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

|  |  |
| --- | --- |
| **N646** | 2022/04/20 |
| **Source** | MMC-DC |
| **Title** | MPAI-MMC Use Cases and Functional Requirements V2 WD1.3 |
| **Target** | MPAI-19 |

Contents

[Abstract 2](#_Toc101376736)

[1 Introduction 2](#_Toc101376737)

[2 Scope of the Use Cases 4](#_Toc101376738)

[3 Terms and Definitions 5](#_Toc101376739)

[4 References 6](#_Toc101376740)

[4.1 Normative References 6](#_Toc101376741)

[4.2 Informative References 7](#_Toc101376742)

[5 Use Case Architectures 8](#_Toc101376743)

[5.1 Conversation About a Scene 8](#_Toc101376744)

[5.1.1 Scope of Use Case 8](#_Toc101376745)

[5.1.2 Reference architecture 8](#_Toc101376746)

[5.1.3 Input/output data 9](#_Toc101376747)

[5.1.4 AI Modules 9](#_Toc101376748)

[5.2 Human-CAV Interaction AIW (HCI) 10](#_Toc101376749)

[5.2.1 Scope of Use Case 10](#_Toc101376750)

[5.2.2 Reference architecture 11](#_Toc101376751)

[5.2.3 Input and output data 12](#_Toc101376752)

[5.2.4 AI Modules 12](#_Toc101376753)

[5.3 Avatar Videoconference 13](#_Toc101376754)

[5.3.1 Scope of Use Case 13](#_Toc101376755)

[5.3.2 TX Client AIW 14](#_Toc101376756)

[5.3.3 Server AIW 17](#_Toc101376757)

[5.3.4 RX Client AIW 19](#_Toc101376758)

[6 Functional Requirements 20](#_Toc101376759)

[6.1 I/O Data summary 20](#_Toc101376760)

[6.1.1 Conversation About a Scene 20](#_Toc101376761)

[6.1.2 Human-CAV Interaction 21](#_Toc101376762)

[6.1.3 Avatar Videoconference 21](#_Toc101376763)

[6.2 Data format requirements 23](#_Toc101376764)

[6.2.1 Ambient Descriptors 23](#_Toc101376765)

[6.2.2 Ambient Model 23](#_Toc101376766)

[6.2.3 Audio 23](#_Toc101376767)

[6.2.4 Audio Scene Descripors 23](#_Toc101376768)

[6.2.5 Audio-Visual Scene Descriptors 24](#_Toc101376769)

[6.2.6 Avatar Descriptors 24](#_Toc101376770)

[6.2.7 Avatar Model 24](#_Toc101376771)

[6.2.8 Emotion (Face) 25](#_Toc101376772)

[6.2.9 Emotion (Speech) 25](#_Toc101376773)

[6.2.10 Emotion (Text) 25](#_Toc101376774)

[6.2.11 Expression (Face) 26](#_Toc101376775)

[6.2.12 Face Objects 26](#_Toc101376776)

[6.2.13 Fused Emotion 26](#_Toc101376777)

[6.2.14 Head and Face Descriptors 26](#_Toc101376778)

[6.2.15 Head Descriptors 27](#_Toc101376779)

[6.2.16 Input Audio 27](#_Toc101376780)

[6.2.17 Input Human Video 27](#_Toc101376781)

[6.2.18 Input Speech 27](#_Toc101376782)

[6.2.19 Input Video 27](#_Toc101376783)

[6.2.20 Intention 28](#_Toc101376784)

[6.2.21 Language Preferences 28](#_Toc101376785)

[6.2.22 LangUnd Text 28](#_Toc101376786)

[6.2.23 Meaning 28](#_Toc101376787)

[6.2.24 Microphone Array Audio 28](#_Toc101376788)

[6.2.25 Object Identifier 29](#_Toc101376789)

[6.2.26 Output Speech 29](#_Toc101376790)

[6.2.27 Physical Gesture Descriptors 29](#_Toc101376791)

[6.2.28 Ranked Text 29](#_Toc101376792)

[6.2.29 Recognised Text 30](#_Toc101376793)

[6.2.30 Scene Descriptor Video 30](#_Toc101376794)

[6.2.31 Spatial coordinates 30](#_Toc101376795)

[6.2.32 Text 30](#_Toc101376796)

[6.2.33 Verbal Human-CAV Interaction 30](#_Toc101376797)

[6.2.34 Video 31](#_Toc101376798)

[6.2.35 Viewpoint 31](#_Toc101376799)

[6.2.36 Visual Scene Descriptors 31](#_Toc101376800)

[Annex 1 – MPAI-wide terms and definitions (Normative) 33](#_Toc101376801)

[Annex 2 - Notices and Disclaimers Concerning MPAI Standards (Informative) 36](#_Toc101376802)

[Annex 3 – The Governance of the MPAI Ecosystem (Informative) 38](#_Toc101376803)

[Annex 2 - Notices and Disclaimers Concerning MPAI Standards (In-formative) 41](#_Toc101376804)

[Annex 3 – The Governance of the MPAI Ecosystem (Informative) 43](#_Toc101376805)

# Abstract

This is a Collection of Use Cases and Functional Requirements being considered for Version 2 of the MPAI-MMC standard. It is a Working Draft expected to be changed in some parts.

# Introduction

In recent years, Artificial Intelligence (AI) and related technologies have been applied to a broad range of applications, have started affecting the life of millions of people and are expected to do so even more in the future. As digital media standards have positively influenced industry and billions of people, so AI-based data coding standards are expected to have a similar positive im-pact. Indeed, research has shown that data coding with AI-based technologies is generally more efficient than with 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 clas-ses of users. Therefore, the need for standardisation is more important and urgent than ever.

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. MPAI Application Standards enable the development of AI-based products, applications and services.

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

For the aforementioned reasons, to fully achieve the MPAI mission, Technical Specifications must be complemented by an ecosystem designed, created and managed to underpin the life cy-cle of MPAI standards through the steps of specification, technical testing, assessment of prod-uct 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 17* if they are common to all MPAI Standards.

The MPAI Ecosystem is specified in [1]. It is composed of:

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

The common infrastructure enabling implementation of MPAI Application Standards and access to the MPAI Store is the AI Framework (AIF) Standard (MPAI-AIF), specified in this document.

Figure 1 depicts the MPAI-AIF Reference Model under which Implementations of MPAI Appli-cation Standards and user-defined MPAI-AIF conforming applications operate.

An AIF Implementation allows execution of AI Workflows (AIW), composed of basic pro-cessing elements called AI Modules (AIM).

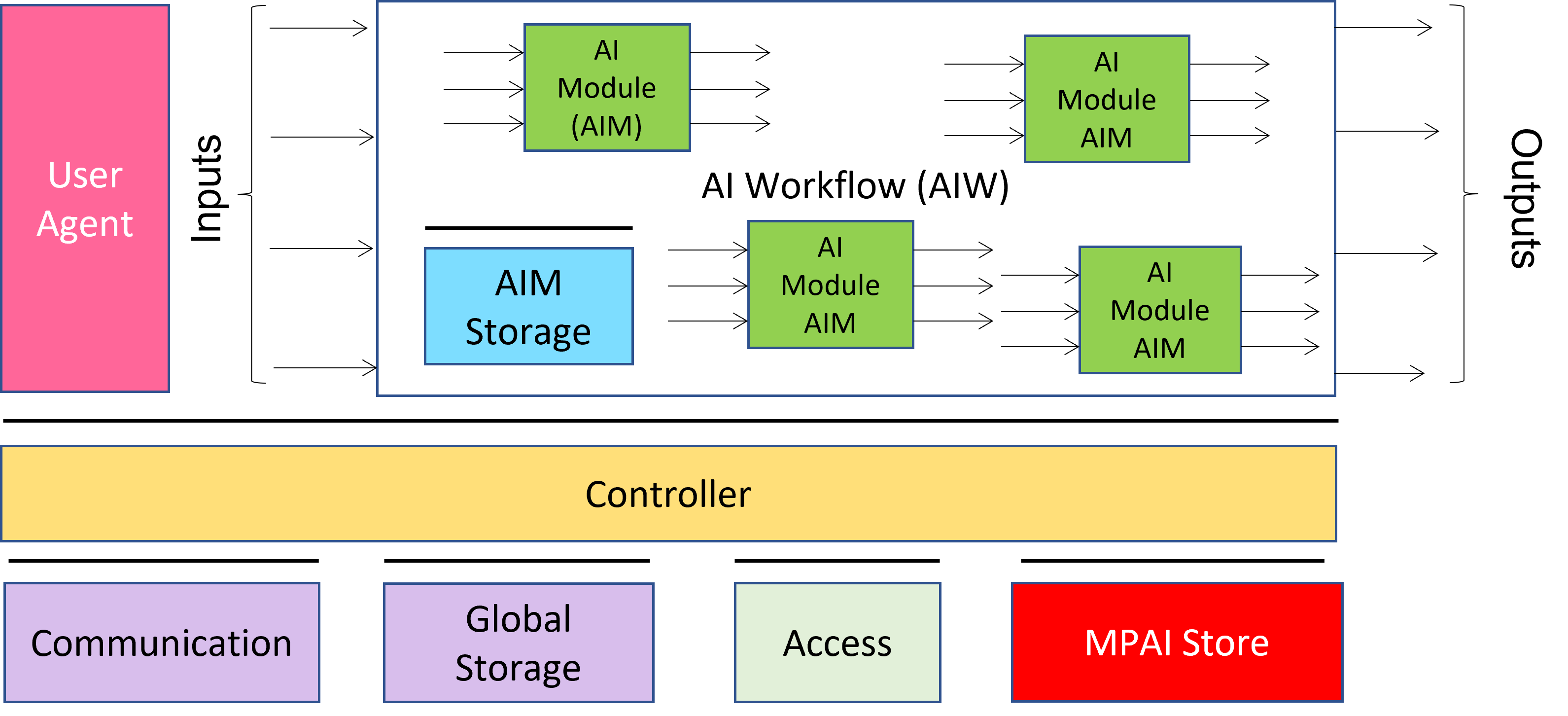


Figure 1 - The AI Framework (AIF) Reference Model and its Components

MPAI Application Standards normatively specify Syntax and Semantics of the input and output data and the Function of the AIW and the AIMs, and the Connections between and among the AIMs of an AIW.

In particular, an AIM is defined by its Function and data, but not by its internal architecture, which may be based on AI or data processing, and implemented in software, hardware or hybrid software and hardware technologies.

MPAI defines Interoperability as the ability to replace an AIW or an AIM Implementation with a functionally equivalent Implementation. MPAI also defines 3 Interoperability Levels of an AIW executed in an AIF:

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

*Level 2 –* Specified by an MPAI Application Standard.

*Level 3 –* Specified by 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 as the Interoperability Level of an Implementation moves from 1 to 3.

Additional information on Interoperability Levels is provided in Annex 3.

# Scope of the Use Cases

The Use Cases of MPAI-MMC V2 add the following functionalities on top of the MPAI-MMC V1 technologies:

1. The user selects:
   1. The Ambient in which the avatars operate.
   2. The avatar model used by the machine to manifest itself.
   3. The Colour (i.e., the characteristics of a speaker) the machine will use to utter speech.
2. The machine locates the visual and speech components of human(s) in the visual and sound space.
3. The machine separates (diarisation):
   1. The visual components of the individual humans from the rest of the visual space (i.e., other visual objects and other visual humans).
   2. The speech components of the individual speaking humans from the rest of sound space (i.e., other sound objects).
4. The machine extracts descriptors of:
   1. Human face.
   2. Physical Gesture (i.e., head, arms, hands and fingers).
   3. Human speech.
5. The machine uses
   1. Face Descriptors to:
      1. Identify the human belonging to a group of a limited number of humans (closed set identification).
      2. Extract the Expression of the face.
      3. Animate the face of an avatar.
   2. Physical Gesture Descriptors to:
      1. Extract the Expression of the Physical Gestures.
      2. Interpret the sign language conveyed by the Physical Gestures (this is culture and language dependent).
      3. Animate the upper part of the body (torso) of an avatar using the required descriptors:
         1. Extracted from a a human.
         2. Synthetically generated.
   3. Speech Descriptos to:
      1. Identify a human belonging to a group composed of a limited number of humans (closed set identification).
      2. Recognise speech (i.e., extract Text).
      3. Extract the Emotion in the speech.
      4. Extract the Colour.
6. The machine holds a conversation with a human or an avatar (other party)
   1. In the context of a specific domain (CAV).
   2. About objects in the visual space indicated by a Physical Gesture.

by:

1. Analysing and interpreting their Expression, Gesture and Emotion.
2. Uttering speech with Emotion, possibly spatially located on the lips of an avatar.
3. Animating:
   * 1. Eyes, lips and facial muscles of a selected avatar to display an Expression.
     2. Lips in sync with an uttered speech.
4. Expressing a sequence of Emotions and displaying a sequence of Expressions that are congruent with:
   * 1. The Text, Expressions and Emotion of the other party, e.g., it expresses satisfaction as it has understood the meaning of the question or sentence.
     2. The machine’s response and its associated Expressions and Emotions.
5. Gazing at the other party it is conversing with.

# 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*

|  |  |  |
| --- | --- | --- |
| **Use Case** | **Term** | **Definition** |
| CAS | Expression (Face) | Attributes of a human face, such as movement of face muscles, head, arms, hands, and fingers reflecting a particular state of mind |
| CAS | Expression (Speech) | Attributes of a speech either inborn or reflecting a particular state of mind (e.g., falsetto, rough). |
| CAS | Descriptor |  |
| CAS | Gesture | A movement of the body or part of it, such as arm, hand and fin-ger, often as a complement to a vocal utterance. |
| CAS | Scene | An environment populated by a human and scattered objects. |
| HCI | Basic World Representation | A description of the Environment based on the CAV sensors and Offline Maps. |
| HCI | Command | High-level instructions whose execution allows a CAV to reach a Goal. |
| HCI | Connected Au-tonomous Vehicle | A vehicle capable to autonomously reach an assigned target by understanding human utterances, planning a route, sensing and interpreting the environment, exchanging information with other CAVs and acting on the CAV’s motion subsystem. |
| HCI | Emotion | An attribute that indicates an emotion out of a finite set of Emotions |
| HCI | Grade | The intensity of an Emotion |
| HCI | Expression (Face) | Attributes of a human face, such as movement of face muscles, head, arms, hands, and fingers reflecting a particular state of mind. |
| HCI | Expression (Speech) | Attributes of a speech either inborn (e.g., falsetto, rough) or reflecting a particular state of mind. |
| HCI | Intention | The result of a question analysis that denotes information on the input question. |
| HCI | Meaning | Information extracted from an input text such as syntactic and se-mantic information. |
| HCI | Physical Gesture | Movement of head, arms, hands and fingers. |
| HCI | Subword Lattice | A directed graph containing speech recognition candidates |
| HCI | Text | A collection of characters drawn from a finite alphabet. |
| HCI | Vocal Gesture | Utterances generally avoiding lexical elements (e.g., cough, laugh, hesitation etc.). |
| HCI | Word Lattice | A directed graph containing speech recognition candidates |
| MCS | Affordance | Quality or property of an object that defines its possible uses or makes clear how it can or should be used. |
| MCS | Ambient | The physical space of a participant and the shared virtual space. |
| MCS | 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. |
| MCS | Audio Object | Audio information with its metadata |
| MCS | Audio Scene | The Audio component of an AV Scene. |
| MCS | Audio-Visual Scene | (AV Scene) The structured composition of Audio and Visual Objects in an Ambient. |
| MCS | Avatar | An animated 3D object representing a particular person in a virtual space. |
| MCS | Avatar Model | An inanimate avatar |
| MCS | Emotion | An attribute that indicates an emotion out of a finite set of Emotions |
| MCS | Face | The portion of a video containing the face of a human. |
| MCS | Physical Gesture | The coordinated movement of hands and possibly arms |
| MCS | Identity | The label associated to a human in a unique way. |
| MCS | Input Audio | Audio entering an MPAI-MCS Client (TX). |
| MCS | Intention | The result of analysis of the goal of an input question. |
| MCS | Meaning | Information extracted from the input such as syntactic and semantic information. |
| MCS | Separated Speech | Speech extracted from Input Audio. |
| MCS | Speech | Digital representation of analogue speech sampled at a frequency between 8 kHz and 96 kHz with a number of bits between 16 bits/sample and 24 bits/sample PCM values. |
| MCS | Visual Object | Visual information with its metadata. |
| MCS | Visual Scene | The Visual component of an AV Scene. |
| MCS | Viewpoint | The audio and visual point in the MCS selected by a human, not necessarily coninciding with the location of avatar |

# References

## Normative References

This document references the following normative documents:

1. Technical Specification: The Governance of the MPAI Ecosystem V1.
2. Technical Specification: AI Framework (MPAI-AIF), MPAI document N359.
3. Technical Specification: Technical Specification: Multimodal Conversation (MPAI-MMC) V1; published at https://mpai.community/standards/resources/1.
4. Technical Specification: Context-based Audio Enhancement (MPAI-CAE) V1.1; published at https://mpai.community/standards/resources/.
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: Ad-vanced 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.

## Informative References

CAV

1. MPAI-CAV Use Cases and Functional Requirements WD0.10, N604
2. An-example-word-lattice; https://www.researchgate.net/figure/An-example-word-lattice\_fig1\_2361715

Ambient

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’s glTF: glTF defines an extensible, transmission-efficient publishing format that streamlines authoring workflows and interactive services by enabling the interoperable use of 3D content across the industry.
4. Pixar's Universal Scene Description: USD is an easily extensible, open-source 3D scene de-scription and file format developed by Pixar for content creation and interchange among dif-ferent tools. As a result of its power and versatility, it’s being widely adopted, not only in the visual effects community, but also in architecture, design, robotics, manufacturing, and other disciplines.
5. Blender; www.blender.org

Face description

1. FACS-based Facial Expression Animation in Unity
2. Facial Action Coding System
3. Facial Action Coding System
4. https://www.noldus.com/applications/facial-action-coding-system
5. https://3d.kalidoface.com/
6. List of motion and gesture file formats
7. https://docs.readyplayer.me/
8. 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. https://docs.unity3d.com/Manual/AmbisonicAudio.html

Presentation/rendering

1. Unity
2. WebXR
3. UnrealEngine
4. Unreal Engine – Audio to Facial Animation SDK
5. https://developer.nvidia.com/nvidia-omniverse-platform
6. Microsoft Mesh

# Use Case Architectures

## Conversation About a Scene

### Scope of Use Case

The human holds a conversation with the machine:

1. The machine sees a single scene containing the human and some scattered objects.
2. The human uses Physical gesture to indicate an object s/he wishes to talk about.
3. The machine gets the human’s Expression, Gesture and Emotion via Video (Face and Gesture) and Audio (Speech).
4. The machine creates an internal representation of the scene and uses it to create a synthetic version of the scene to help the human understand if the machine has interpreted the scene correctly.
5. The machine understands which object the Human is referring to in his/her conversation (because s/he point his/her arm/hand/finger at an object) and makes a pertinent response uttered via synthetic speech accompanied by its avatar face.

### Reference architecture

The Machine operates according to the following workflow:

1. Speech Recognition produces Text and Emotion (Speech) from Input Speech.
2. Language Understanding extracts Emotion (Text) from Text.
3. Video Analysis1 extracts Emotion (Video) from Input Video.
4. Fuses the Emotion (Text), Emotion (Speech) and Emotion (Video).
5. Language Understanding Extracts Meaning from Text.
6. Question Analysis produces Intention from Meaning.
7. Video Analysis3 creates Scene Description from Input Video.
8. Video Analysis 2 creates Gesture Description from Scene Description.
9. Object Identification produces Object ID by combining Scene Description and Gesture Description.
10. Scene2Video produces the 2D video of the Scene Description where the objects are labelled with the machine’s understanding of the objects.
11. Dialogue Processing produces the machine’s Expression (Speech) and Expression (Face) from Fused Emotion, Meaning LangUnd Text and Intention.
12. Speech Synthesis (Expression) produces Output Speech from Expression (Speech).
13. Face Animation (Expression) produces Output Face (Face).

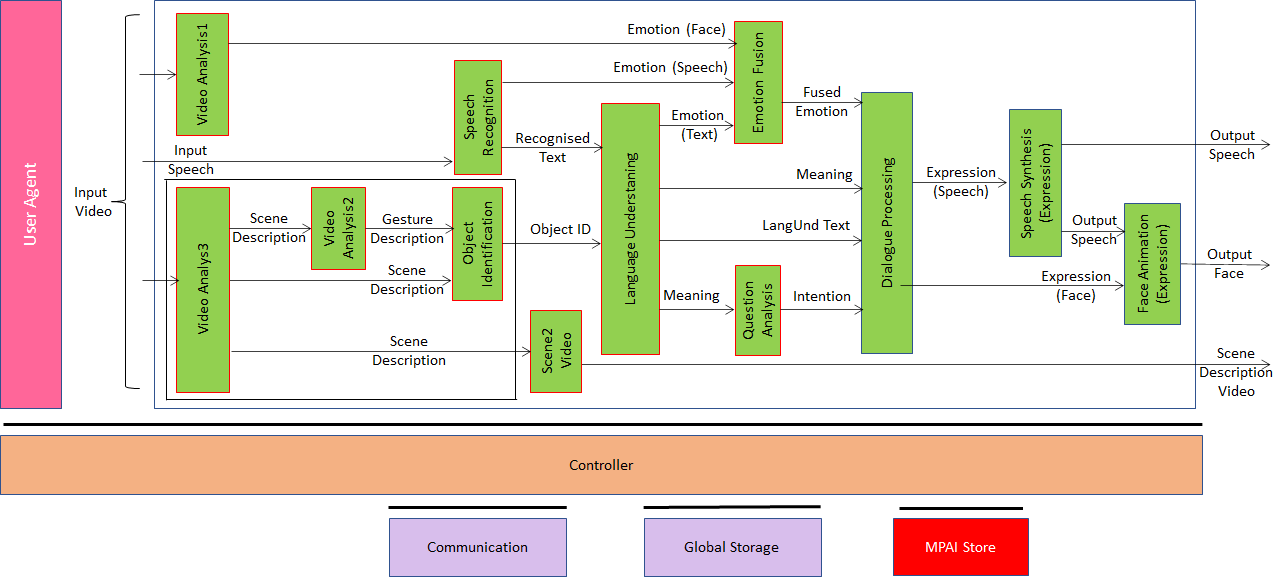


Figure 2 - Conversation About a Scene Reference Model

### Input/output data

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

*Table 2 – 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 |
| **Output data** | **To** | **Comments** |
| Output Speech | Human | Machine’s speech |
| Output Face | Human | Machine’s face |
| Scene Description Video | Human | Video reproduction of scene as seen by machine |

### AI Modules

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

*Table 3 – AI Modules of Conversation About a Scene*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Video Analysis1 | Receives Input Video  Extracts Emotion (Video) |
| Video Analysis2 | Receives Scene Descriptors  Extracts Gesture Descriptors |
| Video Analysis3 | Receives Input Video  Creates Scene Descriptors |
| Speech Recognition | Receives Input Speech  Extracts Recognised Text from Input Speech  Extracts Emotion (Speech) from Input Speech |
| Object Identification | Receives Gesture Description and Scene Description  Produces ID of scene object the human points at |
| Scene2Video | Receives Scene Descriptors (SD)  Converts Scene Descriptors to a Video of the Scene from a selectable viewpoint compatible with Input Video. |
| Language Understanding | Receives Recognised Text and Object  Produces Emotion (Text), Meaning (Text), and LangUnd Text |
| Emotion Fusion | Receives Emotion (Text), Emotion (Speech) and Emotion (Video)  Produces a single Emotion that fuses the 3 Emotions |
| Question Analysis | Receives Meaning  Produces Intention |
| Dialogue Processing | Receives Meaning, Fused Emotion, Land Und Text and Intention  Produces Expression (Text) and Expression (Speech) |
| Speech Synthesis (Expression) | Receives Text with Emotion  Produces Output Speech |
| Lips Animation | Receives Output Speech, Avatar Emotion  Produces Output Face |

## Human-CAV Interaction AIW (HCI)

### Scope of Use Case

This use case is part of the Connected Autonomous Vehicle (CAV) project [10]. A CAV is a system capable of executing a command to move based on 1) an analysis of the data sensed by a range of sensors exploring the environment and 2) information transmitted by other sources in range, e.g., other CAVs and roadside units.

Human-CAV Interaction is based on the principle that the CAV is impersonated by an avatar, se-lected/produced by the CAV rights-holder (owner or renter). The visible features of the avatar are head, face, and torso, and the audible feature is speech, which embeds as much as possible the sentiment, e.g., Emotion and Expression that would be displayed by a human driver in similar conditions.

These are the functions of an HCI considered in this use case:

1. The HCI separates the visual and speech part of the humans from the rest of the environment which can be:
   1. The space close to the CAV (the human gives instructions to the CAV).
   2. The cabin of the CAV (the human holds a conversation with the CAV).
2. The CAV authenticates the humans it is interacting with by recognising the human CAV rights holder (owner or renter) using speech and face information.
3. The HCI separates the speech signals uttered by the humans from the rest of the sound in the cabin and determines the position of the speech-uttering humans.
4. 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, etc.
      2. Other commands, e.g., turn off air conditioning, turn on radio, call a person, open window or door, search for information etc.
      3. (Figure 3 includes the conversion of human commands and CAV response. However, this document does not address the format in which the HCI interacts with the AMS).
   2. By talking and responding to questions from one or more humans about travel-related issues (in-depth domain-specific conversation), e.g.:
      1. CAV offers alternatives to humans, e.g., long but safe way, short but likely to have interruptions.
      2. Humans requestsinformation, e.g., time to destination, route conditions, weather at destination, etc.
5. The HCI senses the humans’ speech, face and gesture and manifests itself as face and shoulder of an avatar capable of
   1. Displaying Expressions in the face.
   2. Animating lips in sync with the speech and in line with the Expression.
   3. Gazing at the human it is conversing with.
   4. Uttering speech with appropriate Emotion.
6. Assuming that the passengers allow the CAV to do so, and the processing is carried out inside the CAV, the HCI follows conversation on travel matters held by humans in the cabin during the travel.
7. Issue regarding the social participation in human activities:
   1. Under what circumstances can the HCI take the initiative in starting and participating in the conversation?

### Reference architecture

Figure 3 represents the Human-CAV Interaction (HCI) Reference Model.

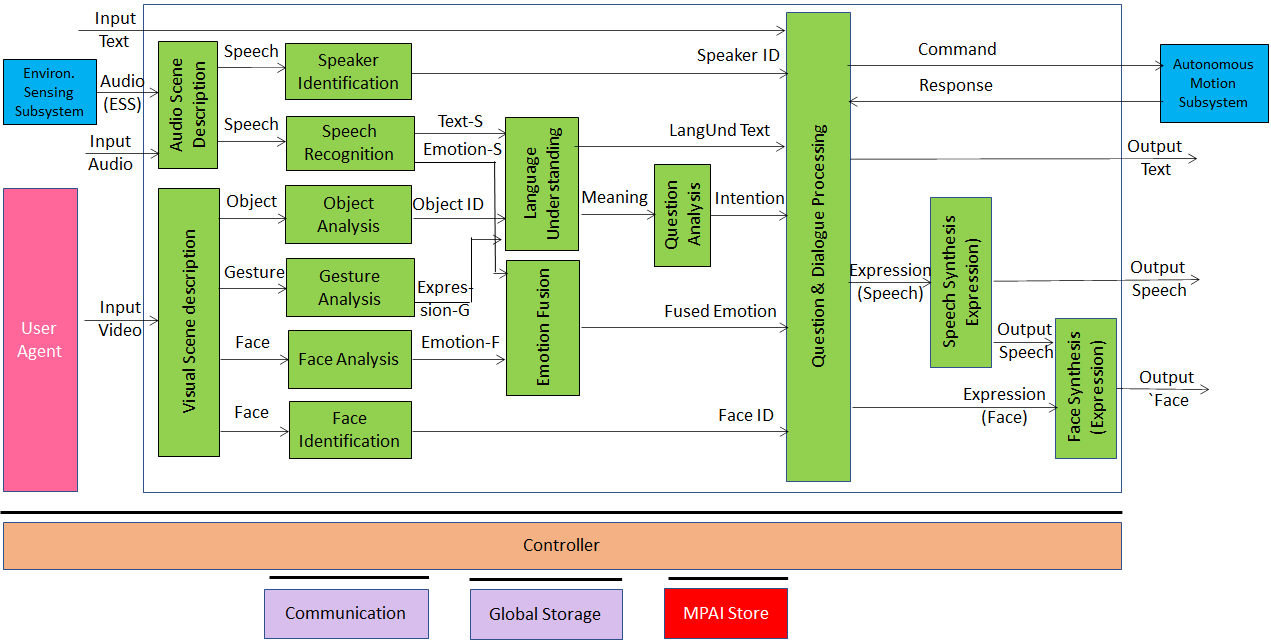


Figure 3 - Human-CAV Interaction Reference Model

HCI operates in two modes:

1. Outdoor: when humans are in the environment, approach the CAV, and:
   1. HCI separates and locates
      1. The human faces from the environment and then identifies them.
      2. The human speech from the environment sound and then recognises the humans for identification.
   2. HCI converses with humans.
2. In the cabin:
   1. HCI separates the speech source from other sounds in the cabin and locates them.
   2. HCI locates the visual elements of humans in the cabin.
   3. HCI recognises the humans for identification.
   4. HCI recognises human gestures.
   5. HCI converses with humans in the cabin.

When conversing with the humans in the cabin, the CAV recognises

1. Emotion in human speech.
2. Expression in face.
3. Gesture,

and responds with an avatar having:

1. Speech containing emotion appropriate to the conversation.
2. Face having
3. Lips moving synchronously with the speech.
4. Face Expression consistent with the Emotion in the CAV speech.
5. Eyes gazing at the relevant human.

### Input and output data

*Table 4* gives the input/output data of Human-CAV Interaction.

*Table 4 – I/O data of Human-CAV Interaction*

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

### AI Modules

*Table 5* gives the AI Modules of the Human-CAV Interaction depicted in Figure 3.

*Table 5 – AI Modules of Human-CAV interaction*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Audio Scene Description | 1. Separates speech signals in Audio (ESS) and Input Audio from non-speech signals. 2. Localises speech sources in space. |
| Speaker identification | Recognises speaker. |
| Speech recognition | 1. Analyses the Speech inputs. 2. Generates text and emotion output. |
| Visual Scene Description | 1. Separates humans from other objects and locates them. 2. Provides:    1. Non-human objects.    2. Visual human elements.    3. Face and attributes. |
| Object analysis | 1. Receives non-human objects. 2. Produces ID of such object in focus. |
| Gesture analysis | 1. Receives human torsoes 2. Produces Expression of Physicsl Gesture. |
| Face identification | 1. Receives Face Objects. 2. Recognises the identities of the humans. |
| Face analysis | 1. Receives Face Objects. 2. Extracts Expression of Face Objects. |
| Language understanding | 1. Analyses natural language expressed as text using an embedded language model. 2. Produces LangUnd Text 3. Produces the Meaning of LangUnd Text. |
| Emotion Fusion | Produces Final Emotion by fusing Emotion (Speech), Emotion (Face) and Emotion (Gesture). |
| Question analysis | Determines Intention by analysing the Meaning of the sentence. |
| Question & Dialogue Processing | 1. Receives Speaker ID and Face ID. 2. If speaker ID and face ID match, then:    1. Produces a command to Autonomous Motion SS    2. Analyses LangUnd Text and user’s Fused Emotion, Intention, Meaning and/or question.    3. Produces Expression (Speech), Expression (Face) and Output Text. 3. Else, responds appropriately. |
| Speech synthesis | Converts Concept (Speech) to Output Speech. |
| Face animation | Converts Concept (Face) to Output Face. |

## Avatar Videoconference

### Scope of Use Case

This Use Case is part of the Mixed-reality Collaborative Space (MCS) project. An MCS is a virtual environment populated by:

1. Virtual twins of humans – embodied in speaking avatars having a specified degree of similarity, in terms of voice and appearance, with their human twins – directed by participating humans to achieve an agreed goal.
2. Human-like speaking avatars possibly without a visual appearance and not representing a human, e.g., a secretary taking notes of the meeting, possibly with the ability to make independent decisions about its actions.

Other examples of collaboration are: attending a lecture, and visiting a virtual place. The space where the collaboration takes place is called Ambient and can be anything from a fictitious space to a replica of a real space.

This Use Case assumes that:

1. A group of geographically dispersed humans participate in a videoconference. Groups of humans may be in the same physical environment.
2. The videoconference room is an MCS whose Ambient is provided by a service provider properly equipped with table and chairs and virtual twins of humans – torsoes of avatars resembling their physical twins – sit at the chairs.
3. Each participant has or selects his/her avatar model.
4. The transmitting part of the MCS client of each participant:
   1. Transmits the participant’s avatar model, speech and language preference.
   2. Extracts Meaning and fused Text, Speech, and Face Emotion from speech and face.
   3. Creates a set of Descriptors of the participant’s face head, arms, hands and fingers using visual technologies integrated with the results of the speech technologies applied in 4.b.
5. The MCS server
   1. Authenticates all participants using speech and face information.
   2. Sends all participants the Ambient Descriptors.
   3. Uses the language preferences to translate all speech signals to the requested languages and send the original and translated speech signals to participant according to their language preferences.
   4. Forwards the avatar models and descriptors to all participants
6. The receiving part of the MCS client of each participant:
   1. Places all avatar models animated with the received descriptors with their speech signals to the chairs selected by the participants.
   2. Selects a Point of View that may be different from the one corresponding to his/her position in the Ambient.
   3. Watches the 3D visual space with the device of his/her choice (HMD or 2D display) and and listens to the spatial audio.

### TX Client AIW

#### Function

The function of the AIW is to

1. Receive:
   1. Input Audio (that includes Audio field created by RX Client) from microphone (ar-ray).
   2. Video of participant from camera (array).
   3. Participant’s Avatar Model.
   4. Participant’s colour and style of synthetic voice (e.g., for use in speech translation).
   5. Participant’s spoken language preferences (e.g., EN-US, IT-CH).
2. Send to server for:
   1. Authentication
      1. Speaker Authentication Descriptors
      2. Face Authentication Descriptors.
   2. Translation
      1. Participant’s colour and style of synthetic voice.
      2. Participant’s spoken language preferences (e.g., EN-US, IT-CH).
   3. Distribution to other Participants:
      1. Speech.
      2. Avatar Model.
      3. Avatar Descriptors.

#### Architecture

Each participant

1. At the start, sends to Server
   1. For distribution to Participants:
      1. Language Preferences.
      2. Avatar model.
   2. For authentication:
      1. Speaker Descriptors for Authentication
      2. Face Descriptors for Authentication
2. During the meeting:
   1. Speech Separation produces
      1. Participant’s Speech
   2. Speech Recognition
      1. Receives Speech
      2. Produces Recognised Text and Emotion (Text)
      3. Emotion Analysis
   3. Emotion Analysis
      1. Receives Speech
      2. Produces Emotion (Speech)
   4. Face Analysis1
      1. Receives Video of Face
      2. Produces Emotion (Video)
   5. Face Analysis2
      1. Receives Video of Face
      2. Produces Face Descriptors
   6. Head Analysis
      1. Receives Video of Face
      2. Produces Head Descriptors
   7. Gesture Analysis
      1. Receives Video of Arms, Hands and Fingers
      2. Produces Gesture Descriptors
   8. Face Analysis3
      1. Receives Video of Face
      2. Produces Face Authentication Descriptors
   9. Language Understanding
      1. Receives Recognised Text
      2. Produces Meaning
   10. Emotion Fusion
       1. Receives Emotion (Text), Emotion (Speech) and Emotion (Face)
       2. Produces Fused Emotion
3. Avatar Description
   1. Receives Fused Emotion, Meaning, Face Descriptors, Head Descriptors and Gesture Descriptors
   2. Produces Avatar Descriptors
4. TX Clients sends to Server for distribution
   1. Speech.
   2. Avatar Descriptors.

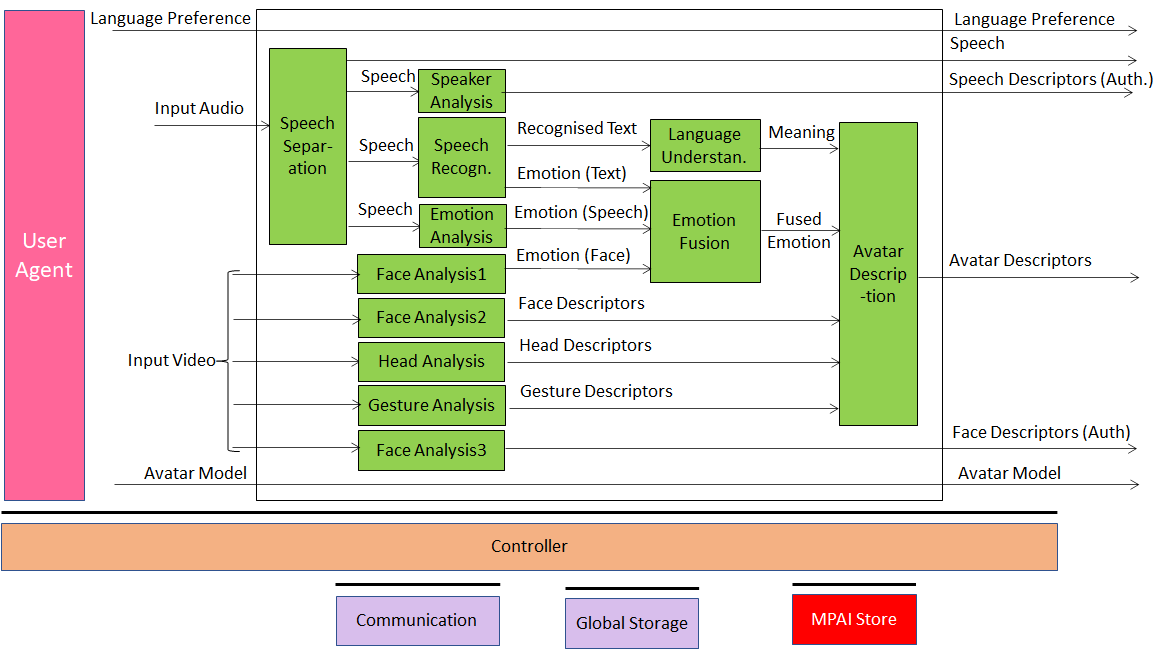


Figure 4 - Reference Model of Avatar Videoconference TX Client

#### Input and output data

*Table 6* gives the input and output data of TX Client AIW:

*Table 6 – Input and output data of Client TX AIW*

|  |  |
| --- | --- |
| **Input** | **Comments** |
| Language Preference | The language Participant wishes to hear in the Ambient. |
| Input Audio | Audio composed of Participant’s Speech, RX Client Audio and other audio. |
| Input Video | Video of Participant’s torso (head, face, arms, hands). |
| Avatar Model | The avatar model selected by Participant. |
| **Output** | **Comments** |
| Language Preference | As in input. |
| Speaker Descriptors (Auth.) | Participant’s Speech Descriptors for Authentication. |
| Participant’s Speech | Speech as separated from Ambient Audio. |
| Avatar Descriptors | As in input. |
| Face Descriptors (Auth.) | Participant’s Face Descriptors for Authentication |

#### AI Modules

*Table 7* gives the AI Modules of TX Client AIW.

*Table 7 – AI Modules of TX Client AIW*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Speech Separation | Provides Speech out of non-speech Sound in Input Audio. |
| Speaker Analysis | Provides Speaker’s Speech Descriptors for Authentication |
| Speech Recognition | Provides Text and Emotion from Separated Speech. |
| Emotion Analysis | Provides Emotion from Speech. |
| Face Analysis1 | Provides Emotion from Input Video (face). |
| Face Analysis2 | Provides Face Descriptors for reproduction of face on avatar. |
| Head Analysis | Provides Head Descriptors for reproduction of head on avatar. |
| Gesture Analysis | Provides Gesture Descriptors for reproduction of arms, hands and fingers of avatar. |
| Face Analysis3 | Provides Face Descriptors for Authentication. |
| Language Understanding | Provides Meaning from Recognised Text. |
| Emotion Fusion | Provides Fused Emotion from Emotion (Speech) and (Emotion Face). |
| Avatar Description | Produces the full set of Avatar Descriptions. |

### Server AIW

#### Function

The function of the Server AIW is:

1. At the start:
   1. Authenticate Participants.
   2. Create Ambient.
   3. Distribute Participants’ avatar models.
2. During the videoconference
   1. Translate and send speech to participants according to their preferences.
   2. Forward Avatars Descriptors to all participants.

#### Architecture

The Server:

1. Receives from:
   1. Server manager:
      1. Selected Ambient.
      2. Participants’ Identities.
   2. Each Participant:
      1. Speaker Descriptors (Auth.)
      2. Face Descriptors (Auth.)
      3. Selected Language
      4. Speech.
      5. Avatar Descriptors.
2. Authenticates Participants
3. Sends:
   1. Ambient Descriptors.
   2. Participants’ IDs.
   3. Participants’ Speech.
   4. Participants’ Avatar Descriptors.

Figure 5 gives the architecture of Server AIW.

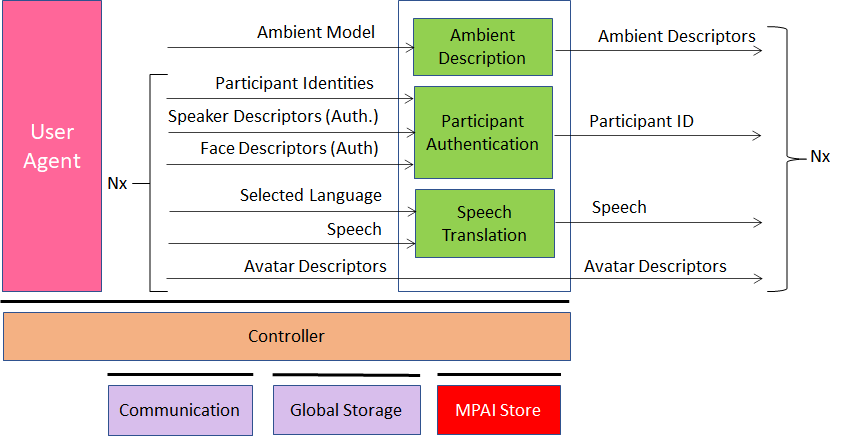


Figure 5 - Reference Model of Avatar Videoconference Server

#### I/O data

*Table 8*gives the input and output data of Server AIW.

*Table 8 – Input and output data of Server AIW*

|  |  |
| --- | --- |
| **Input** | **Comments** |
| Ambient Selection | Set by Conference Manager |
| Participant Identities (xN) | Assigned by Conference Manager |
| Speaker Descriptors (Auth.) (xN) | Participant’s Speech Descriptors for Authentication |
| Face Descriptors (Auth.) (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** |
| Ambient Descriptors (xN) | Static Ambient 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 9* gives the AI Modules of Server AIW.

*Table 9 – AI Modules of Server AIW*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Ambient Description | Creates all static Ambient Descriptors. |
| Participant Authentication | Authenticates Participants using their Speech and Face Descriptors |
| Translation | 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. |

### RX Client AIW

#### Function

The Function of the RX Client AIW is to:

1. Create the Ambient Using Ambient Descriptors.
2. Place Avatars at selected positions
   1. Based on participant’s input
   2. Using Avatar Descriptors.
   3. Adding Speech to each Avatar.
3. Present Ambient and Avatars.
4. Let Participant select the viewpoint in the Audio-Visual Scene.

#### Architecture

The RX Client AIW:

1. Creates the AV Scene using:
   1. The Ambient descriptors.
   2. The Avatars descriptors.
   3. The Speech of each Avatar.
2. Presents the Audio-Visual Scene based on the selected viewpoint in the Ambient.

Figure 6 gives the architecture Client RX AIW.

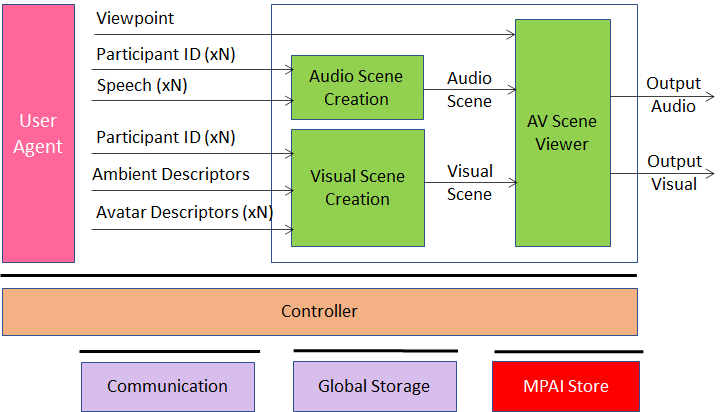


Figure 6 - Client-Based Ambient RX Client Reference Model

#### I/O data

*Table 10* gives the input and output data of RX Client AIW.

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

|  |  |
| --- | --- |
| **Input** | **Comments** |
| Viewpoint | Participant-selected point to see visual objects and hear audio objects in the Ambient |
| Participants’ IDs (xN) | Unique Participants’ IDs |
| Speech (xN) | Participantss Speech |
| Ambient Descriptors | Static Descriptors of Ambient |
| Avatar Descriptors (xN) | Descriptors of Avatars bodies |
| **Output** | **Comments** |
| Output Audio | To be presented using loudspeaker array |
| Output Visual | To be presented using 2D or 3D display |

#### AI Modules

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

*Table 11 – AI Modules of Client-Based Ambient*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Audio Scene Creation | Creates Audio Scene resulting from 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. |

# Functional Requirements

## I/O Data summary

### Conversation About a Scene

For each AIM (1st column), *Table 12* gives the input (2nd column) and the output data (3rd col-umn).

*Table 12 – I/O data of Conversation About a Scene AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| Video Analysis1 | Input Video | Emotion (Face) |
| Video Analysis2 | Scene Descriptors | Gesture Descriptors |
| Video Analysis3 | Input Scene Video | Scene Descriptors |
| Speech Recognition | Input Speech | Recognised Text  Emotion (Speech) |
| Object ID | Gesture Descriptors  Scene Descriptors | Object ID |
| Scene2Video | Scene Descriptors | Output Scene  Video |
| LangUnd Texr | Recognised Text  Object ID | Emotion (Text)  Meaning  LangUnd Text |
| Emotion Fusion | Emotion (Text)  Emotion (Speech)  Emotion (Face) | Fused Emotion |
| Question Analysis | Meaning | Intention |
| Dialogue Processing | Fused Emotion  Meaning  LangUnd Text  Intention | Expression (Speech)  Expression (Face) |
| Speech Synthesis (Expression) | Expression (Speech) | Output Speech |
| Face Animation (Expression) | Expression (Face) | Output Face |

### Human-CAV Interaction

For each AIM (1st column), Table 13 gives the input (2nd column) and the output data (3rd col-umn).

*Table 13 – I/O data of Human-CAV Interaction AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| Speech Separation | Input Audio | Separated Speech |
| Internal AV Scene | Input Video | Face Objects |
| Speaker Recognition | Separated Speech | Speaker ID |
| Speech Recognition | Separated Speech | Emotion (Speech)  Text (Speech) |
| Object and Gesture Analysis | Input Video | Object ID  Emotion (Gesture)  Meaning (Gesture) |
| Face Analysis | Face Objects | Emotion (Face)  Meaning (Face) |
| Face Identification | Face Objects | Face ID |
| Emotion Fusion | Emotion (Speech)  Emotion (Face)  Emotion (Gesture) | Fused Emotion |
| Language Understanding | Text (Speech)  Input Text  Object ID | Text (Language Understanding)  Meaning |
| Question analysis | Meaning | Intention |
| Question and dialogue processing | Input Text  Speaker ID  Fused emotion  Text (Speech)  Meaning  Intention  Face ID  Face Objects | Command/Request  Feedback/Response  Concept (Speech)  Output Text  Concept (Face) |
| Speech synthesis | Concept (Speech) | Output Speech |
| Face animation | Concept (Face) | Output Face |

### Avatar Videoconference

#### TX Client

For each AIM (1st column), *Table 14* gives the input (2nd column) and the output data (3rd col-umn).

*Table 14 – AIMs and Data of TX Client AIW*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
|  | Language Preference | Language Preference |
| Speech Separation | Input Audio | Separatated Speech |
| Speaker Analysis | Speech | Speech Descriptors (Auth.) |
| Speech Recognition | Text  Speech | Recognised Text  Emotion (Text) |
| Emotion Analysis | Speech | Emotion (Speech) |
| Face Analysis1 | Input Video | Emotion (Video)  Meaning (Video) |
| Face Analysis2 | Input Video | Face Descriptors |
| Head Analysis | Input Video | Head Descriptors |
| Gesture Analysis | Input Video | Gesture Descriptors |
| Face Analysis3 | Input Video | Face Descriptors (Auth.) |
| Language Understanding | Recognised Text | Meaning (Text)  Meaning (Speech) |
| Emotion Fusion | Emotion (Text)  Emotion (Speech)  Emotion (Video) | Fused Emotion |
| Question Analysis | Meaning (Text)  Meaning (Speech) | Fused Meaning |
| Avatar Description | Fused Emotion  Fused Meaning  Face Descriptors  Head Descriptors  Gesture Descriptors | Avatar Descriptors |
|  | Avatar Model | Avatar Model |

#### Server

For each AIM (1st column), *Table 15* gives the input (2nd column) and the output data (3rd col-umn).

*Table 15 – AIMs and Data of Server AIW*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| Ambient Description | Ambient Type | Ambient Descriptors |
| Participant Authentication | Participant Identities  Speech Descriptors (Auth.)  Face Descriptors (Auth.) | Participant ID (xN) |
| Translation | Selected Languages  Participants’ Speech | Speech |
|  | Avatar Descriptors | Avatar Descriptors |

#### RX Client

For each AIM (1st column), *Table 16* gives the input (2nd column) and the output data (3rd col-umn).

*Table 16 – AIMs and Data of RX Client AIW*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input** | **Output Data** |
| AV Object Operation | AV Object  AV Object Action | Audio Object  Visual Object |
| Audio Scene Creation | Participant ID  Speech | Audio Scene |
| Visual Scene Creation | Participant ID  Ambient Descriptors  Avatar Descriptors | Visual Scene |
| AV Scene Viewer | Viewpoint  Audio Scene  Visual Scene | Output Audio  Output Visual |

## Data format requirements

### Ambient Descriptors

The set of Descriptors required to represent and render the static components of an Ambient:

1. Visual Objects, e.g., table, swivel chair, walls, furniture, with their affordances.
2. Audio Objects: e.g., an audio source at given coordinates in the Ambient.

The Descriptors should allow:

1. The elements bounding the environment (walls, ceiling, floor, doors., windows).
2. Visual objects in the Ambient (i.e., table, chairs).
3. Audio-visual objects (e.g., a radio generating music).
4. Integration of animated avatars.
5. Association of Audio Objects, e.g., a Speech to a Visual Object, e.g., an Avatar.
6. Viewing and hearing the AV Scene from an arbitrarily selected Viewpoint.

**To Respondents**

MPAI Requests respondents to propose a set of Descriptors that would allow the digital repre-sentation and rendering of a static Ambient with the features described and with the ability to accommodate any other dynamic Audio and Visual Descriptors identified in this document.

### Ambient Model

The Ambient Model is the file format of the Ambient. The Ambient Model has:

1. The visual component represented as Industry Foundation Classes (IFC)/BIM files.
2. The Audio component populated with static audio sources.

**To Respondents**

MPAI requests respondents:

1. To comments on the selected Industry Foundation Classes (IFC)/BIM for the visual compo-nent or to propose alternative formats.
2. To propose a format for the audio component.

### Audio

Monochannel Audio is the digital representation of an analogue audio signal sampled at a fre-quency between 8-192 kHz with a number of bits between 8 bits/sample and 32 bits/sample and a quantisation that is linear or companded.

### Audio Scene Descripors

The descriptors of the structured composition of Audio Objects in two different modalities:

1. The Audio Scene is captured from the real world. Two examples are provided:
   1. CAV: a group of humans approach a CAV or are sitting in the cabin. The CAV separates the different sound sources (speech and other sounds) e spatially locates them.
   2. A group of humans is sitting in a room. A microphone array separates the different sound sources (speech and other sounds) e spatially locates them.
2. The Audio Scene is used to create a sound field in a virtual environment. Two examples are provided:
   1. In Avatar-based Videoconference, avatars are sitting around a table and the mouth of each avatar emits the sound of speech of the human it represents.
   2. In Avatar-based Videoconference, an object (e.g., a radio) emits stereo sound from a given place

MPAI-CAE 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 of this document.

### Audio-Visual Scene Descriptors

The descriptors of the structured composition of Audio-visual Objects for the purpose of representing an audio-visual scene captured from the real world of to build an audio-visual scene in a virtual world. For example:

1. Description of audio-visual objects in the real world: a group of speaking humans with a background noise.
2. Description of audio-visual objects for a virtual world: a set of avatars in a conference rooms.

**To Respondents**

MPAI requests respondents to propose Audio-Visual Scene Descriptors that integrates Audio Scene Descriptors and Visual Scene Descriptors.

### Avatar Descriptors

A set of Descriptors allowing the animation of an Avatar Model based on movement of

1. Muscles of the face (e.g., eyes, lips).
2. Head, arms, hands, and fingers.

**To Respondents**

MPAI requests respondents to propose a set of Avatar Descriptors supporting the stated require-ments.

### Avatar Model

The static model of an Avatar (in the sense that the part of the body below the waist is static). The torso can rotate if the avatar turns it to watch the avatar next to it. Head, arms, hands and fingers are animated, based on the movement of the human it represents.

The model should be extensible to respond to future MPAI Calls for Technology requesting that also the currently static parts of the body are animated, e.g., in case the avatar walks in the Am-bient.

**To Respondents**

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

### Emotion (Face)

The face of a human may express an emotion. MPAI assumes that the same emotion identifiers and semantics defined for text and speech can be used for face [3].

**To Respondents**

Respondents are requested:

1. To comment on the use of [3] to represent the emotion of a face, or
2. To propose extensions or alternative representations of emotion of a face.

### Emotion (Speech)

MPAI-MMC specifies an extensible Basic Emotion Set and Text with Emotion as Reply (speech) format [3].

**To respondents**

Respondents are requested to comment on the suitability of the technology standardised in [3] for the purpose of supporting human-machine dialogue. If a respondent claims that the technolo-gy in [3] is unsuitable, respondent is requested to motivate their claims, and propose an extension of the MPAI Basic Emotion Set or a new solution.

Respondents are requested to propose technology that can synthesise speech having as input:

1. Time-stamped (Sub-)Word Lattices.
2. Sequences of time-stamped Expressions.
3. Duration of transitions between Expressions.

### Emotion (Text)

MPAI-MMC [3] specifies an extensible Basic Emotion Set.

**To respondents**

Respondents are requested to comment on the suitability of the technology standardised in [3.] for the purpose of supporting human-machine dialogue. If a respondent claims that the technolo-gy in [3.] is unsuitable, respondent is requested to motivate their claims, and propose an extension of the MPAI Basic Emotion Set or a new solution.

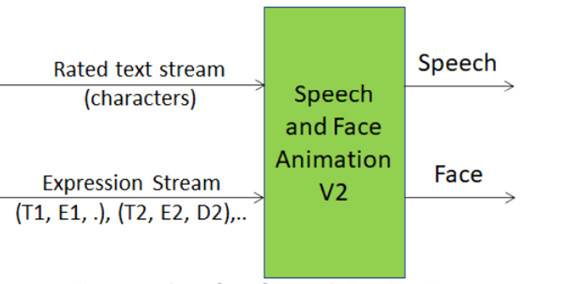


Figure 7 - Speech and Face Animation V2

The dynamic expression representation allows for two different expressions to happen at the same time, possibly with different durations. In other terms, (T1, E1, D1), (T1, T2, D2) is a syntactically correct expression. However, this may not correspond to a combination that is displayed by the face of a human being.

As Expression include Emotions, the same represention of the expressions of a face varing in time can be used to represent emotions of a face varying in time.

### Expression (Face)

MPAI defines Expression (Face) as a set of attributes of the human face:

* Movement of facial muscles (including mouth muscles)
* Movement of eyes
* Movement of head

reflecting a particular state of mind or intention.

Facial Emotion can be extracted from facial expression. Facial Emotion is typically connected with the speech uttered simultaneously with the facial expression. Speech Emotion may coincide with Facial Emotion or be different.

The Facial Action Coding System (FACS) [16] is an anatomically based system capable of de-scribing all facial movements. Facial expressions are broken down into Action Units, elementary components of muscle movement.

**To Respondents**

MPAI is looking for a technology capable to use Head and Face Descriptors to extract the Expression of the Head and Face.

This includes an ontology of expressions and the corresponding Head and Face Descriptors.

MPAI is also looking at a way to represent the expressions of a face varing in time as:

1. Sequences of time-stamped expressions aligned with speech.
2. Duration of transitions between expressions.

The representation should allow for two separate 1 and 2 streams or as a multiplexed 1 and 2 streams.

### Face Objects

Face object is the image of a face that is passed to an AIM for different different types of processing purposes such as:

1. To identify a human.
2. To extract the Emotion or the Expression from the face.

The Face Object should contain the spatial coordinates of the face. One application is to enable an avatar at gaze at the human whose face object has been provided.

**To respondents**

Respondents are invited to propose a Face Object format satisfying the above requirements. This will be used as an input to an AIM, e.g., Face Analysis, Face Identification and Avatar Animation.

### Fused Emotion

Fused Emotion is the Emotion resulting by combining any of the following: Emotion (Text), Emotion (Speech) and Emotion (Video) to improve the animation of an avatar face.

MPAI-MMC has defined a digital representation of Emotion [3.].

**To Respondents**

MPAI requests respondents to comment on the use of [3.] or to propose extensions or alternative Emotion representation formats.

### Head and Face Descriptors

MPAI requests Face Descriptors for the following often 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.
2. To describe the features of a human face and head for the purpose of extracting the facial Expression.
3. To describe head and face movement concomitant with a Physical Gesture.
4. To describe the movement of facial muscles including:
   1. Lip movement.
   2. Eye movement.
5. To animate head and face of an avatar using any of 2, 3, 4.

The notion of keypoints can be used to describe Head, Face and Physical Gesture. Keypoints are independent of the technology (i.e., image processing, or ML, or their combinations) and are used to implement the Keypoint Detection AIM(s).

**To Respondents**

MPAI requests respondents to propose a set of Head and Face Descriptors suitable for the purposes identi-fied above.

### Head Descriptors

The coded representation of the movement of a human head. It should include

1. The coordinates of the point representing the head
2. The trajectory of the point representing a moving head
3. The rotation of a moving head.

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

**To Respondents**

MPAI requests respondents to propose a set of head descriptors satisfying the requirements.

### Input Audio

The digital representation of the audio captured from the real world.

MPAI-CAE has defined a digital representation of a microphone set [4.].

**To Respondents**

MPAI requests:

MPAI requests respondents to

1. To comment on the use of [4.] to capture audio from a Participant’s location for the purpose of separating speech from the captured audio, or
2. To propose extensions to [4.], or
3. To propose alternative representation formats.

### Input Human Video

See Video

### Input Speech

See Speech

### Input Video

Input Video is the digital visual representation of the Participant’s torso, i.e., the upper part of the body for the purpose of digitally representing head, face, arms and hands of a Participant.

**To Respondents**

MPAI requests respondents to propose an Input Video format that facilitates the optimal extrac-tion of head, face, and gesture descriptors.

### Intention

MPAI-MMC [3.] specifies a digital representation format for Intention.

**To respondents**

Respondents are requested to comment on the suitability of the technology standardised in [3.]. Motivated proposals demonstrating superior performance are welcome.

### Language Preferences

Languages are expressed as specified by [6.].

**To Respondents**

MPAI requests respondents to comment on the choice.

### LangUnd Text

See Text

### Meaning

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

**To respondents**

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

### Microphone Array Audio

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 in single or double precision float.

MPAI needs Microphone Array Audio

1. To separate:
   1. The individual Speech Objects from a Composite Audio Object for the following Use Cases:
   2. Enhanced Audioconference Experience (EAE), as standardised in [4.].
   3. Mixed-reality Collaborative Spaces (MCS).
   4. Human-CAV Interaction (HCI)
   5. Individual Audio Objects Objects from a Composite Audio Object for the following Use Cases:
   6. Environment Sensing Subsystem (ESS).
   7. Audio-On-the-Go (AOG).
2. To determine the Audio Scene.
3. To classify the Audio Objects.

**To respondents**

Respondents are requested to comment on the suitability of the technology standardised in [4.].

### Object Identifier

MPAI-MMC [3] specifies a digital representation format for Object Identifier to be used to identify objects held in the hand of a person.

**To respondents**

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

### Output Speech

See Speech

### Physical Gesture Descriptors

Physical Gesture is the coded representation of movement of head, arms, hands, and fingers suit-able for:

1. Recognition of sign language.
2. Recognition of coded hand signs, e.g., to indicate a particular object in a scene.
3. Representation of arbitrary head, arm, hand and finger motion.
4. Culture-dependent signs (e.g, mudra sign).

As an example, recognition of sign language could be implemented as follows:

1. Detect standard key points on head, arm, hands, and fingers for a certain duration of time.
2. Compare the current sequence of key points with a database of key point sequences.
3. Get the phonemes or words represented by the set of key points from a KB.

Point 2 and 3 are language dependent.

The above is depicted by Figure 8.

A picture containing timeline

Description automatically generated

Figure 8 – Sign language keypoint detections

**To Respondents**

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

### Ranked Text

The output of Speech Recognition where characters or words have a probability.

**To respondents**

MPAI requests a format for Ranked Text that

1. Is timestamped.
2. Provides:
   1. Time interval.
   2. Strings of characters.
   3. Corresponding probability.

### Recognised Text

See Text

### Scene Descriptor Video

See Video

### Spatial coordinates

A use case may need to make reference to coordinate system. In Conversation About a Scene the machine needs to know the position of the human because its avatar should gaze at the face of the human when it has a conversation to him/her.

Three different cases are considered:

1. The coordinates are absolute on the surface of the Earth, e.g., in the case of a Connected Autonomous Vehicle.
2. The coordinates are defines with reference to a specific point on the Earth.
3. The coordinates are possibly arbitrarily defined local coordinates, e.g., in the case of Conversation About a Scene. The specific use case defines what is the point with coordinates (0,0,0).
4. The coordinates are in a virtual space. The specific use case defines what is the point with coordinates (0,0,0).

**To respondents**

MPAI requests respondents to provide a coordinate system that can represent a point on a known place on the surface of the Earth, represent a point with reference to a arbitarily defined point on the Earth, a physical point which does not have a defined correspondence with a point on the Earth and a point in a virtual space.

### Text

MPAI-MMC [3] specifies ISO/IEC 10646, Information technology – Universal Coded Charac-ter Set (UCS) [5] as digital Text representation to support most languages in use.

**To respondents**

Respondents are invited to comment on this choice.

### Verbal Human-CAV Interaction

Some commands given to the Autonomous Motion Subsystem are:

1. Go to a Waypoint.
2. How long does it take to get to a waypoint.
3. Park close to a Waypoint.
4. Drive faster.
5. Drive slowly.
6. Display Full World Representation.

Some of the responses of the Autonomous Motion Subsystem are:

1. Enumeration of possible routes with major features of each route.
2. Enumeration of possible parking places with major features of each place.
3. Announcement of obstacles preventing the expeditious accomplishment of the Command.
4. Announcement that the desired Waypoint has been reached.

**To respondents**

Respondents are requested to propose a coded representation of the above Com-mands/Responses. Proposals of coded representation of additional responses are welcome.

### Video

Video is intended for use in the passenger cabin. MPAI-MMC [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 choice for 2D Video.

Respondents are also requested to propose a data format for an array of cameras having vid-eo+depth as the baseline format or other 3D Video data formats.

### Viewpoint

A Viewpoint is

1. The point expressed with one of the spatial coordinates from where a user looks at the space around him/her.
2. The direction the user looks at the space (theta, phi).

**To Respondents**

MPAI requests respondents to propose a way to represent a Viewpoint.

### Visual Scene Descriptors

MPAI requires descriptors to describe the structured composition of Visual Objects. Descriptors should serve two purposes:

1. The Visual Scene is captured from the real world. Three examples are provided:
   1. A group of humans approach a CAV or are sitting in the cabin. The CAV
      1. Separates the different visual objects (humans and other objects),
      2. Identifies face, head, arms, hands and fingers of all humans.
      3. Spatially locates all objects as they move.
   2. A human is sitting in a room. A camera
      1. separates the human from the rest of the environment
      2. spatially locates him/her
      3. captures face, head, arms, hands, and fingers.
   3. A human is part of a scene populated by objects. A camera
      1. separates the human from the rest of the environment
      2. spatially locates the human
      3. captures face, head, arms, hands, and fingers.
2. The Visual Scene Descriptors are also used to create a visual scene in a virtual environment. Two examples are provided:
   1. In Avatar-based Videoconference, avatars are sitting around a table and reproduce faces, head, arms, hands and fingers of the human thei represents.
   2. In Human-CAV Interaction, an avatar represents the virtual CAV driver displaying face characteristics and head movement congruent with the speech it utters.

**To Respondents**

MPAI requests respondents to propose Visual Scene Descriptors satisfying the needs of the Use Cases of this document.

# Annex 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 17.

*Table 17 – 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. |
| 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. |

# Annex 2 - 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/governance/) 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.

# Annex 3 – 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.

|  |  |
| --- | --- |
| 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 **

*Figure 7* 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 7 – The MPAI ecosystem operation*

# Annex 2 - Notices and Disclaimers Concerning MPAI Standards (In-formative)

The notices and legal disclaimers given below shall be borne in mind when downloading and us-ing approved MPAI Standards.

In the following, “Standard” means the collection of four MPAI-approved and published docu-ments: “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. An MPAI Standard may only be developed when a Framework Licence has been adopted. MPAI Standards are devel-oped by especially established MPAI Development Committees who operate on the basis of con-sensus, as specified in Annex 1 of the MPAI Statutes. While the MPAI General Assembly and the Board of Directors administer the process of the said Annex 1, MPAI does not independent-ly 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 to deter-mine the status of any given published MPAI Standard.

Comments on MPAI Standards are welcome from any interested parties, whether MPAI mem-bers 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. Comments will be reviewed by the appropriate committee for their tech-nical relevance. However, MPAI does not provide interpretation, consulting information, or ad-vice 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 express-ly 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, especial-ly 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, INCI-DENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: THE NEED TO PROCURE SUBSTITUTE GOODS OR SER-VICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOW-EVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARIS-ING 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 par-ties. 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 cor-rect implementation of the MPAI Standard. Moreover, positive Performance Assessment of an implementation applies exclusively in the context of the MPAI Governance 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 regula-tory 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 copy-right to its Standards. For inquiries regarding the copyright of MPAI standards, please contact the MPAI Secretariat.

The Reference Software of an MPAI Standard is released with the MPAI Modified Berkeley Software Distribution 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.

# Annex 3 – 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 in-ter¬connected 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.

Implementers’ benefits Upload to the MPAI Store and have globally distributed Implementations of

- AIFs conforming to MPAI-AIF.

- AIWs and AIMs performing prop¬rietary 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 Implem¬entation, the AIW must be an Implementation of an MPAI Use Case and the AIMs must con¬form with an MPAI Applicati¬on 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’ bene-fits - 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 Func¬tions and the data Formats are known.

Market’s ben-efits - Open AIW and AIM markets foster competition leading to better prod-ucts.

- 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. Howev-er, the performance of an AIM is typically higher if the data used for training are in greater quan-tity and more in tune with the scope. Training data that have large variety and cover the spec¬trum 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’ bene-fits Get assurance that the Implementation being used performs correctly, e.g., it has been properly trained.

Market’s ben-efits 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 Figure 9 is a high-level description of the MPAI ecosystem operation applicable to fully conform-ing 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 Implement-ers (step 5a) and MPAI Store.
6. Implementers submit Implementations to the MPAI Store tested for Confor¬mance and securi-ty.
7. Users download and use Implementations, and submit experience scores.

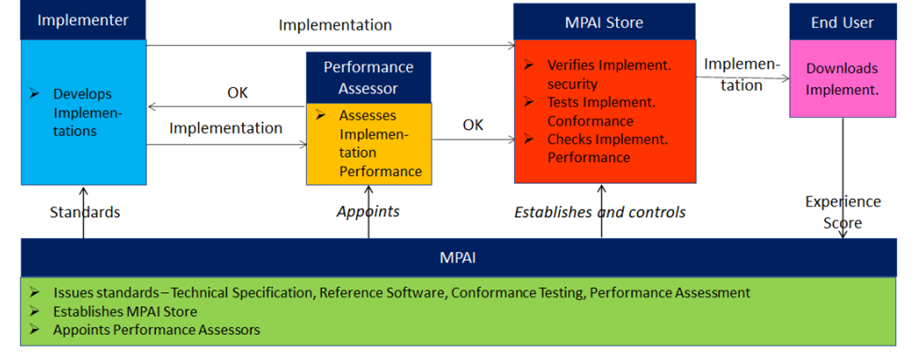


Figure 9 – The MPAI ecosystem operation