 Avatar-Based Videoconference 21](#_Toc109166638)

[6.3.1 Scope of Use Case 21](#_Toc109166639)

[6.3.2 Transmitting Client AIW 22](#_Toc109166640)

[6.3.3 Server AIW 25](#_Toc109166641)

[6.3.4 Virtual Secretary AIW 27](#_Toc109166642)

[6.3.5 Receiving Client AIW 29](#_Toc109166643)

[7 I/O Data summary 31](#_Toc109166644)

[7.1 Personal Status Extraction 31](#_Toc109166645)

[7.2 Personal Status Display 31](#_Toc109166646)

[7.3 Conversation About a Scene 31](#_Toc109166647)

[7.4 Human-CAV Interaction 32](#_Toc109166648)

[7.5 Avatar Videoconference 33](#_Toc109166649)

[7.5.1 Transmitting Client 33](#_Toc109166650)

[7.5.2 Server 33](#_Toc109166651)

[7.5.3 Virtual Secretary 33](#_Toc109166652)

[7.5.4 Receiving Client 34](#_Toc109166653)

[8 Functional Requirements 34](#_Toc109166654)

[8.1 Digital representation of analogue signals 35](#_Toc109166655)

[8.1.1 Microphone Array Audio 35](#_Toc109166656)

[8.1.2 2D Video 35](#_Toc109166657)

[8.1.3 3D Video 35](#_Toc109166658)

[8.2 Natively digital data formats 36](#_Toc109166659)

[8.2.1 Text 36](#_Toc109166660)

[8.2.2 Recognised Text 36](#_Toc109166661)

[8.2.3 Language Understanding (Text) 36](#_Toc109166662)

[8.2.4 Summary 36](#_Toc109166663)

[8.2.5 Environment Model 36](#_Toc109166664)

[8.2.6 Avatar Model 37](#_Toc109166665)

[8.2.7 Human Object 38](#_Toc109166666)

[8.2.8 Face Object 38](#_Toc109166667)

[8.2.9 Head Object 38](#_Toc109166668)

[8.2.10 Arm Object 38](#_Toc109166669)

[8.2.11 Hand Object 38](#_Toc109166670)

[8.2.12 Finger Object 39](#_Toc109166671)

[8.2.13 AV Scene Renderer Format 39](#_Toc109166672)

[8.2.14 Spatial Attitude 39](#_Toc109166673)

[8.2.15 Point of View 39](#_Toc109166674)

[8.3 Descriptors 39](#_Toc109166675)

[8.3.1 Audio Scene Descriptors 39](#_Toc109166676)

[8.3.2 Visual Scene Descriptors 40](#_Toc109166677)

[8.3.3 Audio-Visual Scene Descriptors 41](#_Toc109166678)

[8.3.4 Avatar Descriptors 41](#_Toc109166679)

[8.3.5 Face Descriptors 42](#_Toc109166680)

[8.3.6 Gesture Descriptors 42](#_Toc109166681)

[8.3.7 Head Descriptors 42](#_Toc109166682)

[8.3.8 Arm Descriptors 43](#_Toc109166683)

[8.3.9 Hand Descriptors 43](#_Toc109166684)

[8.3.10 Finger Descriptors 43](#_Toc109166685)

[8.3.11 Speech Descriptors 43](#_Toc109166686)

[8.3.12 Text Descriptors 44](#_Toc109166687)

[8.4 Interpretations 44](#_Toc109166688)

[8.4.1 Emotion 44](#_Toc109166689)

[8.4.2 Cognitive State 44](#_Toc109166690)

[8.4.3 Personal Attitude 44](#_Toc109166691)

[8.4.4 Personal Status 45](#_Toc109166692)

[8.4.5 Meaning 45](#_Toc109166693)

[8.4.6 Object Identifier 45](#_Toc109166694)

[Annex A – MPAI-wide terms and definitions (Normative) 46](#_Toc109166695)

[Annex B - Notices and Disclaimers Concerning MPAI Standards (Informative) 49](#_Toc109166696)

[Annex C – The Governance of the MPAI Ecosystem (Informative) 51](#_Toc109166697)

# Introduction

In recent years, Artificial Intelligence (AI) and related technologies have been applied to a broad range of applications, have started affecting the lives of millions of people and are expected to do even more so in the future. As digital media *standards* have positively influenced industry and billions of people, AI-based data coding *standards* are expected to have a similar positive impact. Indeed, research has shown that data coding using AI-based technologies is generally more efficient than by using existing technologies for, e.g., compression and feature-based description.

However, some AI technologies may carry inherent risks, e.g., in terms of bias toward some classes of users. Therefore, the need for standardisation is more important and urgent than ever and will extend far beyond the traditional scope of standardisation.

The international, unaffiliated, not-for-profit MPAI – Moving Picture, Audio and Data Coding by Artificial Intelligence – Standards Developing Organisation has the mission to develop AI-enabled data coding standards with associated clear licensing framework. In conjunction with MPAI Systems Standards, MPAI Application Standards enable the development of AI-based products, applications, and services.

As a part of its mission, MPAI has developed standard operating procedures to enable users of MPAI implementations to make an informed decision about their applicability. Central to this is the notion of Performance, defined as a set of attributes characterising a reliable and trustworthy implementation of MPAI standards.

Therefore, to fully achieve the MPAI mission, Technical Specifications must be complemented by an ecosystem designed, established, and managed to underpin the life cycle of MPAI standards through the steps of specification, technical testing, assessment of product safety and security, and distribution.

In the following, Terms beginning with a capital letter are defined in *Table 1* if they are specific to this Standard and in *Table 26* if they are common to all MPAI Standards.

The Governance of the MPAI Ecosystem (MPAI-GME), specified in [1], is composed of:

* MPAI as the provider of Technical, Reference Software, Conformance and Performance Specifications.
* Implementers of MPAI standards.
* MPAI-appointed Performance Assessors.
* The MPAI Store which takes care of the secure distribution of validated Implementations.
* Users of Implementations.

The common infrastructure enabling the implementation of MPAI Application Standards and access to the MPAI Store is the AI Framework (AIF) Standard (MPAI-AIF). Figure 1 depicts the MPAI-AIF Reference Model under which Implementations of MPAI Application Standards and user-defined MPAI-AIF conforming applications operate.

An AIF Implementation allows the execution of AI Workflows (AIW), composed of basic processing elements called AI Modules (AIM).

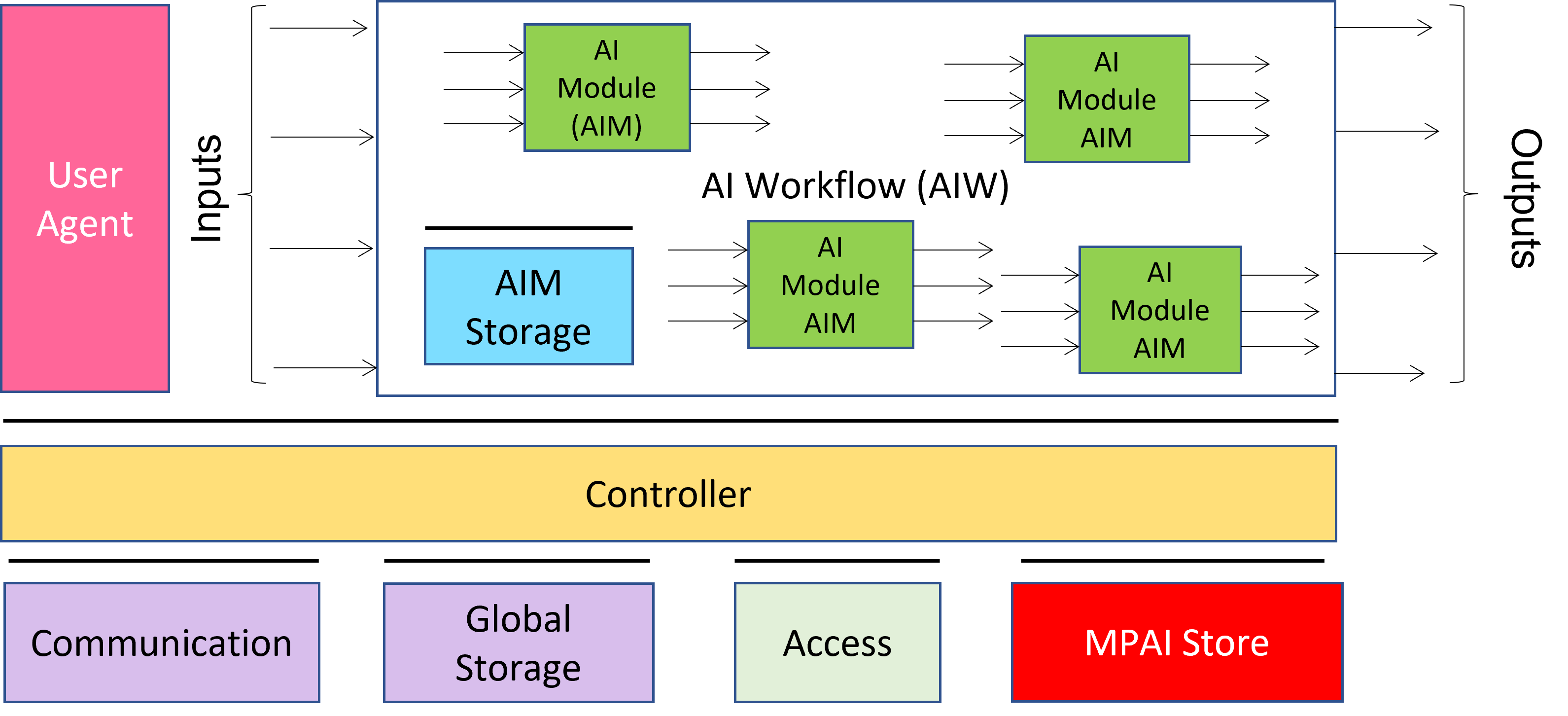


Figure 1 - The AI Framework (AIF) Reference Model and its Components

MPAI Application Standards normatively specify the Function of the AIW and its AIMs, the Syntax and Semantics of the input and output Data of the AIW and its AIMs, and the Connections between and among the AIMs of an AIW. MPAI Application Standards do not specify the internal architecture of the AIMs, which may be based on AI or data processing technologies, and be implemented in software, hardware or mixed software and hardware technologies.

MPAI defines Interoperability as the ability to replace an AIW or an AIM Implementation with a functionally equivalent Implementation. Three Interoperability Levels of an AIW executed in an AIF are defined:

*Level 1 –* Implementer-specific and conforming with the MPAI-AIF Standard.

*Level 2 –* Conforming with an MPAI Application Standard.

*Level 3 –* Conforming with an MPAI Application Standard and certified by a Performance As­sessor.

MPAI offers Users access to the promised benefits of AI with a guarantee of increased transparency, trust and reliability of Implementations as the Interoperability Level of an Implementation moves from 1 to 3.

Additional information on Interoperability Levels is provided in Annex C.

# Scope of the Use Cases

## Use Case description

MPAI-MMC V1 [3] includes 5 Use Cases referenced as Conversation with Emotion (CWE), Multimodal Question Answering (MQA), Unidirectional Speech Translation (UST), Bidirectional Speech Translation (BST) and One-to-Many Speech Translation (MST).

MPAI-MMC V2 adds three new Use Cases to those of MPAI-MMC V1. Some can be considered as extensions of existing Use Cases, and some are new. All share a wide range of technologies already existing or to be standardised by MPAI. However, MPAI wishes to be made aware of new technologies that perform the functionalities of MPAI-MMC Version 2 even if the technologies with those functionalities have already been specified in MPAI-MMC Version 1.

MPAI has identified 2 Composite AIMs that are used in more than one Use Case

1. *Personal Status Extraction (PSE)* performs the following:
   1. Analyses the Manifestations of the Personal Status[[1]](#footnote-1) conveyed by Text, Speech, Face, and Gesture – of a human or an avatar.
   2. Provides an estimate of the said Personal Status (see 2.2.3).
2. *Personal Status Display (PSD)* performs the following:
   1. Receives Text and a Personal Status that may have been generated by a machine, e.g., by conversing with a human or an avatar, or resulting from Personal Status Extraction.
   2. Animates an Avatar Model (from the waist up) using Text and Personal Status.

A brief description of the three MPAI-MMC V2 Use Cases is provided by:

1. *Conversation About a Scene (CAS)*. A human and a machine converse about the objects in a room considered as noiseless:
   1. The human uses a finger to indicate their interest in a particular object.
   2. The machine understands the Personal Status displayed by the human in their Speech, Face, and Gestures, e.g., the human expresses satisfaction because the machine understands their question.
   3. The machine manifests itself as the head-and-shoulders of an avatar whose Speech, Face, and Gesture (limited to Head in this Use Case) convey the machine’s Personal Status resulting from the conversation in a way that is congruent with the speech it utters.
2. *Human-Connected Autonomous Vehicle Interaction (HCI)*. A group of humans converse with a Connected Autonomous Vehicle (CAV) on a domain-specific subject (travel by car):
   1. Humans can converse with the CAV:
      1. Outside the CAV when they ask to be let into the CAV.
      2. Inside the CAV when the humans sit in the cabin.

The two Environments are assumed to be noisy.

* 1. The machine understands the Speech, and the human’s Personal Status shown on their Text, Speech, and Face.
  2. The machine appears as the head and shoulders of an avatar whose Text, Speech, Face, and Gesture (limited to Head in this Use Case) convey a Personal Status congruent with the Speech it utters.

1. *Avatar-Based Videoconference (ABV)*. Avatars representing geographically distributed humans or groups of humans participate in a videoconference:
   1. The avatars reproduce the movements of the upper part of the human participants (from the waist up) with a high degree of accuracy.
   2. The *Virtual Secretary* is a special participant not representing a human participant in the Virtual Environment where the Videoconference is held with the role to:
      1. Make and visually share a summary of what the other avatars say.
      2. Receive comments.
      3. Process the vocal and textual comments considering the avatars’ Personal Status showing in their Text, Speech, Face, and Gestures.
      4. Edit the summary accordingly.
      5. Display the summary.
   3. Each human participant and the meeting manager assign the position to each avatar and the associated speech as they see fit in the meeting’s Virtual Environment.

The functionalities of the Use Cases identified in this document are not intended to exclude support of other functionalities or to preclude the possibility of reducing the identified functionalities. The partitioning of the AIWs in AIMs and the data exchanged between AIMs described in this document are not intended to preclude the possibility of adding or removing data flows or AIMs. MPAI is willing to consider motivated proposals of different AIMs and AIWs if a level of detail comparable to this document is provided.

## Key technologies

The Use Cases of MPAI-MMC V2 require a wide range of advanced technologies. Some have already been standardised in approved MPAI Standards, some are extensions, and some are entirely new. The MPAI-MMC V2 standard will specify some technologies starting from the technologies proposed in response to the companion Call for Technologies [10] document whose functional requirements are provided in Chapter 8. The technologies proposed shall satisfy both these Functional Requirements and the Commercial Requirements given in [11].

Some of the key MPAI-MMC V2 technologies are briefly introduced in the following subsections.

### Audio and Visual Scene Description

This technology concerns the creation of the **Audio-Visual Scene Descriptions** satisfying the requirements of the specific Use Cases, namely:

1. (CAS) A human in a room with a few objects around them converses with a machine indicating the objects they want to talk about. This Use Case requires only the Visual Scene Descriptors.
2. (HCI) Humans converse with the CAV while they are outside it. The CAV individually identifies and locates their utterances in the space outside the CAV, and individually identifies and describes their bodies and speech sources.
3. (HCI) The CAV individually identifies and locates the utterances of the humans inside the cabin and individually identifies, locates, and describes their bodies and speech sources.
4. (ABV) (Groups of) humans, some at the same geographical location, participate in a Videocon­ference represented by their avatars. The participants’ client devices perform the following:
   1. Individually identify and locate their utterances in the Physical Environment providing them as Audio Scene Descriptors.
   2. Individually identify, locate in the Physical Environment, and describe their bodies, providing them as Visual Scene Descriptors.

### User identification

This technology concerns the **identification of a limited number of humans** using the speech and face of humans in the following Use Cases:

1. (HCI) A group of humans outside the CAV wants to be let into the CAV.
2. (HCI) A group of humans inside the CAV (cabin) wants to converse with the CAV.
3. (ABV) Geographically distributed (groups of) humans want to participate in a virtual videoconference represented by their avatars.

Both visual and speech technologies are used for this purpose.

### Personal Status Extraction

As introduced earlier in this document, MPAI defines *Personal Status* as a set of internal characteristics of a person. Currently, Emotion, Cognitive State, and Attitude are considered:

1. Emotion and Cognitive State result from the interaction of a human with the Environment or subsets of it.
   1. *Cognitive State* is more rational (e.g., “Confused”, “Dubious”, “Convinced”).
   2. *Emotion* is less rational (e.g., “Angry”, “Sad”, “Determined”).
2. *Personal Attitude* (called Attitude when there is no confusion with Spatial Attitude) is the stance that a human takes when s/he has developed an Emotion and a Cognitive State (e.g., “Respectful”, “Confrontational”, “Soothing”).

The Personal Status of a human can be displayed in one of the following Modalities: *Text, Speech,* *Face,* or *Gesture*. More Modalities are possible, e.g., the body itself as in body language, dance, song, etc. They are not considered in this document because MPAI has not considered them mature for standardisation. It should be noted that Personal Status may be shown only by one of the four Modalities or by two, by three or by all four simultaneously.

MPAI is seeking technologies that allow a machine to extract the Personal Status of a human or an avatar conveyed by their Text, Speech, Face, and Gesture.

MPAI identifies three steps in the process that allows a machine to extract the Personal Status conveyed by a channel:

1. *Data Capture* (e.g., characters and words, a digitised speech segment, the digital video containing the hand of a person, etc.).
2. *Descriptor Extraction* (e.g., pitch and intonation of the speech segment, thumb of the hand raised, the right eye winking, etc.).
3. *Status Interpretation* (i.e., the *values* of one, two, or all three of Emotion, Cognitive State, and Attitude).

As for any information transmitted over a physical channel, there is no guarantee that the machine will be able to correctly decode the message conveyed. The complexity of the problem may also be compounded by the potential intention of the human transmitter to disguise their Personal Status.

### Personal Status Display

A machine conversing with a human generates a **synthetic Personal Status** conveyed by the *Text, Speech, Face, and Gesture* of an avatar.

This document identifies a Composite Use Case called Personal Status Display and develops the Functional Requirements of the components able to generate Text, Speech, Face, and Gesture from Text, PS-Speech, PS-Face, and PS-Gesture. Note that PS- followed by any of Text, Speech, Face, and Gesture refers humans while PS (), where the bracket includes any of Text, Speech, Face, and Gesture, refers to avatars.

### Collaborative Virtual Environments

To support the Avatar-Based Videoconference Use Case, standards are required for the creation of:

1. *Environment Model* representing the static audio-visual features of a typical Environment where collaborative work is done, e.g., a meeting room.
2. *Audio-Visual Scene Descriptors* describing the dynamic audio-visual parts of the Environment.
3. *Avatar Model* faithfully representing the features of a human specific to the upper part of a body (from the waist up).
4. *Avatar Descriptors* faithfully describing the movement of a human’s face and gesture.

that an *independent* third-party can use to reproduce:

1. A Virtual Environment with the Audio-Visual Scene in it.
2. Avatars whose *Face and Gesture Descriptors* faithfully reproduce the time-dependent changes transmitted by the other party as intended by the sender without requiring access to additional information.

The Use Cases considered in this document do not require that an avatar be able to move itself (e.g., walk) in the Environment.

## Summary of functionalities

A summary of the functionality of the use cases is given by the following list:

1. The **user** selects:
   1. The virtual space (Environment) where the avatars operate.
   2. The avatar model to be animated.
   3. The positions occupied by the avatars in the Virtual Environment.
   4. The user’s Point of View when watching and hearing the 3D Audio-Visual Scene in the Virtual Environment.
2. The **machine**
   1. perceives the Audio-Visual components of the Real Environment and creates:
      1. An Audio Scene Description composed of independent audio objects and their locations.
      2. A Visual Scene Description composed of independent visual objects and their locations.
      3. A complete Audio-Visual Scene Description.
   2. Identifies a human belonging to a group composed of a limited number of humans (closed set identification) using
      1. Speech Descriptors.
      2. Face Descriptors.
   3. Decodes a human’s Personal Status by:
      1. Extracting the Descriptors of the Manifestations of a human’s Personal Status (Text, Speech, Face, and Gesture).
      2. Extracting a human’s Personal Status from Text, Speech, Face, and Gesture Descriptors.
      3. Fusing the different Personal Status in Text, Speech, Face, and Gesture into Personal Status.
   4. Animates a speaking avatar by:
      1. Synthesising speech using Text, Speech, and Personal Status of Text (PS-Text) and Personal Status of Speech (PS-Speech).
      2. Animating the Face using Text, Synthesised Speech, Personal Status of Face (PS-Face).
      3. Animating the Gesture of an avatar using Text and Personal Status of Gesture (PS-Gesture).
   5. Converses with another party (human or avatar) by:
      1. Decoding the other party’s Personal Status (Personal Status).
      2. Animating an avatar representing it (see 5.).
   6. Summarises and refines the speech of other parties by analysing and interpreting their Text, Speech, and Personal Status Manifestations.

# Terms and Definitions

The terms used in this document whose first letter is capital have the meaning defined in Table 1.

*Table 1 – Table of terms and definitions*

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Audio | Digital representation of an analogue audio signal sampled at a frequency between 8-192 kHz with a number of bits/sample between 8 and 32, and non-linear and linear quantisation. |
| Audio Object | Coded representation of Audio information with its metadata. An Audio Object can be a combination of Audio Objects. |
| Audio Scene | The Audio Objects of an Environment with Object location metadata. |
| Audio-Visual Object | Coded representation of Audio-Visual information with its metadata. An Audio-Visual Object can be a combination of Audio-Visual Objects. |
| Audio-Visual Scene | (AV Scene) The Audio-Visual Objects of an Environment with Object location metadata. |
| Avatar | An animated 3D object representing a real or fictitious person in a Virtual Space. |
| Avatar Model | An avatar exposing handles for animation. |
| Cognitive State | An element of the internal status reflecting the way a human or avatar understands the Environment, such as “Confused”, “Dubious”, “Convinced”. |
| Colour (of speech) | The timber of an identifiable voice independent of a current Personal Status and language. |
| Connected Autonomous Vehicle | A vehicle able to autonomously reach an assigned geographical position by:   1. Understanding human utterances. 2. Planning a route. 3. Sensing and interpreting the Environment. 4. Exchanging information with other CAV. 5. Acting on the CAV’s motion actuation subsystem. |
| Descriptor | Coded representation of text, audio, speech, or visual feature. |
| Emotion | An element of the internal status resulting from the interaction of a human or avatar with the Environment or subsets of it, such as “Angry”, “Sad”, “Determined”. |
| Environment | A Physical or Virtual Space containing a Scene. |
| Environment Model | The static audio and visual components of the Environment, e.g., walls, table. and chairs. |
| Face | The portion of a 2D or 3D digital representation corresponding to the face of a human. |
| Factor | One of Emotion, Cognitive State or Attitude. |
| Grade | The intensity of an Emotion. |
| Identity | The label uniquely associated with a human or an avatar. |
| Manifestation | The manner of showing the Personal Status, or a subset of it, in any one of Speech, Face, and Physical Gesture. |
| Meaning | Semantic structure extracted from Text and Personal Status. |
| Modality | One of Text, Speech, Face, or Gesture. |
| Orientation | The 3 yaw, pitch, and roll (α,β,γ) angles of a representative point of an object in the Real and Virtual Space. |
| Personal Attitude | An element of the internal status related to the way a human or avatar intends to position vis-à-vis the Environment or subsets of it, e.g., “Respectful”, “Confrontational”, “Soothing”. |
| Personal Status | The ensemble of information internal to a person, including Emotion, Cognitive State, and Attitude. |
| Personal Status (Face) | The Personal Status, or a subset of it, conveyed by Face. |
| Personal Status (Gesture) | The Personal Status, or a subset of it, conveyed by Gesture. |
| Personal Status (Speech) | The Personal Status, or a subset of it, conveyed by Speech. |
| Physical Gesture | A movement of the body or part of it, such as the head, arm, hand, and finger, often a complement to a vocal utterance. |
| Position | The 3 coordinates (x,y,z) of a representative point of an object in the Real and Virtual Space. |
| Rendered Scene | The output produced by an audio-visual renderer, which renders the Audio-Visual Scene from a selected Point of View. |
| Scene | An Environment populated by humans and real objects (Real Environment) or by avatars and virtual objects (Virtual Environment). |
| Scene Descriptor | An individual attribute of the coded representation of an object in a Scene, including their Spatial Attitude. |
| Spatial Attitude | Position and Orientation and their velocities and accelerations of a Human and Physical Object in a Real or Virtual Environment. |
| Speech | Digital representation of analogue speech sampled at a frequency between 8 kHz and 96 kHz with a number of bits/sample of 8, 16 and 24, and non-linear and linear quantisation. |
| Subword Lattice | A directed graph containing speech recognition sub-word candi­dates. |
| Text | A series of characters drawn from a finite alphabet. |
| Point of View | The Spatial Attitude of a human or avatar looking at the Environment. |
| Virtual Secretary | An avatar in a virtual conference not representing an actual participating human whose role is to:   1. Make and visually share a summary of what other avatars say. 2. Receive comments via Text and Speech. 3. Process the textual comments using the avatars’ Personal Status showing in their Text, Speech, Face, and Gestures. 4. Edit the summary accordingly. 5. Continually display the summary. |
| Virtual Twin | An Audio-Visual Object in a Virtual Environment representing an Audio-Visual Object in a Real Environment. |
| Visual Object | Coded representation of Visual information with its metadata. A Video Object can be a combination of Video Objects. |
| Vocal Gesture | Utterance, such as cough, laugh, hesitation, etc. Lexical elements are excluded. |
| Word Lattice | A directed graph containing speech recognition word candidates. |

# References

## Normative References

This document references the following normative documents:

1. MPAI; Technical Specification: The Governance of the MPAI Ecosystem V1: <https://mpai.community/standards/resources/#GME>
2. MPAI; Technical Specification: AI Framework (MPAI-AIF) V1.1; <https://mpai.community/standards/resources/#AIF>
3. MPAI; Technical Specification: Specification: Multimodal Conversation (MPAI-MMC) V1.2; [https://mpai.community/standards/resources/#MMC](https://mpai.community/standards/resources/" \l "MMC)
4. MPAI; Technical Specification: Context-based Audio Enhancement (MPAI-CAE) V1.3; [https://mpai.community/standards/resources/#CAE](https://mpai.community/standards/resources/" \l "CAE)
5. Universal Coded Character Set (UCS): ISO/IEC 10646; December 2020
6. ISO 639-1:2002 Codes for the Representation of Names of Languages – Part 1: Alpha-2 Code
7. ISO/IEC 14496-10; Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding.
8. ISO/IEC 23008-2; Information technology – High-efficiency coding and media delivery in heterogeneous environments – Part 2: High-Efficiency Video Coding.
9. ISO/IEC 23094-1; Information technology – General video coding – Part 1: Essential Video Coding.
10. MPAI; MPAI-MMC V2 Call for Technologies; <https://mpai.community/standards/mpai-mmc/call-for-technologies-2/mpai-mmc-v2-call-for-technologies/>
11. MPAI; MPAI-MMC V2 Framework Licence; <https://mpai.community/standards/mpai-mmc/framework-licence/mpai-mmc-v2-framework-licence/>

## Informative References

The references below are provided as examples of technologies potentially relevant to this document. MPAI does not endorse any of these technologies as suitable responses to the “requests to respondents” sections of this document.

**General**

1. An example word lattice; <https://www.researchgate.net/figure/An-example-word-lattice_fig1_2361715>

**CAV**

1. MPAI-CAV Use Cases and Functional Requirements WD0.12, [https://mpai.community/standards/mpai-cav/.](https://mpai.community/standards/mpai-cav/)

**Environment**

1. ISO 16739-1:2018 Industry Foundation Classes (IFC) For Data Sharing in The Construction And Facility Management Industries — Part 1: Data Schema
2. <https://technical.buildingsmart.org/standards/ifc/>
3. Khronos; glTF Runtime 3d Asset Delivery; <https://www.khronos.org/gltf/>.
4. Pixar; Universal Scene Description; <https://graphics.pixar.com/usd/release/index.html>.
5. Blender; [www.blender.org](http://www.blender.org).

**Face description**

1. FACS-based Facial Expression Animation in Unity
2. Noldus; Facial Action Coding System; <https://www.noldus.com/applications/facial-action-coding-system>
3. Kalidoface; <https://3d.kalidoface.com/>
4. <https://docs.readyplayer.me/>
5. <https://zivadynamics.com/zrt-face-trainer>

**Audio objects**

1. <https://en.m.wikipedia.org/wiki/Ambisonics>
2. <https://docs.enklu.com/docs/Assets/Audio>
3. Unity Audio Source; <https://docs.unity3d.com/Manual/class-AudioSource.html>
4. Unreal Engine – Audio to Facial Animation SDK

**Presentation/rendering**

1. W3D; WebXR Device API; <https://www.w3.org/TR/webxr/>
2. Unity; <https://unity.com/>
3. UnrealEngine; <https://www.unrealengine.com/>
4. NVIDIA Omniverse™; <https://developer.nvidia.com/nvidia-omniverse-platform>
5. Microsoft Mesh; <https://www.microsoft.com/en-us/mesh>

# Composite AIM Architectures

## Personal Status Extraction

### Scope of Composite AIM

Personal Status Extraction (PSE) is a composite AIM that analyses the Manifestation of a Personal Status conveyed by Text, Speech, Face, and Gesture – of a human or an avatar – and provides the Personal Status estimate. It is used in the Use Case-related figures of this document as a replacement for the combination of the AIMs depicted in Figure 2.

### Reference architecture

Personal Status Extraction produces the estimate of the Personal Status of a human or an avatar by analysing each Modality in 3 steps:

1. *Data Capture* (e.g., characters and words, a digitised speech segment, the digital video containing the hand of a person, etc.).
2. *Descriptor Extraction* (e.g., pitch and intonation of the speech segment, thumb of the hand raised, the right eye winking, etc.).
3. *Personal Status Interpretation* (i.e., one, two, or all three of Emotion, Cognitive State and Attitude).

Figure 2 depicts the Personal Status estimation process used across different Use Cases considered in this document: Descriptors are extracted from Text, Speech, Face Object, and Human Object. Descriptors are interpreted and the specific Manifestations of the Personal Status in the Text, Speech, Face, and Gesture channels derived. Finally, the different estimations of the Personal Status are fused, to the Personal Status.

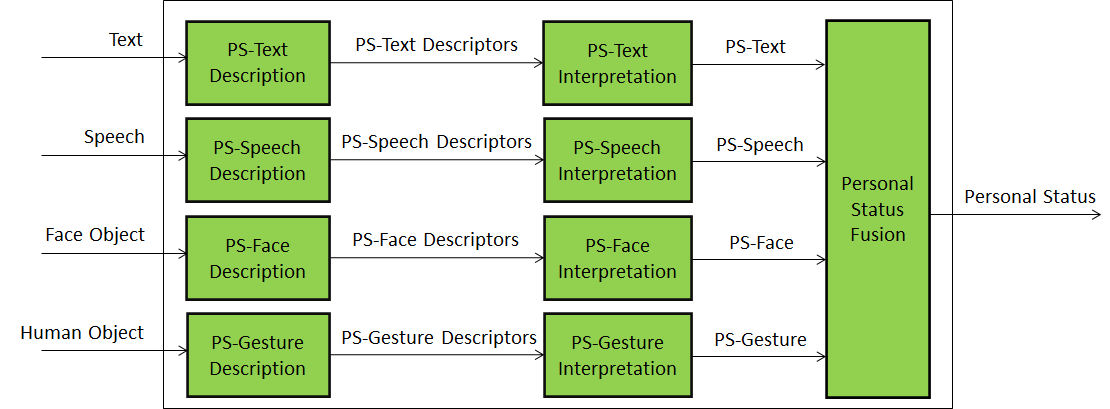


Figure 2 – Reference Model of Personal Status Extraction

An implementation can combine, e.g., the PS-Gesture Description and PS-Gesture Interpretation AIMs into one AIM, and directly provide PS-Gesture from a Human Object without exposing PS-Gesture Descriptors.

### Input/output data

*Table 2* gives the input/output data of Conversation with Emotion.

*Table 2 – I/O data of Personal Status Extraction*

|  |  |  |
| --- | --- | --- |
| **Input data** | **From** | **Comment** |
| Text | Human | Text from a keyboard or recognised speech. |
| Speech | Microphone | Speech of human. |
| Face Object | Camera | The face of the human. |
| Human Object | Camera | The upper part of the body. |
| **Output data** | **To** | **Comments** |
| Personal Status | An AIM | For further processing |

### AI Modules

*Table 3* gives the list of the AIMs with their functions.

*Table 3 – AI Modules of Personal Status Extraction*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| PS-Text Description | Receives Text.  Produces PS-Text Descriptors. |
| PS-Speech Description | Receives Speech.  Produces PS-Speech Descriptors. |
| PS-Face Description | Receives Face.  Produces PS-Face Descriptors. |
| PS-Gesture Description | Receives Gesture.  Produces PS- Gesture Descriptors. |
| PS-Text Interpretation | Receives PS-Text Descriptors.  Produces PS-Text. |
| PS-Speech Interpretation | Receives PS-Speech Descriptors.  Produces PS-Speech. |
| PS-Face Interpretation | Receives PS-Face Descriptors.  Produces PS-Face. |
| PS-Gesture Interpretation | Receives PS-Gesture Descriptors.  Produces PS-Gesture. |
| Personal Status Fusion | Receives PS-Text, PS-Speech, PS-Face, PS-Gesture.  Produces Personal Status. |

## Personal Status Display

### Scope of Composite AIM

A Personal Status Display (PSD) is a Composite AIM receiving Text and Personal Status and generating an avatar producing Text and uttering Speech with the intended Personal Status while the avatar’s Face and Gesture show the intended Personal Status. The Personal Status driving the avatar can be extracted from a human or can be synthetically generated by a machine as a result of its conversation with a human or another avatar. It is used in the Use Case-related figures of this document as a replacement for the combination of the AIMs depicted in Figure 3.

### Reference Architecture

Figure 3 represents the AIMs required to implement Avatar Animation.

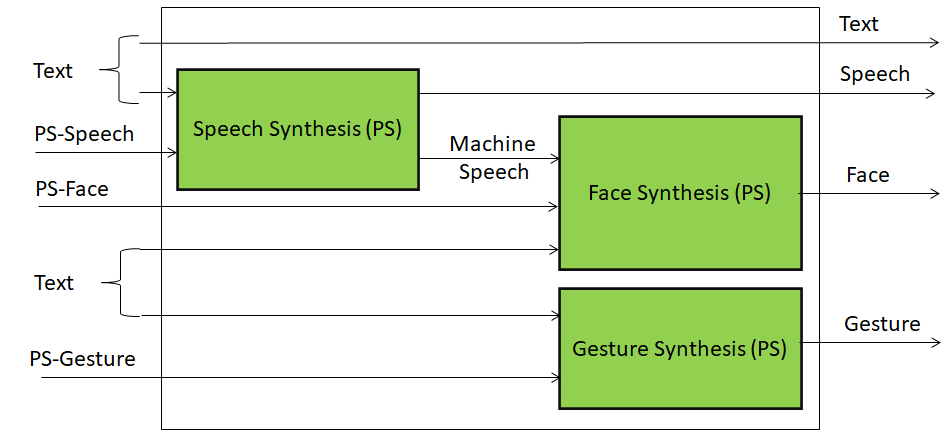


Figure 3 – Reference Model of Personal Status Display

The Personal Status Display operates as follows:

1. Text is synthesised as Machine Speech using the Personal Status provided by PS (Speech). Text is also passed as output.
2. Machine Speech and PS (Face) are used to produce the animated Face.
3. PS (Gesture) is used to synthesise the animated Gesture.

### Input/output data

*Table 4* gives the input/output data of Personal Status Display.

*Table 4 – I/O data of Personal Status Display*

|  |  |  |
| --- | --- | --- |
| **Input data** | **From** | **Comment** |
| Text | Keyboard or Machine |  |
| PS (Speech) | Personal Status Extractor or Machine |  |
| PS (Face) | Personal Status Extractor or Machine |  |
| PS (Gesture) | Personal Status Extractor or Machine |  |
| **Output data** | **To** | **Comments** |
| Machine Text | Human or Avatar |  |
| Machine Speech | Human or Avatar |  |
| Machine Face | Human or Avatar |  |
| Machine Gesture | Human or Avatar |  |

### AI Modules

*Table 5* gives the list of AIMs with their functions.

*Table 5 – AI Modules of Personal Status Extraction*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Speech Synthesis (PS) | Receives Text and PS-Speech.  Produces Machine Speech. |
| Face Synthesis (PS) | Receives Machine Speech and PS-Face.  Produces Machine Face. |
| Gesture Synthesis (PS) | Receives Text and Machine PS-Gesture.  Produces Machine-Gesture. |

# Use Case Architectures

## Conversation About a Scene

### Scope of Use Case

A human holds a conversation with a machine:

1. The machine perceives (sees and hears) the Environment containing a speaking human, and some scattered objects.
2. The machine understands the human’s Speech and gets the human’s Personal Status by capturing Speech, Face, and Gesture, extracting Descriptors, and interpreting the Descriptors.
3. The human converses with the machine and indicates the object in his/her Environment s/he wishes to talk to or ask questions about using Speech, Face, and Physical Gesture.
4. The machine understands which object the human is referring to and generates an avatar that:
   1. Utters Speech conveying a synthetic Personal Status that is pertinent with the human’s Personal Status Manifested in his/her Speech, Face, and Gesture, and
   2. Shows a face conveying a Personal Status that is pertinent with the human’s Personal Status and the Personal Status resulting from the conversation.
5. The machine generates the Scene Presentation of its internally developed representation of the Audio-Visual Scene of the Environment. The human can use the Scene Presentation to understand how the machine sees the Environment. The objects are labelled with the machine’s understanding of their semantics.

### Reference architecture

The Machine operates according to the following workflow:

1. Visual Scene Description produces Face Object and Human Object from Input Video.
2. Speech Recognition produces Recognised Text from Input Speech.
3. Gesture Description produces Gesture Descriptors from Gesture Descriptors.
4. Object Description produces Object Descriptors from Object.
5. Object Identification produced Object ID from Object Descriptors.
6. Personal Status Extraction produces Personal Status.
7. Language Understanding produces Meaning and Language Understanding Text from Personal Status, Recognised Text, and Object ID.
8. Question and Dialogue Processing produces Output Text, PS (Speech), PS (Face) and PS (Gesture) from Personal Status, Meaning, and Language Understanding Text.
9. Avatar Animation produces Machine Text, Machine Speech, Machine Face, Machine Gesture from Output Text, PS (Speech), PS (Gesture).
10. Scene Rendering uses the Visual Scene Descriptors to produce the Rendered Scene as seen from the user-selected Point of View. The rendering is constantly updated as the machine improves its understanding of the scene and its objects.

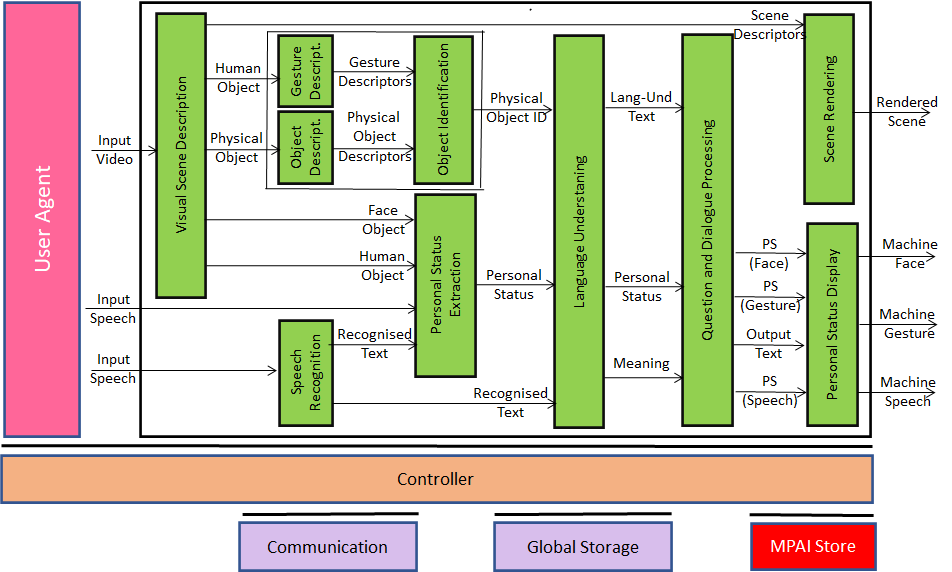


Figure 4 – Reference Model of Conversation About a Scene

### Input/output data

*Table 6* gives the input/output data of Conversation About a Scene

*Table 6 – I/O data of Conversation About a Scene*

|  |  |  |
| --- | --- | --- |
| **Input data** | **From** | **Comment** |
| Input Video | Camera | Points to human and scene. |
| Input Speech | Microphone | Speech of human. |
| Scene Rendering Point of View | Human | The point of view of the scene displayed by Scene Rendering. |
| **Output data** | **To** | **Comments** |
| Output Speech | Human | Machine’s speech. |
| Output Face | Human | Machine’s face. |
| Output Gesture | Human | Machine’s head. |
| Scene Rendering | Human | Reproduction of the scene containing labelled objects as seen by the machine. |

### AI Modules

*Table 7* gives the list of AIMs with their functions.

*Table 7 – AI Modules of Conversation About a Scene*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Visual Scene Description | Receives Input Video.  Produces Face Object and Human Object. |
| Gesture Description | Receives Gesture Descriptors.  Produces Gesture Descriptors. |
| Object Description | Receives Object.  Produces Object Descriptors. |
| Speech Recognition | Receives Input Speech.  Produces Recognised Text. |
| Physical Object Identification | Receives Object Descriptors.  Produces Object ID Object Descriptors. |
| Personal Status Extraction | Receives Face Object, Human Object, Input Speech and Recognised Text.  Produces Personal Status. |
| Language Understanding | Receives Personal Status, Recognised Text, and Object ID.  Produces Meaning and Lang-Und-Text. |
| Question and Dialogue Processing | Receives Personal Status, Meaning, and Lang-Und-Text.  Produces Output Text, PS (Speech), PS (Face) and PS (Gesture). |
| Scene Rendering | Receives Scene Descriptors.  Produces Rendered Scene. |
| Personal Status Display | Receives Output Text, PS (Speech), PS (Face) and PS (Gesture).  Produces Speech, Gesture, and Face. |

## Human-CAV Interaction AIW (HCI)

### Scope of Use Case

This use case is part of the Connected Autonomous Vehicle (CAV) project [13]. A CAV is a system able to execute a command to displace itself based on 1) analysis and interpretation of the data sensed by a range of onboard sensors exploring the environment and 2) information transmitted by other sources in range, e.g., other CAVs, traffic lights and roadside units.

Figure 5 depicts the four subsystems of a CAV.

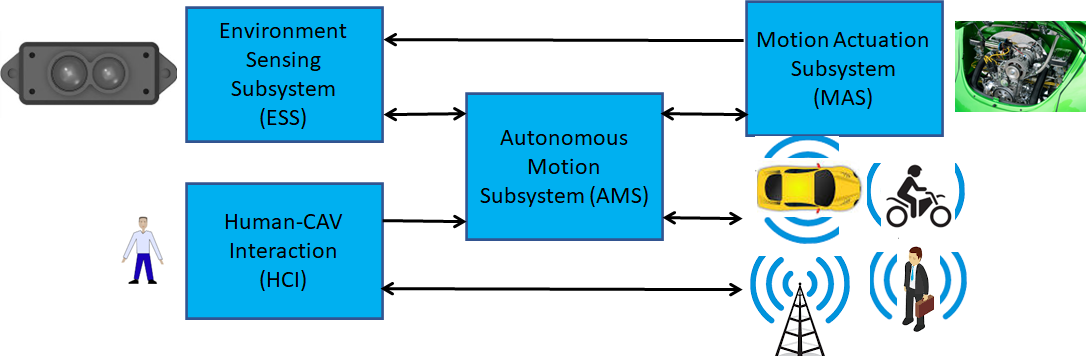


Figure 5 – The Connected Autonomous Vehicle Reference Model

**Human-CAV interaction (HCI)** recognises the human owner or renter, responds to humans’ commands and queries and senses human activities during the travel and may activate other Subsystems in response to humans’ requests. The data exchanged between HCI and the Autonomous Motion Subsystem (AMS) is depicted in Figure 6 but the requirements of the format of the data exchanged between HCI and AMS are not part of this document.

**Environment Sensing Subsystem (ESS)** acquires information from the Environment via a variety of sensors and produces a representation of the Environment (Basic World Representation), i.e., its best understanding of the Environment based on the sensed data.

**Autonomous Motion Subsystem (AMS)** computes the route to destination, uses different sources of information – CAV sensors, other CAVs and transmitting units – to produce a Full World Representation and issues commands to drive the CAV to the intended destination.

**Motion Actuation Subsystem (MAS)** provides non-electromagnetic and non-acoustical environment information¸ and receives and actuates motion commands in the physical world.

Human-CAV Interaction assumes that the CAV is impersonated by an avatar, whose model is selected/produced by the CAV owner or renter. The visible features of the avatar are head, face, and shoulders, and the audible feature is speech, which embeds as much as possible the Personal Status that would be displayed by a human driver in similar conditions.

This use case includes the following functions:

1. Outside the CAV, when a group of humans approaches the CAV:
   1. The HCI creates AV Scene Descriptors providing access to the individual audio and visual objects, in particular the speech and the face of the humans from the rest of the Environment (close to the CAV).
   2. The HCI authenticates the humans it is interacting with by recognising the human CAV rights holder using Speech and Face Descriptors.
2. Inside the CAV, when a group of humans is sitting in the seats:
   1. The HCI creates an AV scene description that separates and locates the visual and speech part of the humans from the rest of the Environment (cabin).
3. The HCI interacts with the humans in the cabin in two ways:
   1. By responding to commands and queries from one or more humans at the same time, e.g.:
      1. Commands to go to a waypoint, park at a place, etc.
      2. Commands with an effect in the cabin, e.g., turn off air conditioning, turn on the radio, call a person, open window or door, search for information etc.

Note: For completeness, Figure 6 includes the conversion of human commands and responses from the CAV, even though this document does not address the format in which HCI interacts with AMS (Autonomous Motion Subsystem).

* 1. By conversing with and responding to questions from one or more humans at the same time about travel-related issues (in-depth domain-specific conversation), e.g.:
     1. Humans request information, e.g., time to destination, route conditions, weather at destination, etc.
     2. Humans ask questions about objects in the cabin.
     3. CAV offers alternatives to humans, e.g., long but safe way, short but likely to have interruptions.

1. The HCI captures the humans’ speech and face, and extracts Text and Personal Status from them.
2. The HCI shows itself as the face and shoulder of an avatar able to:
   1. Utter speech conveying PS (Speech).
   2. Display face conveying PS (Face).
   3. Animate lips in sync with the speech and in line with PS (Face).
   4. Display head conveying PS (Gesture).
3. The HCI may follow the conversation on travel matters held by humans in the cabin if 1) the passengers allow the HCI to do so, and 2) the processing is carried out inside the CAV.
   1. Issue regarding the HCI’s social participation in human activities: under what circum­stances can the HCI take the initiative in starting and participating in the conversation?

### Reference architecture

Figure 6 represents the Human-CAV Interaction (HCI) Reference Model.

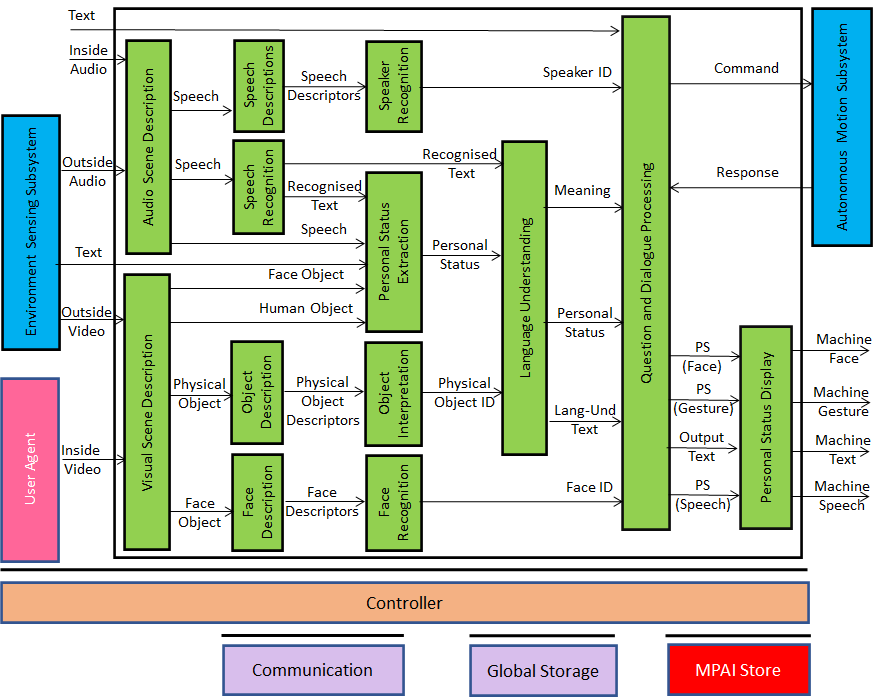


Figure 6 – Human-CAV Interaction Reference Model

HCI operates in two modes:

1. Outdoor: when humans are in the environment and approach the CAV:
   1. HCI separates and locates
      1. The human Face Objects from the Environment.
      2. The human Speech Object from the Environment sound.
   2. HCI identifies the humans using the Face and Speech Objects.
   3. HCI converses with humans on “travel-by-car”-related matters.
2. In the cabin:
   1. HCI separates the Speech Objects from other sounds in the cabin and locates them.
   2. HCI locates the visual elements of humans in the cabin.
   3. HCI identifies the humans.
   4. HCI extracts Personal Status from Speech, Face, and Gesture.
   5. HCI converses with humans in the cabin.

When conversing with the humans in the cabin, the CAV responds by generating an avatar generated by the Personal Status Display.

### Input and output data

*Table 8* gives the input/output data of Human-CAV Interaction.

*Table 8 – I/O data of Human-CAV Interaction*

|  |  |  |
| --- | --- | --- |
| **Input data** | **From** | **Comment** |
| Audio (ESS) | Users in the outside Environment | User authentication  User command  User conversation |
| Audio | Cabin Passengers | User’s social life  Commands/interaction with CAV |
| Text | Cabin Passengers | User’s social life  Commands/interaction with CAV |
| Video (ESS) | Users in the outside Environment | Commands/interaction with CAV |
| Video | Cabin Passengers | User’s social life  Commands/interaction with CAV |
| **Output data** | **To** | **Comments** |
| Output Speech | Cabin Passengers | CAV’s response to passengers |
| Output Face | Cabin Passengers | CAV’s face when speaking |
| Output Gesture | Cabin Passengers | CAV’s head when speaking |
| Output Text | Cabin Passengers | CAV’s response to passengers |

### AI Modules

*Table 9* gives the AI Modules of the Human-CAV Interaction depicted in Figure 3.

*Table 9 – AI Modules of Human-CAV interaction*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Audio Scene Description** | Receives Environment Audio (outside inside)  Produces Audio Scene Descriptors in the outdoor and indoor scenarios. |
| **Visual Scene Description** | Receives Environment Video (outside and inside)  Produces Visual Scene Descriptors in the outdoor and indoor scenarios. |
| **Speech Description** | Receives Speech.  Produces Speech Descriptors. |
| **Speech Recognition** | Receives Speech.  Produces Recognised Text. |
| **Physical Object Description** | Receives Physical Object.  Produces Object Descriptors. |
| **Face Description** | Receives Face Object.  Produces Face Descriptors. |
| **Speaker Recognition** | Receives Speech Descriptors.  Produces Speaker ID. |
| **Personal Status Extraction** | Receives Recognised Text, Speech, Text, Face Object, Human Object.  Produces Personal Status |
| **Object Interpretation** | Receives Physical Object Descriptors  Produces Object ID. |
| **Face Recognition** | Receives Face Descriptors.  Produces Face ID. |
| **Language Understanding** | Receives Recognised Text, Personal Status, and Object ID.  Produces Meaning, Personal Status, Lang-Und-Text. |
| **Question and Dialogue Processing** | Receives Speaker ID, Personal Status, Meaning, Lang-Und-Text, and Face ID.  Produces PS (Face), PS (Gesture) Output Text, and PS (Speech). |
| **Personal Status Display** | Receives PS (Face), PS (Gesture) Output Text, and PS (Speech).  Produces Machine (Face), Machine (Gesture), Machine (Text), and Machine (Speech). |

## Avatar-Based Videoconference

### Scope of Use Case

This Use Case is part of the Mixed-reality Collaborative Space (MCS) project. An MCS is a Virtual Environment where:

1. Virtual Twins of humans – embodied in speaking avatars having a high level of similarity, in terms of voice and appearance, with their Human Twins – are directed by Human Twins to achieve an agreed goal.
2. Human-like speaking avatars, possibly without a visual appearance, not representing a human, e.g., a secretary taking notes of the meeting, answering questions, etc.

The Environment can be anything from a fictitious space to a replica of a real space.

This Use Case addresses the Avatar-Based Videoconference where participants entrust the image of the upper part of their bodies to avatars who utter the participants’ real voice.

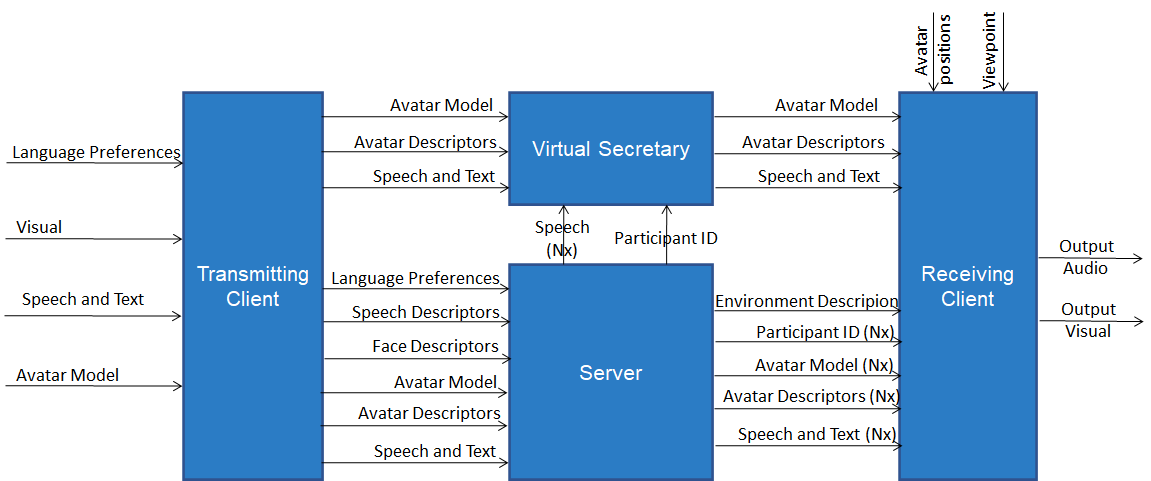


Figure 7 – Avatar-Based Videoconference end-to-end diagram

With reference to Figure 7, the elements characterising the Use Case are:

1. Geographically dispersed humans, some of which are co-located, participate in a videoconference represented by avatar with high similarity with their human twins. The members of a groups have an individual participation in the Virtual Environment where the avatar-based videoconference takes place.
2. The videoconference room in the Virtual Environment is equipped with a table and an appropriate number of chairs hosting:
   1. Virtual Twins of human resembling their Human Twins represented as the upper part of the avatars (waist up).
   2. Avatars not corresponding to any participant, in particular a Virtual Secretary.
3. Each participant selects his/her avatar model.
4. Each participant uses the transmitting part of their client (Transmitting Client) to send the Server:
   1. The participant’s avatar model, language preference, and speech.
   2. The Avatar Descriptors of the participant’s Face and Gesture.
5. The Server
   1. Authenticates all participants using Speech and Face Descriptors.
   2. Sends the Descriptors of the Visual Objects in the Environment, e.g., walls, furniture, etc. to all participants.
   3. Forwards the avatar models and descriptors to all participants.
   4. Uses the language preferences to translate active speech signals to the requested languages.
   5. Sends the original and translated speech signals to participants according to their language preferences.
6. The Virtual Secretary
   1. Collects the statements made by participating avatars while monitoring the avatars’ Personal Statuses conveyed by their speech, face, and gesture.
   2. Makes a summary by combining all information.
   3. Displays the summary in the Environment for avatars to read and edit the Summary directly.
   4. Edits the Summary based on conversations with avatars using Text, Speech, and Personal Status conveyed by Text, Speech, Face and Gesture.
7. Each participant uses the receiving part of their client (Receiving Client) to:
   1. Place all avatar models on a chair of their choice.
   2. Animate the avatar models with the received descriptors.
   3. Attach the speech signals to the mouths of the corresponding avatars, so that each speech is properly located in the Environment (Spatial Audio).
   4. Select a Point of View, possibly different from the position assigned to his/her avatar in the Environment.
   5. Watch the Virtual Environment with the device of his/her choice (HMD or 2D display) and listen to the resulting spatial audio with the device of his/her choice (HMD or earpads).

It should be noted that Figure 7 depicts a particular implementation of the AVB Use Case where the clients execute all functionalities that a client can meaningfully provide. Obviously, this is not the only possible partitioning, but it is the one that best preserves participants’ privacy. Other configurations are possible. For instance, a participant could use a dumb client which only controls scene creation and presents the rendered audio-visual scene generated by the server. Ditto for the transmitting part of the client where the Avatar Descriptors could be created by the server at the cost of sending high bitrate private data.

### Transmitting Client AIW

#### Function

The function of the Transmitting Client AIW is to:

1. Receive:
   1. Input Audio from the microphone (array).
   2. Input Video of the participant from the camera (array).
   3. Participant’s Avatar Model.
   4. Participant’s spoken language preferences (e.g., EN-US, IT-CH).
2. Send to the Server:
   1. Speech Descriptors (for Authentication).
   2. Face Descriptors (for Authentication).
   3. Participant’s spoken language preferences.
   4. Speech (for distribution to all participants).
   5. Avatar Model (for distribution to all participants).
   6. Avatar Descriptors (for distribution to all participants).

#### Architecture

At the start:

1. Each participant sends to the Server:
   1. Language preferences
   2. Avatar model.
2. Speaker Identification sends to Server: Descriptors for Authentication.
3. Face Identification sends to Server: Face Descriptors for Authentication.

During the meeting the following AIMs of the Transmitting Client produce:

1. Audio Scene Description: participants’ speech and location.
2. Visual Scene Description: Face Object and Human Object.
3. Speech Description: Speech Descriptors for Authentication.
4. Speech Recognition: Recognised Text.
5. Face Description1: Face Descriptors.
6. Gesture Description: Gesture Descriptors.
7. Face Description2: Face Descriptors for Authentication.
8. Personal Status Extraction: Personal Status.
9. Language Understanding: Meaning.
10. Avatar Description: Avatar Descriptors.

During the meeting Transmitting Client of each participant sends to Server for distribution to all participants:

1. Speech.
2. Avatar Descriptors.

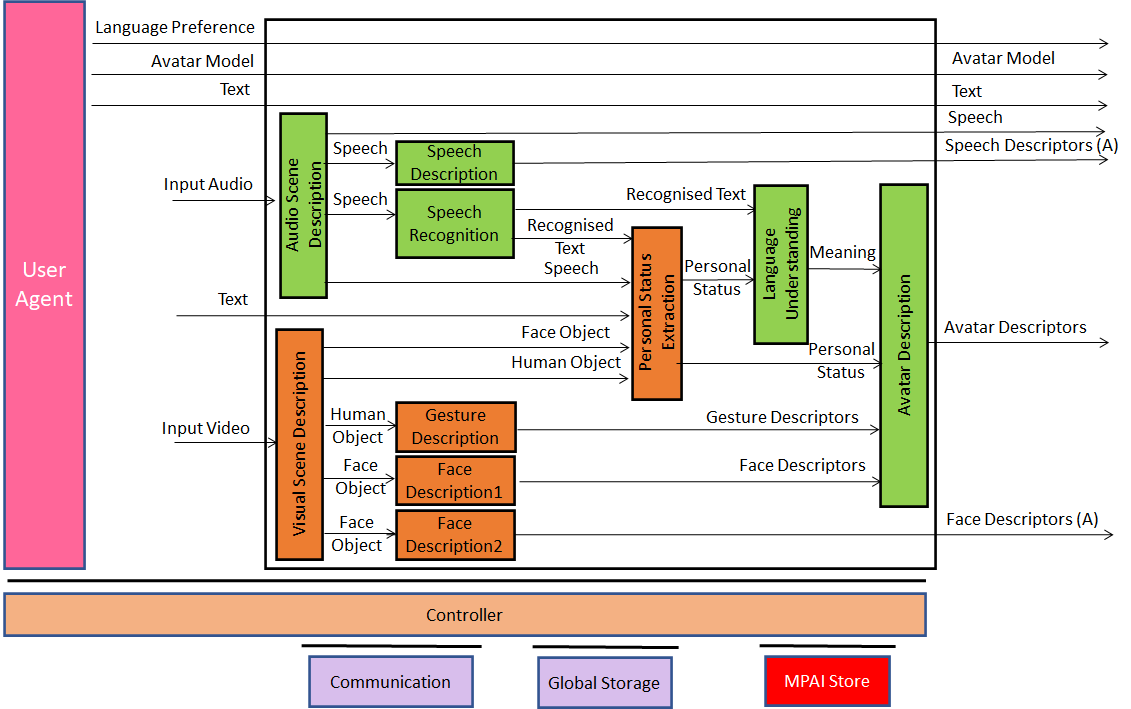


Figure 8 – Reference Model of Avatar Videoconference Transmitting Client

#### Input and output data

*Table 10* gives the input and output data of Transmitting Client AIW:

*Table 10 – Input and output data of Client Transmitting AIW*

|  |  |
| --- | --- |
| **Input** | **Comments** |
| Text | Chat text used to communicate with Virtual Secretary or other participants |
| Language Preference | The language participant wishes to speak and hear at the videoconference. |
| Input Audio | Audio of participant’s Speech and Environment Audio. |
| Input Video | Video of participants’ upper part of the body. |
| Avatar Model | The avatar model selected by the participant. |
| **Output** | **Comments** |
| Language Preference | As in input. |
| Speaker Descriptors (A) | Participants’ Speech Descriptors for Authentication. |
| Participant’s Speech | Speech as separated from Environment Audio. |
| Avatar Descriptors | Descriptors produced by Transmitting Client. |
| Face Descriptors (A) | Participant’s Face Descriptors for Authentication |

#### AI Modules

*Table 11* gives the AI Modules of Transmitting Client AIW.

*Table 11 – AI Modules of Transmitting Client AIW*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Audio Scene Description | Receives Input Audio.  Provides Speech. |
| Visual Scene Description | Receives Input Video  Provides Face Objects and Human Objects. |
| Speech Description | Receives Speech.  Provides Speech Descriptors for Authentication. |
| Speech Recognition | Receives Speech.  Provides Recognised Text. |
| Face Description1 | Receives Face Objects.  Provides the Face channel’s Status. |
| Gesture Description | Receives Human Objects.  Provides the Gesture channel’s Status. |
| Face Description2 | Provides Face Descriptors for Authentication. |
| Personal Status Extraction | Receives Recognised Text, Speech, Text, Face Object, Human Objects.  Provides the Personal Status |
| Language Understanding | Receives Recognised Text and Personal Status.  Provides Meaning. |
| Avatar Description | Receives Meaning, Personal Status, Face Description, and Gesture Description.  Produces the full set of Avatar Descriptions. |

### Server AIW

#### Function

The function of the Server AIW is:

1. At the start:
   1. To Authenticate Participants.
   2. To distribute the Environment Model.
   3. To distribute participants’ Avatar Models.
2. During the videoconference
   1. To Translate and send speech to participants according to their preferences.
   2. To Forward Avatars Descriptors to all participants.

#### Architecture

The Server:

1. Receives from:
   1. Server manager:
      1. Selected Environment.
      2. Participants’ Identities.
   2. Each Participant:
      1. Speech Descriptors (A).
      2. Face Descriptors (A).
      3. Language Preferences.
      4. Speech.
      5. Avatar Descriptors.
2. Authenticates Participants.
3. Translates Participants’ speech according to preferences.
4. Sends:
   1. Environment Descriptors.
   2. Participants’ IDs.
   3. Participants’ speech and translated participants’ speech.
   4. Participants’ Avatar Descriptors.

Figure 5 gives the architecture of Server AIW.

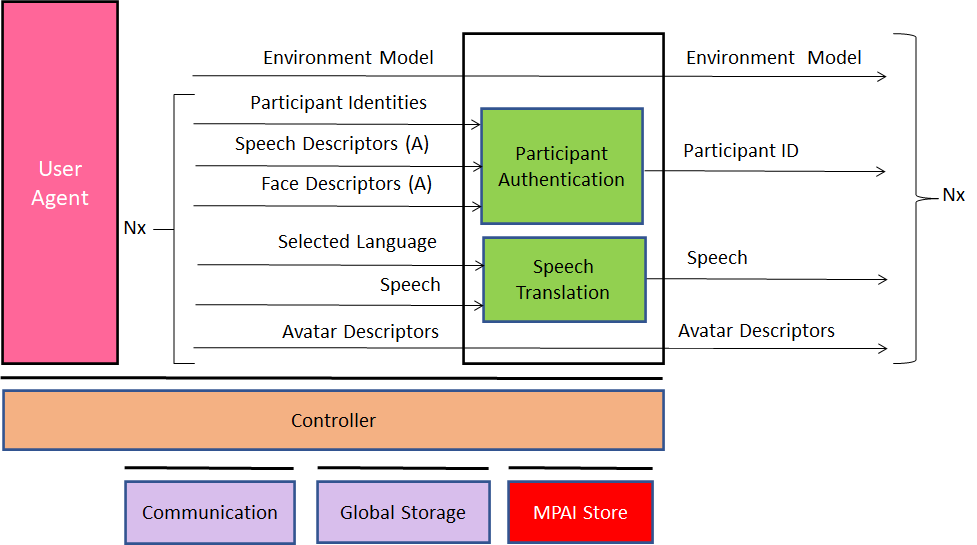


Figure 9 – Reference Model of Avatar-Based Videoconference Server

#### I/O data

*Table 12* gives the input and output data of Server AIW.

*Table 12 – Input and output data of Server AIW*

|  |  |
| --- | --- |
| **Input** | **Comments** |
| Environment Selection | Set by Conference Manager |
| Participant Identities (xN) | Assigned by Conference Manager |
| Speech Descriptors (A) (xN) | Participant’s Speech Descriptors for Authentication |
| Face Descriptors (A) (xN) | Participant’s Face Descriptors for Authentication |
| Selected Language (xN) | From all participants |
| Speech (xN) | From all participants |
| Avatar Model (xN) | From all participants |
| Avatar Descriptors (xN) | From all participants |
| **Outputs** | **Comments** |
| Environment Model (xN) | Static Environment Descriptors |
| Participant ID (xN) | As in input |
| Speech (xN) | As in input |
| Avatar Model (xN) | As in input |
| Avatar Descriptors (xN) | As in input |

#### AI Modules

*Table 13* gives the AI Modules of Server AIW.

*Table 13 – AI Modules of Server AIW*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Environment Description | Creates all static Environment Descriptors. |
| Participant Authentication | Authenticates Participants using their Speech and Face Descriptors |
| Translation | For all participants   1. Selects an active speaker. 2. Translates the Speech of that speaker to the set of translated Speech in the Selected Languages. 3. Assigns a translated Speech to the appropriate set of Partici­pants. |

### Virtual Secretary AIW

#### Architecture

The role of the Virtual Secretary is to:

1. Listen to the Speech of each avatar.
2. Monitor their Personal Status.
3. Draft a Summary using emojis and text in the meeting’s common language (decided outside).

The Summary can be handled in two different ways:

1. Transferred to an external application so that participants can edit the Summary.
2. Displayed to avatars:
   1. Avatars make Speech comments or Text comments (offline, i.e., via chat).
   2. The Virtual secretary edits the Summary interpreting Speech, Text, and the avatars’ Personal Status.

Figure 10 depicts the architecture of the Virtual Secretary AIW.

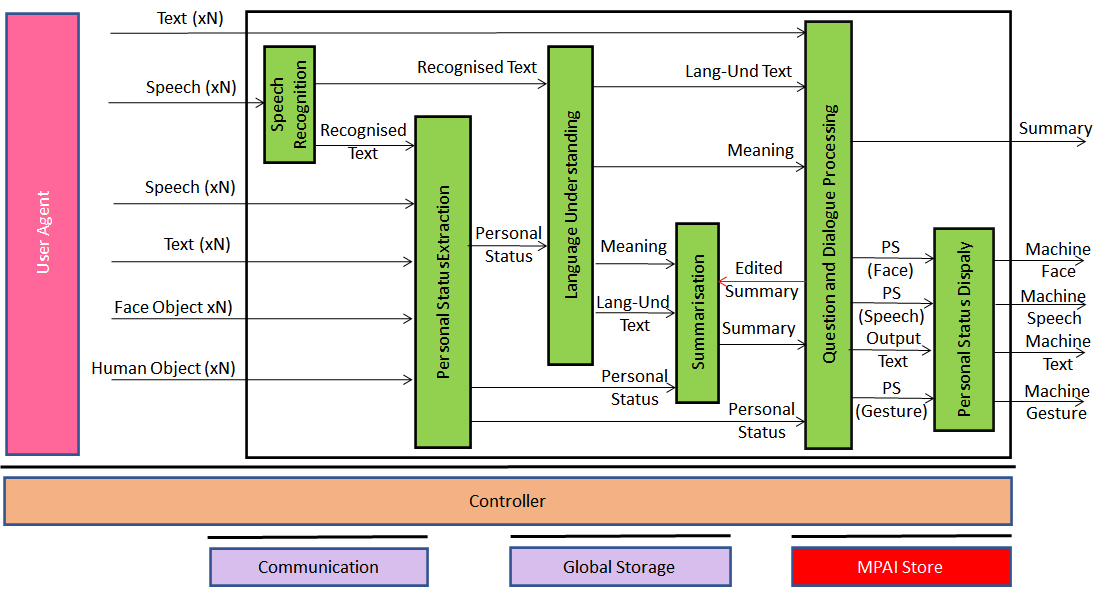


Figure 10 – Reference Model of Virtual Secretary

The Virtual Secretary workflow operates as follows (note that the Virtual Secretary processes one avatar at a time):

1. Speech Recognition extracts Text from an avatar speech.
2. Personal Status Extraction extracts Personal Status from Recognised Text, Speech, Face Object, and Human Object.
3. Language Understanding:
   1. Receives Personal Status and Recognised Text.
   2. Creates
      1. The final version of recognised text (Language Understanding-Text).
      2. Meaning of the sentence uttered by an avatar.
4. Summarisation
   1. Receives:
      1. Language Understanding -Text.
      2. Personal Status.
      3. Meaning.
   2. Creates Summary using emojis and text in the meeting’s common language.
   3. Receives Edited Summary from Question and Dialogue Processing.
5. Question and Dialogue Processing
   1. Receives
      1. Language Understanding -Text.
      2. Text from an avatar via chat.
      3. Meaning.
      4. Summary.
   2. Creates edited Summary.
   3. Sends edited Summary back to Summarisation.
   4. Outputs Text and Personal Status.
6. Speech Synthesis (PS)
   1. Receives Output Text and PS (Speech).
   2. Produces Output Speech with conveyed Personal Status.
7. Personal Status Display generates a manifestation of the machine as an avatar.

#### I/O Data

*Table 14 – I/O data of Virtual Secretary*

|  |  |  |
| --- | --- | --- |
| **Input data** | **From** | **Comment** |
| Text (xN) | Avatars | Remarks on the summary, etc. |
| Speech (xN) | Avatars | Utterances by avatars |
| Face (xN) | Avatars | Faces of avatars |
| Gesture (xN) | Avatars | Gestures of avatars |
| **Output data** | **To** | **Comments** |
| Machine Speech | Avatars | Speech to avatars |
| Machine Face | Avatars | Face to avatars |
| Machine Gesture | Avatars | Gesture to avatars |
| Machine Text | Avatars | Response to chat. |
| Summary | Avatars | Summary of avatars’ interventions |

#### AI Modules

*Table 15* gives the AI Modules of the Virtual Secretary depicted in Figure 10.

*Table 15 – AI Modules of Virtual Secretary*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Speech Recognition | 1. Receives Speech. 2. Produces Recognised Text. |
| Personal Status Extraction | 1. Receives PS-Speech Descriptors, PS-Face Descriptors, PS-Gesture Description. 2. Produces Personal Status. |
| Language understanding | 1. Receives Recognised Text and Personal Status. 2. Produces Language Understanding -Text and Meaning. |
| Summarisation | 1. Receives Meaning and Language Understanding -Text, and edited Summary 2. Produces Summary. |
| Question & Dialogue Processing | 1. Receives Language Understanding -Text, Personal Status, Intention, Meaning. 2. Produces PS (Speech), PS (Face), PS (Gesture), and Output Text, and edited Summary. |
| Personal Status Display | 1. Receives Output Text, PS (Speech), PS (Face), and PS (Gesture). 2. Produces Machine-Text, Machine-Speech, Machine-Face, Machine-Gesture |

### Receiving Client AIW

#### Function

The Function of the Receiving Client AIW is to:

1. Create the Environment Using Environment Descriptors.
2. Place and animate the Avatar Models at positions selected by the participant.
3. Add the relevant Speech to each Avatar.
4. Present the Audio-Visual Scene as seen from the Point of View selected by the participant.

#### Architecture

The Receiving Client AIW:

1. Creates the AV Scene using:
   1. The Environment Model and Descriptors.
   2. The Avatar Models and Descriptors.
   3. The Speech of each Avatar.
2. Presents the Audio-Visual Scene based on the selected viewpoint in the Environment.

Figure 6 gives the architecture of Client Receiving AIW.

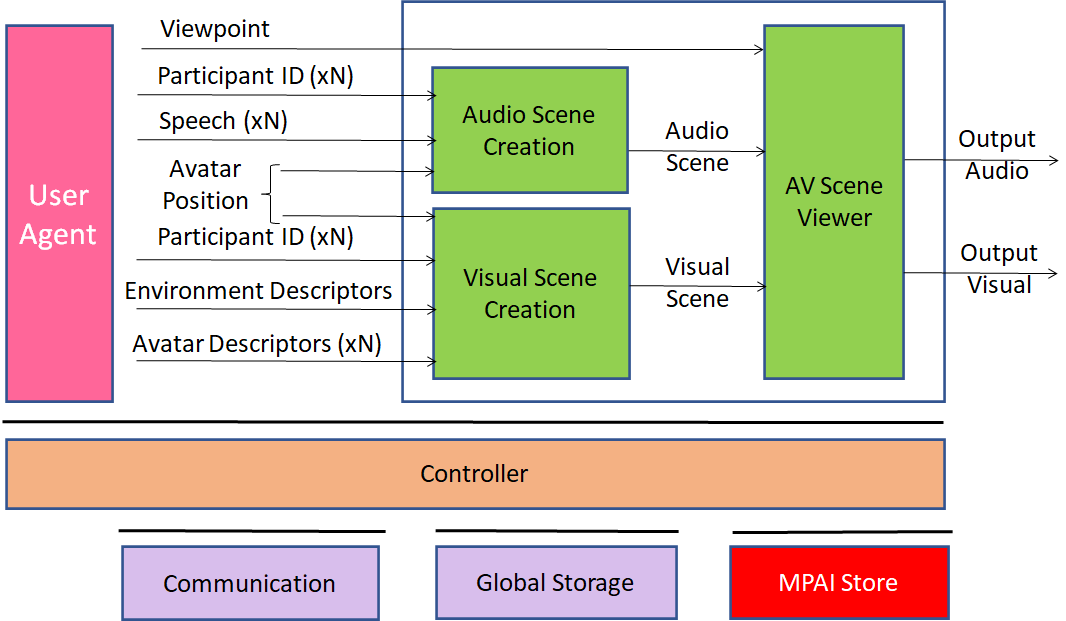


Figure 11 – Reference Model of Avatar-Based Videoconference Receiving Client

#### I/O Data

*Table 16* gives the input and output data of Receiving Client AIW.

*Table 16 – Input and output data of Receiving Client AIW*

|  |  |
| --- | --- |
| **Input** | **Comments** |
| Viewpoint | Participant-selected point to see visual objects and hear audio objects in the Virtual Environment. |
| Avatar Position | Coordinates of Avatars in the Environment |
| Participants’ IDs (xN) | Unique Participants’ IDs |
| Speech (xN) | Possibly translated Participant’s Speech. |
| Environment Descriptors | Static Descriptors of Environment. |
| Avatar Descriptors (xN) | Descriptors of Avatar’s body animation. |
| **Output** | **Comments** |
| Output Audio | Presented using loudspeaker (array). |
| Output Visual | Presented using 2D or 3D display. |

#### AI Modules

*Table 17* gives the AI Modules of Receiving Client AIW.

*Table 17 – AI Modules of Client-Based Environment*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Audio Scene Creation | Creates Audio Scene by combining the speech of speaking Avatars at the respective locations of the Visual Scene. |
| Visual Scene Creation | Creates Visual Scene composed of static Visual Scene Descriptors and Avatars. |
| AV Scene Viewer | Displays Participant’s Audio-Visual Scene from selected Viewpoint. |

# I/O Data summary

## Personal Status Extraction

For each AIM (1st column), *Table 18* gives the input (2nd column) and the output data (3rd column).

*Table 18 – I/O data of Personal Status Extraction AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| PS-Text Description | Text. | PS-Text Descriptors. |
| PS-Speech Description | Speech. | PS-Speech Descriptors. |
| PS-Face Description | Face. | PS-Face Descriptors. |
| PS-Gesture Description | Gesture. | PS- Gesture Descriptors. |
| PS-Text Interpretation | PS-Text Descriptors. | PS-Text. |
| PS-Speech Interpretation | PS-Speech Descriptors. | PS-Speech. |
| PS-Face Interpretation | PS-Face Descriptors. | PS-Face. |
| PS-Gesture Interpretation | PS-Gesture Descriptors. | PS-Gesture. |
| Personal Status Fusion | PS-Text.  PS-Speech.  PS-Face.  PS-Gesture. | Personal Status. |

## Personal Status Display

For each AIM (1st column), *Table 19* gives the input (2nd column) and the output data (3rd column).

*Table 19 – I/O data of Avatar Animation AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| Speech Synthesis (PS) | Text  PS-Speech. | Machine Speech. |
| Face Synthesis (PS) | Machine Speech.  PS-Face.  Text. | Machine Face. |
| Gesture Synthesis (PS) | Text.  PS-Gesture. | Machine Gesture. |

## Conversation About a Scene

For each AIM (1st column), *Table 20* gives the input (2nd column) and the output data (3rd column).

*Table 20 – I/O data of Conversation About a Scene AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| Visual Scene Description | Input Video. | Face Object and Human Object. |
| Speech Recognition | Input Speech. | Recognised Text. |
| Gesture Description | Gesture Descriptors. | Gesture Descriptors. |
| Object Description | Object. | Object Descriptors. |
| Object Identification | Object Descriptors. | Object ID Object Descriptors. |
| Personal Status Extraction | Face Object.  Human Object.  Input Speech.  Recognised Text. | Personal Status. |
| Language Understanding | Personal Status.  Recognised Text.  Object ID. | Meaning.  Lang-Und-Text. |
| Question and Dialogue Processing | Personal Status.  Meaning.  Lang-Und-Text. | Output Text.  PS (Speech).  PS (Face).  PS (Gesture). |
| Personal Status Display | Output Text.  PS (Speech).  PS (Face).  PS (Gesture). | Machine-Text.  Machine-Speech.  Machine-Face.  Machine-Gesture. |
| Scene Renderer | Scene Description. | Rendered Scene. |

## Human-CAV Interaction

For each AIM (1st column), *Table 21* gives the input (2nd column) and the output data (3rd column).

*Table 21 – I/O data of Human-CAV Interaction AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| Audio Scene Description | Environment Audio. | Audio Scene Descriptors. |
| Visual Scene Description | Environment Video. | Visual Scene Descriptors. |
| Speech Description | Speech. | Speech Descriptors. |
| Speech Recognition | Speech. | Recognised Text. |
| Object Description | Object. | Object Descriptors. |
| Face Description1 | Face Descriptors. | Face ID. |
| Speaker Recognition | Speech Descriptors. | Speaker ID. |
| Personal Status Extraction | Recognised Text.  Speech.  Face Object.  Human Object. | Personal Status. |
| Object Interpretation | Object Descriptor. | Object ID. |
| Face Recognition | Face Descriptors. | Face ID. |
| Language Understanding | Recognised Text.  Personal Status.  Object ID. | Meaning.  Personal Status.  Lang-Und-Text |
| Question and Dialogue Processing | Speaker ID.  Personal Status.  Meaning.  Lang-Und-Text.  Face ID. | Output Text.  PS (Speech).  PS (Face).  PS (Gesture). |
| Personal Status Display | Output Text.  PS (Speech).  PS (Face).  PS (Gesture). | Machine-Text.  Machine-Speech.  Machine-Face.  Machine-Gesture |

## Avatar Videoconference

### Transmitting Client

For each AIM (1st column), *Table 22* gives the input (2nd column) and the output data (3rd col-umn).

*Table 22 – AIMs and Data of Transmitting Client AIW*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| Audio Scene Description | Input Audio. | Speech. |
| Visual Scene Description | Input Video | Face Objects.  Human Objects. |
| Speech Description | Speech. | Speech Descriptors (A). |
| Speech Recognition | Speech. | Recognised Text. |
| Face Description1 | Face Objects. | Face Descriptors |
| Gesture Description | Human Objects. | Gesture Descriptors |
| Face Description2 | Face Objects. | Face Descriptors (A). |
| Personal Status Extraction | Recognised Text.  Speech.  Text.  Face Object.  Human Objects. | Personal Status |
| Language Understanding | Recognised Text.  Personal Status. | Meaning. |
| Avatar Description | Meaning.  Personal Status.  Face Description.  Gesture Description. | Avatar Descriptions. |

### Server

For each AIM (1st column), *Table 23* gives the input (2nd column) and the output data (3rd col-umn).

*Table 23 – AIMs and Data of Server AIW*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| Environment Description | Environment Type | Environment Descriptors |
| Participant Authentication | Participant Identities (xN)  Speech Descriptors (A.) (xN)  Face Descriptors (A.) (xN) | Participant ID (xN) |
| Translation | Language Preferences (xN)  Participants’ Speech (xN) | Speech (xN) |
|  | Avatar Descriptors (xN) | Avatar Descriptors (xN) |

### Virtual Secretary

For each AIM (1st column), *Table 24* gives the input data (2nd column) and the output data (3rd column).

*Table 24 – AIMs and Data of Virtual Secretary AIW*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| Speech Recognition | Speech. | Recognised Text. |
| PS-Speech Description | Speech. | PS-Speech Descriptors. |
| PS-Face Description | Face Object. | PS-Face Descriptors. |
| PS-Gesture Description | Human Object. | PS-Gesture Descriptors. |
| Personal Status Extraction | Speech.  Text.  Face Object.  Human Object. | Personal Status. |
| Language understanding | Recognised Text.  Personal Status. | Lang-Und-Text.  Meaning. |
| Summarisation | Meaning.  Lang-Und-Text.  Edited Summary | Summary |
| Question & Dialogue Processing | Lang-Und-Text.  Personal Status.  Meaning.  Edited Summary. | PS (Speech).  PS (Face).  PS (Gesture).  Output Text.  Summary. |
| Personal Status Display | Output Text.  PS (Speech).  PS (Face).  PS (Gesture). | Machine-Text.  Machine-Speech.  Machine-Face.  Machine-Gesture |

### Receiving Client

For each AIM (1st column), *Table 25* gives the input (2nd column) and the output data (3rd col-umn).

*Table 25 – AIMs and Data of Receiving Client AIW*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input** | **Output Data** |
| Audio Scene Creation | Participant ID  Speech | Audio Scene |
| Visual Scene Creation | Participant ID  Environment Descriptors  Avatar Descriptors | Visual Scene |
| AV Scene Viewer | Viewpoint  Audio Scene  Visual Scene | Output Audio  Output Visual |

# Functional Requirements

This chapter presents the functional requirements of the data formats required by the Use Cases identified in this document. They belong to the following categories:

1. Digital representations of analogue signals: e.g., Multichannel Audio, Video.
2. Natively digital data formats: e.g., Text or data formats specific to the Virtual Environment.
3. Descriptors, i.e., the features extracted from Face, Physical Gesture, Speech, and Audio and Visual Scenes.
4. Interpretations, i.e., the semantics of Descriptors and their combinations.

## Digital representation of analogue signals

### Microphone Array Audio

MPAI needs Microphone Array Audio to support the requirements of the following existing and new Use Cases:

1. Enhanced Audioconference Experience (EAE) [4].
2. Conversation About a Scene (CAS).
3. Human-CAV Interaction (HCI).
4. Avatar-Based Videoconference (ABV).
5. Audio-On-the-Go (AOG).

In [4] MPAI has standardised Microphone Array Audio defined as Interleaved Multichannel Audio whose channels are sampled at a minimum of 5.33 ms (e.g., 256 samples at 48 kHz) to a maximum of 85.33 ms (e.g., 4096 samples at 48 kHz) and each sample is quantised with 16 or 24 or 32 bits.

**To respondents**

Respondents are requested to propose extensions to [4] or to propose alternative representation formats supporting the Use Cases described in this document:

1. HCI – Humans approach a CAV and converse with it.
2. HCI – Humans are in the cabin and converse with the CAV.
3. ABV – Human/s is/are in a room attending a virtual videconference

### 2D Video

Video may be used to capture visual information to create a Visual Scene as identified in this document.

MPAI [3] specifies Video as:

1. Pixel shape: square
2. Bit depth: 8 or 10 bits/pixel
3. Aspect ratio: 4/3 or 16/9
4. 640 < # of horizontal pixels <1920
5. 480 < # of vertical pixels <1080
6. Frame frequency 50-120 Hz
7. Scanning: progressive
8. Colorimetry: ITU-R BT709 or BT2020
9. Colour format: RGB or YUV
10. Compression, either:
    1. Uncompressed.
    2. Compressed according to one of the following standards: MPEG-4 AVC [7], MPEG-H HEVC [8], MPEG-5 EVC [9]

**To respondents**

Respondents are invited to comment on MPAI’s parameter choice for 2D Video.

### 3D Video

3D Video may be used to capture visual information to facilitate the creation of a Visual Scene. This can result from one camera an array of cameras having video + depth as the baseline format or other 3D Video data formats.

**To respondents**

Respondents are requested to provide a 3D Video format to create Visual Scene Descriptors as identified in this document:

1. CAS – A human in a room with objects.
2. HCI – Humans around a CAV.
3. HCI – Humans sitting in the cabin.
4. ABV – Humans sitting in a room attending a conference.

## Natively digital data formats

### Text

For all instances of text, MPAI specifies ISO/IEC 10646, Information technology – Universal Coded Character Set (UCS) [5] as Text representation because of its ability to represent the characters used by most languages currently in use.

**To respondents**

Respondents are invited to comment on this choice.

### Recognised Text

Recognised Text can be represented as simple Text or as a Word or as Sub-word lattice.

**To respondents**

Respondents are requested to propose a format for Recognised Text:

1. Whose start is timestamped in relation to the selected unit, e.g., word, syllable, phoneme.
2. That provides:
   1. Time interval (duration).
   2. Strings of characters or phonemic symbols.
   3. Corresponding probability.

### Language Understanding (Text)

Language Understanding (Text) is Text revised by the Language Understanding Module.

**To Respondents**

Respondents are requested to comment on the above.

### Summary

Summary is data structure composed of Text possibly enhanced by characters expressing a Personal Status in a format suitable for human editing.

**To respondents**

Respondents are invited to propose a format satisfying the requirements.

### Environment Model

The Environment Model is the data format describing a Virtual Environment.

An Environment Model can be:

1. Captured from a Real Environment.
2. Synthetically generated.

To produce a Virtual Environment:

1. Captured from a Real Environment, or
2. Synthetically generated.

The Environment Model shall have a Visual and Audio component and satisfy the following functional requirements:

1. The Visual component shall:
   1. Describe the elements bounding the Environment (e.g., walls, ceiling, floor, doors, windows).
   2. Describe the Visual Objects in the Environment (e.g., table, swivel chair, and furniture) at given coordinates.
   3. Support uniform lighting in the environment.
   4. Support integration of avatars with animated face and gesture.
2. The Audio component shall:
   1. Describe the acoustic characteristics of the environment, e.g., reverberation time.
   2. Describe the Audio Objects in the Environment.
   3. Support the integration of Audio Objects (e.g., the speech associated to an avatar) in the environment at given coordinates and directions.
3. It shall be possible to associate Audio Objects to Visual Objects, e.g., a Speech to an Avatar and vice-versa.
4. The party receiving an Environment Model shall be able to use it without requiring additional information from the party who created it.

**To Respondents**

MPAI requests respondents:

1. To propose a format for the Environment Model that would allow the digital representation and rendering of a static Environment with the features described and with the ability to accommodate at runtime any other dynamic Audio and Visual Descriptors identified in this document.
2. To comments on the use of the Industry Foundation Classes (IFC)/BIM [14] file for the visual component or to propose alternative formats.

### Avatar Model

The Avatar Model is the data format describing a static avatar from the waist up displaying movements in face and gesture.

An Avatar Model can be:

1. Captured from an instance of a human.
2. Synthetically generated.

With the goal of producing an Avatar Model:

1. Captured from an instance of a human, or
2. Synthetically generated.

The Avatar Model shall be able to:

1. Represent the face, head, arms, hands, and fingers specific of a human.
2. Display a given Personal Status in face, head, arms, hands, and fingers.
3. Animate the lips based on Text, Speech, and Personal Status.
4. Animate head, arms, hands, and fingers based on Text and Personal Status.

The Avatar Model shall have a standard format, i.e., a party shall be able to use an Avatar Model as intended by the party who created it without requiring access to additional information.

**To Respondents**

MPAI requests respondents to propose an Avatar Model satisfying the stated requirements.

### Human Object

Human Object is an object in a Visual Scene corresponding to a human satisfying the following requirements:

1. It represents the upper part of the body (from the waist up) with high accuracy.
2. It gives access to the following components of the human body: face, head, arms, hands, and fingers.

**To respondents**

Respondents are invited to propose a Human Object format that supports the requirements.

### Face Object

Face Object is the 2D image captured from a camera or the 3D image of a face extracted from a Visual Scene Description that can be used for different purposes, such as:

1. To extract the identity from the Face Object.
2. To extract the Personal Status from the Face Object.
3. To extract the spatial coordinates of the Face Object.

The Face Object shall be organically part of a Human Object.

**To respondents**

Respondents are invited to propose a Face Object format satisfying the above requirements.

### Head Object

Head Object is the 2D image captured from a camera or the 3D image of a head extracted from a Visual Scene Description that

1. Is organically part of the Human Object.
2. Can be used to extract Head Descriptors.

**To respondents**

Respondents are invited to propose a Head Object format satisfying the above requirements.

### Arm Object

Arm Object is the 2D image captured from a camera or the 3D image of an arm extracted from a Visual Scene Description that

1. Is organically part of the Human Object.
2. Can be used to extract Arm Descriptors.

**To respondents**

Respondents are invited to propose an Arm Object format satisfying the above requirements.

### Hand Object

Hand Object is the 2D image captured from a camera or the 3D image of an arm extracted from a Visual Scene Description that

1. Is organically part of the Arm Object.
2. Can be used to extract Hand Descriptors.

**To respondents**

Respondents are invited to propose a Hand Object format satisfying the above requirements.

### Finger Object

Finger Object is the 2D image captured from a camera or the 3D image of an arm extracted from a Visual Scene Description that

1. Is organically part of the Hand Object.
2. Can be used to extract Finger Descriptors.

**To respondents**

Respondents are invited to propose a Finger Object format satisfying the above requirements.

### AV Scene Renderer Format

The format used by the audio-visual renderer to present the Audio-Visual Scene as used, e.g., in the HCI Use Case and in the receiving client of the ABV Use Case. This would be the format that a server performing the function of the client would send to a render-only client.

**To Respondents**

MPAI requests respondents to propose a suitable data format for the purpose indicated.

### Spatial Attitude

Spatial Attitude is the set of 18 values (x,y,z,α,β,γ; their 1st order derivatives; their 2nd order derivatives) corresponding to the Position and Orientation and their derivatives of an Object in a Real or Virtual Environment. The position of the object corresponds to that of a representative point of the object.

Use Cases in this document need to refer to a coordinate system to be able to express the spatial position and orientation of a real or virtual object.

Four different cases are considered:

1. The coordinates are absolute on the surface of the Earth, e.g., in the case of the objects in a World Representation used by a Connected Autonomous Vehicle.
2. The coordinates are defined with reference to a specific point on the Earth.
3. The coordinates are arbitrarily defined local coordinates. The Use Case defines what is the point with coordinates (0,0,0) and angles with values (0,0,0).
4. The coordinates are in a virtual space. The Use Case defines what is the point with coordinates (0,0,0) and angles with values (0,0,0).

**To Respondents**

MPAI requests respondents to comment on the above and propose a format.

### Point of View

The point expressed as coordinates from where a user looks at the space around him/her (x,y,z) or hears the sound field in a given direction (α,β,γ). Point of View is a special case of Spatial Attitude.

**To Respondents**

MPAI requests respondents to comment on the above.

## Descriptors

### Audio Scene Descriptors

Audio Scene Descriptors describe the structured composition of Audio Objects from two different perspectives:

1. The Audio Scene is *captured from the real world*, e.g.:
   1. A group of humans approach a CAV or are sitting in the CAV cabin. A microphone array in the CAV separates and spatially locates the different sound sources (speech and noise).
   2. A group of humans is sitting in a room. A microphone array in the room separates and spatially locates the different sound sources (speech and noise).
2. The Audio Scene is *used to create a sound field* in an Environment, e.g.:
   1. In Avatar-Based Videoconference, avatars are sitting around a table at the Spatial Attitude decided by the human participant and the mouth of each avatar emits the speech of the human it represents.

Audio Scene Descriptors shall have a standard format, i.e., a party shall be able to use the Audio Scene Descriptors as intended by the party who created it without requiring access to additional information. MPAI-CAE V1 has standardised an audio scene description for 1.b [4].

**To Respondents**

MPAI requests respondents to propose Audio Scene Descriptors satisfying the needs of the Use Cases described above and other Use Cases of this document.

The Audio Scene Description should provide:

1. Access to the individual objects (Speech Objects).
2. The spatial coordinates and the directions of the individual objects in a specified coordinate system.

### Visual Scene Descriptors

Visual Scene Descriptors describe the structured composition of Visual Objects in a Visual Scene serving two purposes:

1. *The Visual Scene is captured from an Environment* in three use cases:
   1. (CAS) A human is part of a scene populated by objects where a camera:
      1. Separates the human from the rest of the environment.
      2. Spatially locates the human.
      3. Captures face and gesture of the human.
   2. (HCI) A group of humans approach a CAV or are sitting in the cabin and the CAV:
      1. Separates the different visual objects (humans and other objects),
      2. Identifies the faces of all humans.
      3. Spatially locates all objects as they move.
   3. (ABV) A human/a group of humans is/are sitting in a room and a camera:
      1. Separates the humans from the rest of the environment.
      2. Spatially locates them.
      3. Captures their faces and gestures.
2. *The Visual scene is created in an Environment*, Virtual (VR) and Real (AR) in three use cases:
   1. (CAS) An avatar displays the face of a machine conversing with the human manifesting its Personal Status.
   2. (HCI) An avatar represents the virtual CAV driver displaying its Personal Status (face and head) congruent with the speech it utters.
   3. (ABV) Avatars sit around a table and reproduce faces, heads, arms, hands, and fingers of the humans they represent.

A party shall be able to use the Visual Scene Descriptors as intended by the party who created it without requiring access to additional information.

**To Respondents**

MPAI requests respondents to propose Visual Scene Descriptors representing the Spatial Attitude of the objects relevant to the 3 Use Cases of this document, i.e.:

1. Human Objects (in CAS, HCI, and ABV).
2. Face Objects (in CAS, HCI, and ABV).
3. Physical Objects (in CAS and HCI).

The Objects may have a static or quasi-static Spatial Attitude. However, face, head, arms, hands, and fingers can be change over time.

The Visual Scene Description should provide:

1. A reference set of spatial coordinates.
2. Access to the individual objects of a scene for specific processing such as the extraction of Face Object or the extraction Face and Gesture Descriptors from the Human Object.

### Audio-Visual Scene Descriptors

Audio-Visual Scene Descriptors describe the structured composition of the Audio-Visual Objects in an audio-visual scene.

The Audio-Visual Scene Descriptors can be:

1. Captured from an instance of a human.
2. Synthetically generated

with the goal of producing an Audio-Visual Scene:

1. Captured from an instance of a human, or
2. Synthetically generated.

Or reproducing a Physical Environment.

The Audio-Visual Scene Descriptors shall:

1. Have a common time base.
2. Enable the association of co-located audio and visual objects if both are available.
3. Support the physical displacement and interrelation (e.g., occlusion) of audio and visual objects over time.

For example:

1. Description of the audio-visual objects in the Real Environment where a group of speaking humans with a background noise.
2. Description of the audio-visual objects in a Virtual Environment (a conference room) attended by a set of speaking avatars representing humans and a Virtual Secretary.

The Audio-Visual Scene Descriptors shall have a standard format, i.e., a party shall be able to use Audio-Visual Scene Descriptors as intended by the party who created them without requiring access to additional information.

**To Respondents**

MPAI requests respondents to propose Audio-Visual Scene Descriptors that integrate Audio Scene Descriptors and Visual Scene Descriptors serving the needs of the Use Cases of this document.

### Avatar Descriptors

Avatar Descriptors represent the instantaneous alterations of the face, head, arms, hands and fingers of an Avatar Model.

Avatar Descriptors can be:

1. Captured from an instance of a human.
2. Synthetically generated.

with the goal of producing Avatar Descriptors of an avatar:

1. Captured from an instance of a human, or
2. Synthetically generated.

Avatar Descriptors shall have a standard format, i.e., a party shall be able to use Avatar Descriptors as intended by the party who created them without requiring access to additional information.

**To Respondents**

MPAI requests respondents to propose a set of Avatar Descriptors supporting the stated requirements.

### Face Descriptors

MPAI needs Face Descriptors for the following possibly overlapping purposes:

1. To recognise the identity of a limited number of humans, e.g.,
   1. Members of a family.
   2. Customers of a CAV-renting company.
   3. Participants in a meeting.
2. To describe the features of a face for the purpose of extracting:
   1. The physical features of a face.
   2. The movement of the mobile parts of a face, e.g., lips and eyes.
   3. The Personal Status conveyed by a Face Object.
3. To animate the face of an Avatar Model using any of the points in 2.
4. To specify the coordinates of the point representing the face assumed to be a rigid body.
5. To specify the trajectory of the point representing a moving face.

For instance, the x(t), y(t), z(t) coordinates of the point representing the face and the yaw, pitch, roll.

Key Points are examples of Descriptors. They should be technology-independent (i.e., image processing, or ML, or their combinations).

**To Respondents**

Respondents are requested to propose Facial Descriptions that can be used for the purposes identified above.

### Gesture Descriptors

Gesture is the ensemble of head, arm, hand, and finger movement and Gesture Descriptors are the organised composition of Head Descriptors, Arm Descriptors, Hand Descriptors, and Finger Descriptors.

Gesture Descriptors are used to:

1. Represent arbitrary movement of head, arms, hands, and fingers.
2. Recognise:
   1. Sign language.
   2. Culture-dependent signs (e.g., mudra sign).
   3. Coded hand signs, e.g., to indicate a particular object in a scene.
   4. A human’s Personal Status conveyed by Gesture.
3. Animate an avatar’s head, arms, hands, and fingers.

**To Respondents**

MPAI requests syntax and semantics of arm, hand, finger key points that are independent of the technology used to implement the Key Point Detection AIM – i.e., image processing, or ML, or combinations of the two.

### Head Descriptors

Head Descriptors are used to describe:

1. The Head Object for the purpose of extracting:
   1. The physical features of a head.
   2. The Manifestation of the Personal Status in the head.
2. The coordinates of the point representing the head assumed to be a rigid body.
3. The trajectory of the point representing a moving head.
4. The rotation of a moving head.

For instance, the x(t), y(t), z(t) coordinates of the point representing the head and the yaw, pitch, and roll rotation.

**To Respondents**

MPAI requests respondents to propose a set of Head Descriptors satisfying the requirements.

### Arm Descriptors

Arm Descriptors are used to describe:

1. The Arm Object for the purpose of extracting:
   1. The physical features of an arm.
   2. The Manifestation of Personal Status in the arm.
2. The coordinates of the 2 components of an arm assumed to be rigid bodies.
3. The trajectory of the points representing a moving arm.

**To Respondents**

MPAI requests respondents to propose a set of Arm Descriptors satisfying the requirements.

### Hand Descriptors

Hand Descriptors are used to describe:

1. The Hand Object for the purpose of extracting:
   1. The physical features of a hand.
   2. The Manifestation of Personal Status in the hand.
2. The coordinates of the hand assumed to be a rigid body.
3. The trajectory of the points representing a moving hand.

**To Respondents**

MPAI requests respondents to propose a set of Hand Descriptors satisfying the requirements.

### Finger Descriptors

Finger Descriptors are used to describe:

1. The Finger Object for the purpose of extracting:
   1. The physical features of a finger.
   2. The Manifestation of Personal Status in the fingers.
2. The coordinates of
   1. The thumb assumed to be composed of two rigid bodies
   2. The four fingers assumed to be composed of three rigid bodies.
3. The trajectory of the points representing a moving finger.

**To Respondents**

MPAI requests respondents to propose a set of Finger Descriptors satisfying the requirements.

### Speech Descriptors

MPAI-MMC V1.2 [3] and MPAI-CAE V1.3 have standardised Speech Descriptors (Features). Speech Descriptors shall satisfy the requirement Ability to recognise the identity of a limited number of humans, e.g.:

1. Members of a family.
2. Customers of a CAV-renting company.
3. Participants in a meeting.
4. Ability to describe the features of a speech segment for the purpose of extracting the Personal Status conveyed by the speech segment.

**To Respondents**

MPAI requests respondents to propose Speech Descriptors satisfying the above requirements.

### Text Descriptors

To support the Use Cases of this document, Text Descriptors are needed to describe elements of e.g., sentiment analysis of a text, such as positive /negative words, emojis.

**To Respondents**

MPAI requests respondents to propose Text Descriptors satisfying the above requirements.

## Interpretations

### Emotion

In [3] MPAI has standardised a set of basic Emotions and their semantics to represent emotion conveyed by text, speech, face, and gesture. The same set is assumed to represent the emotion conveyed by any Modality.

**To Respondents**

Respondents are requested to propose the following:

1. To comment on the use of [3] to represent emotion. If a respondent claims that Basic Emotion Set of [3] is unsuitable for the Use Cases of this document, respondents are requested to motivate their claims, and propose an extension of the MPAI Basic Emotion Set or a new solution.
2. To comment on the use of the same set of emotions for text, speech, face and gesture.

### Cognitive State

The internal status of a human reflecting his/her understanding of the Environment, such as “Confused” or “Dubious” or “Convinced”.

**To Respondents**

Respondents are requested to propose the following:

1. A basic set of Cognitive States suitable for the Text, Speech, Face, and Gesture Modalities preferably using the same structure adopted by MPAI for Emotion [3].
2. Motivate the use of different basic sets of Cognitive States for the different Modalities.

### Personal Attitude

The internal status of a human related to the way s/he intends to position him/herself vis-à-vis the Environment, especially social, or subsets of it, e.g., “Confrontational”, “Respectful”, “Soothing”.

**To Respondents**

Respondents are requested to propose the following:

1. A basic set of Attitudes suitable for the Text, Speech, Face, and Gesture Modalities preferably using the same structure adopted by MPAI for Emotion [3].
2. Motivate the use of different basic sets of Attitudes for the different Modalities.

### Personal Status

Personal Status (PS) indicates a set of 3 Factors {Emotion, Cognitive State, Attitude} conveyed by one or more Modalities {Text, Speech, Face, and Gesture Modalities}. Each Factor may be present or absent, and may be time dependent, i.e., it could have a timestamp, say, T1 and a duration D after which a different Personal Status may be displayed at time T2=T1+D.

Personal Status may be the result of fusing the Personal Statuses conveyed by at least two of the following channels: Text, Speech, Face, and Gesture.

**To Respondents**

Respondents are requested to propose the following:

1. A Personal Status format capable of describing the evolution of Personal Status over time.
2. A Fused Personal Status format supporting the requirements to:
   1. Include the Emotion, Cognitive Status, and Attitude making up a Personal Status.
   2. Retain information on the measured values of the different factors in a Personal Status conveyed by the different Modalities.
   3. Describe the evolution of Personal Status over time.

### Meaning

Meaning is information extracted from an input text and Personal Status such as question, statement, exclamation, expression of doubt, request, invitation. MPAI-MMC [3] specifies a digital representation format for Meaning.

**To respondents**

Respondents are requested to comment on the suitability of the technology standardised in [3].

### Object Identifier

MPAI-MMC [3] specifies an Object Identifier to be used to identify a visual object.

**To respondents**

Respondents are requested to comment on the suitability of the technology standardised in [3] for the Conversation About a Scene and Human-CAV Interaction Use Cases.

1. – MPAI-wide terms and definitions (Normative)

The Terms used in this standard whose first letter is capital and are not already included in Table 1 are defined in *Table 26*.

*Table 26 – MPAI-wide Terms*

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Access | Static or slowly changing data that are required by an application such as domain knowledge data, data models, etc. |
| AI Framework (AIF) | The environment where AIWs are executed. |
| AI Module (AIM) | A processing element receiving AIM-specific Inputs and producing AIM-specific Outputs according to according to its Function. An AIM may be an aggregation of AIMs. |
| AI Workflow (AIW) | A structured aggregation of AIMs implementing a Use Case receiving AIW-specific inputs and producing AIW-specific inputs according to its Function. |
| AIF Metadata | The data set describing the capabilities of an AIF set by the AIF Implementer. |
| AIM Metadata | The data set describing the capabilities of an AIM set by the AIM Implementer. |
| Application Programming Interface (API) | A software interface that allows two applications to talk to each other |
| Application Standard | An MPAI Standard specifying AIWs, AIMs, Topologies and Formats suitable for a particular application domain. |
| Channel | A physical or logical connection between an output Port of an AIM and an input Port of an AIM. The term “connection” is also used as a synonym. |
| Communication | The infrastructure that implements message passing between AIMs. |
| Component | One of the 9 AIF elements: Access, AI Module, AI Workflow, Communication, Controller, Internal Storage, Global Storage, MPAI Store, and User Agent. |
| Composite AIM | An AIM that includes AIMs and their Connections and Topology. |
| Conformance | The attribute of an Implementation of being a correct technical Implementation of a Technical Specification. |
| Conformance Tester | An entity authorised by MPAI to Test the Conformance of an Implementation. |
| Conformance Testing | The normative document specifying the Means to Test the Conformance of an Implementation. |
| Conformance Testing Means | Procedures, tools, data sets and/or data set characteristics to Test the Conformance of an Implementation. |
| Connection | A channel connecting an output port of an AIM and an input port of an AIM. |
| Controller | A Component that manages and controls the AIMs in the AIF, so that they execute in the correct order and at the time when they are needed. |
| Data | Information in digital form. |
| Data Format | The standard digital representation of Data. |
| Data Semantics | The meaning of Data. |
| Device | A hardware and/or software entity running at least one instance of an AIF. |
| Ecosystem | The ensemble of the following actors: MPAI, MPAI Store, Implementers, Conformance Testers, Performance Testers and Users of MPAI-AIF Implementations as needed to enable an Interoperability Level. |
| Event | An occurrence acted on by an Implementation. |
| Explainability | The ability to trace the output of an Implementation back to the inputs that have produced it. |
| Fairness | The attribute of an Implementation whose extent of applicability can be assessed by making the training set and/or network open to testing for bias and unanticipated results. |
| Function | The operations effected by an AIW or an AIM on input data. |
| Global Storage | A Component to store data shared by AIMs. |
| Identifier | A name that uniquely identifies an Implementation. |
| Implementation | 1. An embodiment of the MPAI-AIF Technical Specification, or 2. An AIW or AIM of a particular Level (1-2-3). |
| Internal Storage | A Component to store data of the individual AIMs. |
| Interoperability | The ability to functionally replace an AIM/AIW with another AIM/AIW having the same Interoperability Level |
| Interoperability Level | The attribute of an AIW and its AIMs to be executable in an AIF Implementation and to be:   1. Implementer-specific and satisfying the MPAI-AIF Standard *(Level 1)*. 2. Specified by an MPAI Application Standard (*Level 2)*. 3. Specified by an MPAI Application Standard and certified by a Performance Assessor (*Level 3)*. |
| Knowledge Base | Structured and/or unstructured information made accessible to AIMs via MPAI-specified interfaces |
| Message | A sequence of Records. |
| Normativity | The set of attributes of a technology or a set of technologies specified by the applicable parts of an MPAI standard. |
| Performance | The attribute of an Implementation of being Reliable, Robust, Fair and Replicable. |
| Performance Assessment | The normative document specifying the procedures, the tools, the data sets and/or the data set characteristics to Assess the Grade of Performance of an Implementation. |
| Performance Assessment Means | Procedures, tools, data sets and/or data set characteristics to Assess the Performance of an Implementation. |
| Performance Assessor | An entity authorised by MPAI to Assess the Performance of an Implementation in a given Application domain |
| Port | A physical or logical communication interface of an AIM. |
| Profile | A particular subset of the technologies used in MPAI-AIF or an AIW of an Application Standard and, where applicable, the classes, other subsets, options and parameters relevant to that subset. |
| Record | Data with a specified structure. |
| Reference Model | The AIMs and theirs Connections in an AIW. |
| Reference Software | A technically correct software implementation of a Technical Specification containing source code, or source and compiled code. |
| Reliability | The attribute of an Implementation that performs as specified by the Application Standard, profile and version the Implementation refers to, e.g., within the application scope, stated limitations, and for the period of time specified by the Implementer. |
| Replicability | The attribute of an Implementation whose Performance, as Assessed by a Performance Assessor, can be replicated, within an agreed level, by another Performance Assessor. |
| Robustness | The attribute of an Implementation that copes with data outside of the stated application scope with an estimated degree of confidence. |
| Scope | The domain of applicability of an MPAI Application Standard. |
| Service Provider | An entrepreneur who offers an Implementation as a service (e.g., a recommendation service) to Users. |
| Specification | A collection of normative clauses. |
| Standard | The ensemble of Technical Specification, Reference Software, Conformance Testing and Performance Assessment of an MPAI application Standard. |
| Technical Specification | (Framework) the normative specification of the AIF.  (Application) the normative specification of the set of AIWs belon­ging to an application domain along with the AIMs required to Im­plement the AIWs that includes:   1. The formats of the Input/Output data of the AIWs implementing the AIWs. 2. The Connections of the AIMs of the AIW. 3. The formats of the Input/Output data of the AIMs belonging to the AIW. |
| Testing Laboratory | A laboratory accredited by MPAI to Assess the Grade of Performance of Implementations. |
| Time Base | The protocol specifying how AIF Components can access timing information. |
| Topology | The set of AIM Connections of an AIW. |
| Use Case | A particular instance of the Application domain target of an Application Standard. |
| User | A user of an Implementation. |
| User Agent | The Component interfacing the user with an AIF through the Controller |
| Version | A revision or extension of a Standard or of one of its elements. |
| Zero Trust | A cybersecurity model primarily focused on data and service protection that assumes no implicit trust. |

1. - Notices and Disclaimers Concerning MPAI Standards (Informative)

The notices and legal disclaimers given below shall be borne in mind when [downloading](https://www.mpai.community/resources/) and using approved MPAI Standards.

In the following, “Standard” means the collection of four MPAI-approved and [published](https://www.mpai.community/resources/) documents: “Technical Specification”, “Reference Software” and “Conformance Testing” and, where applicable, “Performance Testing”.

Life cycle of MPAI Standards

MPAI Standards are developed in accordance with the [MPAI Statutes](https://mpai.community/statutes/). An MPAI Standard may only be developed when a Framework Licence has been adopted. MPAI Standards are developed by especially established MPAI Development Committees who operate on the basis of consensus, as specified in Annex 1 of the [MPAI Statutes](https://mpai.community/statutes/). While the MPAI General Assembly and the Board of Directors administer the process of the said Annex 1, MPAI does not independently evaluate, test, or verify the accuracy of any of the information or the suitability of any of the technology choices made in its Standards.

MPAI Standards may be modified at any time by corrigenda or new editions. A new edition, however, may not necessarily replace an existing MPAI standard. Visit the [web page](https://mpai.community/resources/) to determine the status of any given published MPAI Standard.

Comments on MPAI Standards are welcome from any interested parties, whether MPAI members or not. Comments shall mandatorily include the name and the version of the MPAI Standard and, if applicable, the specific page or line the comment applies to. Comments should be sent to the [MPAI Secretariat](mailto:secretariat@mpai.community). Comments will be reviewed by the appropriate committee for their technical relevance. However, MPAI does not provide interpretation, consulting information, or advice on MPAI Standards. Interested parties are invited to join MPAI so that they can attend the relevant Development Committees.

Coverage and Applicability of MPAI Standards

MPAI makes no warranties or representations of any kind concerning its Standards, and expressly disclaims all warranties, expressed or implied, concerning any of its Standards, including but not limited to the warranties of merchantability, fitness for a particular purpose, non-infringement etc. MPAI Standards are supplied “AS IS”.

The existence of an MPAI Standard does not imply that there are no other ways to produce and distribute products and services in the scope of the Standard. Technical progress may render the technologies included in the MPAI Standard obsolete by the time the Standard is used, especially in a field as dynamic as AI. Therefore, those looking for standards in the Data Compression by Artificial Intelligence area should carefully assess the suitability of MPAI Standards for their needs.

IN NO EVENT SHALL MPAI BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: THE NEED TO PROCURE SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE PUBLICATION, USE OF, OR RELIANCE UPON ANY STANDARD, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

MPAI alerts users that practicing its Standards may infringe patents and other rights of third parties. Submitters of technologies to this standard have agreed to licence their Intellectual Property according to their respective Framework Licences.

Users of MPAI Standards should consider all applicable laws and regulations when using an MPAI Standard. The validity of Conformance Testing is strictly technical and refers to the correct implementation of the MPAI Standard. Moreover, positive Performance Assessment of an implementation applies exclusively in the context of the [MPAI Governance](https://mpai.community/standards/mpai-gme/) and does not imply compliance with any regulatory requirements in the context of any jurisdiction. Therefore, it is the responsibility of the MPAI Standard implementer to observe or refer to the applicable regulatory requirements. By publishing an MPAI Standard, MPAI does not intend to promote actions that are not in compliance with applicable laws, and the Standard shall not be construed as doing so. In particular, users should evaluate MPAI Standards from the viewpoint of data privacy and data ownership in the context of their jurisdictions.

Implementers and users of MPAI Standards documents are responsible for determining and complying with all appropriate safety, security, environmental and health and all applicable laws and regulations.

Copyright

MPAI draft and approved standards, whether they are in the form of documents or as web pages or otherwise, are copyrighted by MPAI under Swiss and international copyright laws. MPAI Standards are made available and may be used for a wide variety of public and private uses, e.g., implementation, use and reference, in laws and regulations and standardisation. By making these documents available for these and other uses, however, MPAI does not waive any rights in copyright to its Standards. For inquiries regarding the copyright of MPAI standards, please contact the [MPAI Secretariat](mailto:secretariat@mpai.community).

The Reference Software of an MPAI Standard is released with the [MPAI Modified Berkeley Software Distribution licence](https://mpai.community/about/licence/). However, implementers should be aware that the Reference Software of an MPAI Standard may reference some third party software that may have a different licence.

1. – The Governance of the MPAI Ecosystem (Informative)

**Level 1 Interoperability**

With reference to Figure 1, MPAI issues and maintains a standard – called MPAI-AIF – whose components are:

1. An environment called AI Framework (AIF) running AI Workflows (AIW) composed of inter­connected AI Modules (AIM) exposing standard interfaces.
2. A distribution system of AIW and AIM Implementation called MPAI Store from which an AIF Implementation can download AIWs and AIMs.

A Level 1 Implementation shall implement the MPAI-AIF Technical Specification executing AIWs composed of AIMs able to call the MPAI-AIF APIs.

|  |  |
| --- | --- |
| Implementers’ benefits | Upload to the MPAI Store and have globally distributed Implementations of   * AIFs conforming to MPAI-AIF. * AIWs and AIMs performing prop­rietary functions executable in AIF. |
| Users’ benefits | Rely on Implementations that have been tested for security. |
| MPAI Store’s role | * Tests the Conformance of Implementations to MPAI-AIF. * Verifies Implementations’ security, e.g., absence of malware. * Indicates unambiguously that Implementations are Level 1. |

**Level 2 Interoperability**

In a Level 2 Implementation, the AIW must be an Implementation of an MPAI Use Case and the AIMs must conform with an MPAI Application Standard.

|  |  |
| --- | --- |
| Implementers’ benefits | Upload to the MPAI Store and have globally distributed Implementations of:   * AIFs conforming to MPAI-AIF. * AIWs and AIMs conforming to MPAI Application Standards. |
| Users’ benefits | * Rely on Implementations of AIWs and AIMs whose Functions have been reviewed during standardisation. * Have a degree of Explainability of the AIW operation because the AIM Functions and the data Formats are known. |
| Market’s benefits | * Open AIW and AIM markets foster competition leading to better products. * Competition of AIW and AIM Implementations fosters AI innovation. |
| MPAI Store’s role | * Tests Conformance of Implementations with the relevant MPAI Standard. * Verifies Implementations’ security. * Indicates unambiguously that Implementations are Level 2. |

**Level 3 Interoperability**

MPAI does not generally set standards on how and with what data an AIM should be trained. This is an important differentiator that promotes competition leading to better solutions. However, the performance of an AIM is typically higher if the data used for training are in greater quantity and more in tune with the scope. Training data that have large variety and cover the spectrum of all cases of interest in breadth and depth typically lead to Implementations of higher “quality”.

For Level 3, MPAI normatively specifies the process, the tools and the data or the characteristics of the data to be used to Assess the Grade of Performance of an AIM or an AIW.

|  |  |
| --- | --- |
| Implementers’ benefits | May claim their Implementations have passed Performance Assessment. |
| Users’ benefits | Get assurance that the Implementation being used performs correctly, e.g., it has been properly trained. |
| Market’s benefits | Implementations’ Performance Grades stimulate the development of more Performing AIM and AIW Implementations. |
| MPAI Store’s role | * Verifies the Implementations’ security * Indicates unambiguously that Implementations are Level 3. |

**The MPAI ecosystem**

The following is a high-level description of the MPAI ecosystem operation applicable to fully conforming MPAI implementations:

1. MPAI establishes and controls the not-for-profit MPAI Store.
2. MPAI appoints Performance Assessors.
3. MPAI publishes Standards.
4. Implementers submit Implementations to Performance Assessors.
5. If the Implementation Performance is acceptable, Performance Assessors inform Implementers and MPAI Store.
6. Implementers submit Implementations to the MPAI Store tested for Conformance and security.
7. Users download and use Implementations, and submit experience scores.



*Figure 12 – The MPAI ecosystem operation*

1. Personal Status is defined as a set of internal characteristics of a person or avatar. MPAI currently considers Emotion, Cognitive State, and Attitude. It can be conveyed by the following Modalities: Text, Speech, face and Gesture. Other Modalities are possible but currently not considered. [↑](#footnote-ref-1)