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

|  |  |
| --- | --- |
|  | **Public document** |
| **N694** | 2022/05/18 |
| **Source** | MMC-DC |
| **Title** | MPAI-MMC Use Cases and Functional Requirements V2 WD1.4 |
| **Target** | MPAI Members |

Contents

[Abstract 2](#_Toc103229600)

[1 Introduction 2](#_Toc103229601)

[2 Scope of the Use Cases 4](#_Toc103229602)

[3 Terms and Definitions 5](#_Toc103229603)

[4 References 7](#_Toc103229604)

[4.1 Normative References 7](#_Toc103229605)

[4.2 Informative References 7](#_Toc103229606)

[5 Use Case Architectures 8](#_Toc103229607)

[5.1 Conversation About a Scene 8](#_Toc103229608)

[5.1.1 Scope of Use Case 8](#_Toc103229609)

[5.1.2 Reference architecture 8](#_Toc103229610)

[5.1.3 Input/output data 9](#_Toc103229611)

[5.1.4 AI Modules 9](#_Toc103229612)

[5.2 Human-CAV Interaction AIW (HCI) 10](#_Toc103229613)

[5.2.1 Scope of Use Case 10](#_Toc103229614)

[5.2.2 Reference architecture 11](#_Toc103229615)

[5.2.3 Input and output data 12](#_Toc103229616)

[5.2.4 AI Modules 13](#_Toc103229617)

[5.3 Avatar-Based Videoconference 14](#_Toc103229618)

[5.3.1 Scope of Use Case 14](#_Toc103229619)

[5.3.2 Transmitting Client AIW 15](#_Toc103229620)

[5.3.3 Server AIW 17](#_Toc103229621)

[5.3.4 Virtual Secretary AIW 19](#_Toc103229622)

[5.3.5 Receiving Client AIW 22](#_Toc103229623)

[6 Functional Requirements 23](#_Toc103229624)

[6.1 I/O Data summary 23](#_Toc103229625)

[6.1.1 Conversation About a Scene 23](#_Toc103229626)

[6.1.2 Human-CAV Interaction 24](#_Toc103229627)

[6.1.3 Avatar Videoconference 25](#_Toc103229628)

[6.2 Data format requirements 26](#_Toc103229629)

[6.2.1 Environment Descriptors 26](#_Toc103229630)

[6.2.2 Environment Model 27](#_Toc103229631)

[6.2.3 Audio 27](#_Toc103229632)

[6.2.4 Audio Scene Descripors 27](#_Toc103229633)

[6.2.5 Audio-Visual Scene Descriptors 28](#_Toc103229634)

[6.2.6 Avatar Descriptors 28](#_Toc103229635)

[6.2.7 Avatar Model 28](#_Toc103229636)

[6.2.8 Avatar Position 28](#_Toc103229637)

[6.2.9 Emotion Stream 29](#_Toc103229638)

[6.2.10 Emotion (Face) 29](#_Toc103229639)

[6.2.11 Emotion (Speech) 30](#_Toc103229640)

[6.2.12 Emotion (Text) 30](#_Toc103229641)

[6.2.13 Face Descriptors 30](#_Toc103229642)

[6.2.14 Face Objects 30](#_Toc103229643)

[6.2.15 Fused Emotion 31](#_Toc103229644)

[6.2.16 Head Descriptors 31](#_Toc103229645)

[6.2.17 Input Audio 31](#_Toc103229646)

[6.2.18 Input Human Video 31](#_Toc103229647)

[6.2.19 Input Speech 31](#_Toc103229648)

[6.2.20 Input to Translation 31](#_Toc103229649)

[6.2.21 Input Video 32](#_Toc103229650)

[6.2.22 Intention 32](#_Toc103229651)

[6.2.23 Language Preferences 32](#_Toc103229652)

[6.2.24 LangUnd (Text) 32](#_Toc103229653)

[6.2.25 Meaning 32](#_Toc103229654)

[6.2.26 Microphone Array Audio 32](#_Toc103229655)

[6.2.27 Object Identifier 33](#_Toc103229656)

[6.2.28 Output Speech 33](#_Toc103229657)

[6.2.29 Physical Gesture Descriptors 33](#_Toc103229658)

[6.2.30 Ranked Text 34](#_Toc103229659)

[6.2.31 Recognised Text 34](#_Toc103229660)

[6.2.32 Scene Descriptor Video 34](#_Toc103229661)

[6.2.33 Spatial coordinates 34](#_Toc103229662)

[6.2.34 Speech Features 34](#_Toc103229663)

[6.2.35 Summary 34](#_Toc103229664)

[6.2.36 Text 35](#_Toc103229665)

[6.2.37 Verbal Human-CAV Interaction 35](#_Toc103229666)

[6.2.38 Video 35](#_Toc103229667)

[6.2.39 Viewpoint 36](#_Toc103229668)

[6.2.40 Visual Scene Descriptors 36](#_Toc103229669)

[Annex 1 – MPAI-wide terms and definitions (Normative) 37](#_Toc103229670)

[Annex 2 - Notices and Disclaimers Concerning MPAI Standards (Informative) 40](#_Toc103229671)

[Annex 3 – The Governance of the MPAI Ecosystem (Informative) 42](#_Toc103229672)

[Annex 2 - Notices and Disclaimers Concerning MPAI Standards (Informative) 44](#_Toc103229673)

[Annex 3 – The Governance of the MPAI Ecosystem (Informative) 46](#_Toc103229674)

# Abstract

This is a Collection of Use Cases and Functional Requirements of the data types potentially requested in the development of Version 2 of the MPAI-MMC Technical Specification. It is a Working Draft that is expected to change in some parts before it is published in final form as an attachment to the MPAI-MMC V2 Call for Technologies together with the Commercial Requirements.

# 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 even more so 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 using AI-based technologies is generally more efficient than 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. 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 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 implementation of MPAI standards.

For the aforementioned reasons, 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 20 if they are common to all MPAI Standards.

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

* MPAI as provider of Technical, Reference Software, 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). 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 execution of AI Workflows (AIW), composed of basic processing elements called AI Modules (AIM).

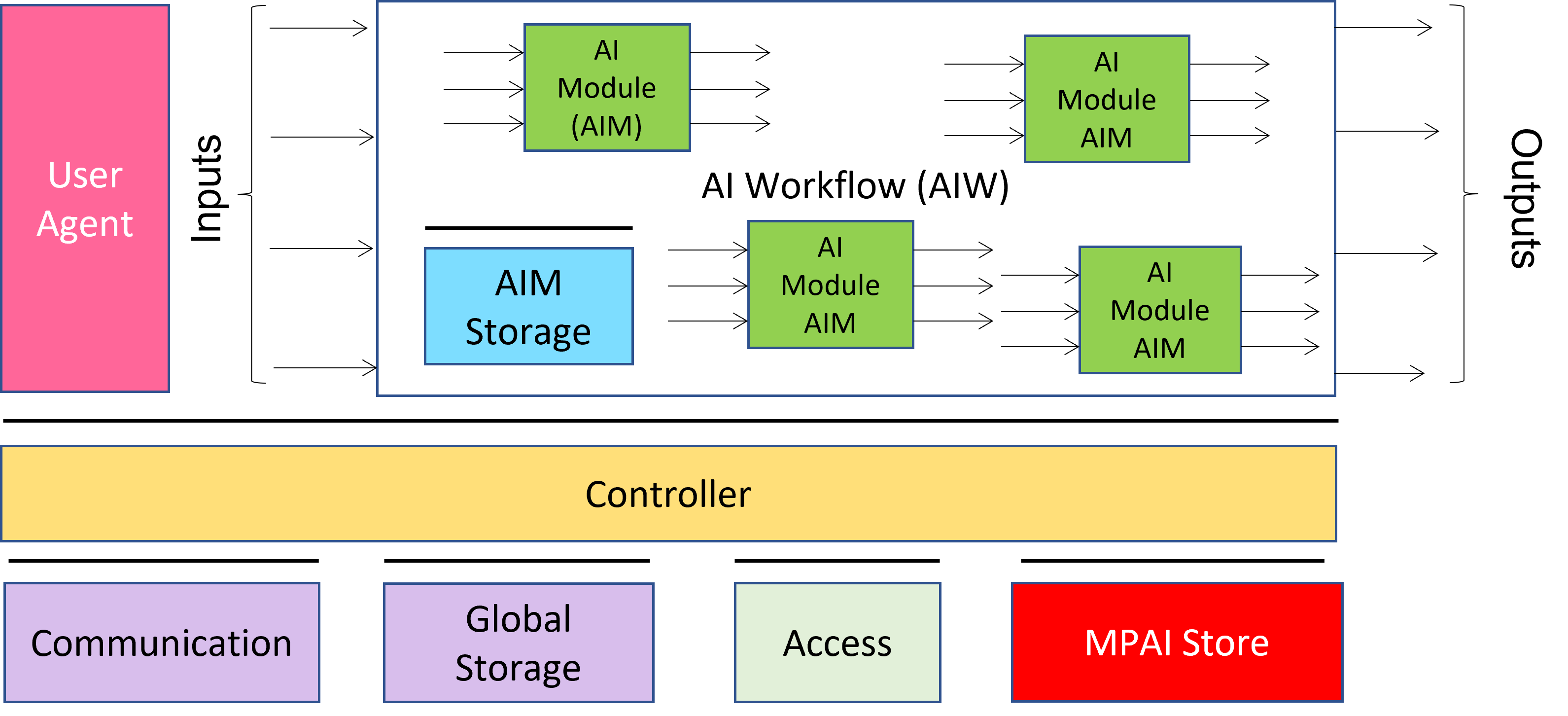


Figure 1 - The AI Framework (AIF) Reference Model and its Components

MPAI Application Standards normatively specify the Syntax and Semantics of the input and output data and the Function of the AIW and of the 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 technologies, 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. Three Interoperability Levels of an AIW executed in an AIF are defined:

*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 of Implementations 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

MPAI-MMC V2 adds the following Use Cases to those of MPAI-MMC V1:

1. *Conversation About a Scene (CAS)*: a human holds a conversation with a machine about objects in a scene of which the human is part. While conversing, the human uses a finger to indicate their interest in particular objects.
2. *Human-Connected Autonomous Vehicle Interaction (HCI)*: a group of humans holds a conversation with a Connected Autonomous Vehicle (CAV) on a domain-specific subject (travel by car). The machine understands the utterances, the emotion in their speech and in their faces, and the expression in their gestures. The machine manifests itself as the head and shoulders of an avatar whose face and head convey emotions and expressions congruent with the speech it utters.
3. *Avatar-Based Videoconference (ABV)*. Avatars representing humans participating in a videoconference reproduce the movements of the upper part of the human participants (from the waist up) with a high degree of accuracy, possibly including a *Virtual Secretary*, i.e., an avatar not representing a human participant who makes and displays a summary of what other ABV avatars say and may receive and process comments while monitoring avatars’ emotion in speech and face and expression in face and gestures.

The Use Cases of MPAI-CAE V2 require the capability to create audio scene descriptions satisfying the specific requirements of the new Use Cases. 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 colour in the avatar’s speech.
   4. The position occupied by avatars in the Environment.
   5. The point of view used in reproducing a 3D visual scene.
2. The **machine** perceives a real visual and sound space and creates:
   1. An audio scene composed of independent audio objects and their locations.
   2. A visual scene composed of independent visual objects and their locations.
   3. A complete audio-visual scene.
3. The **machine** extracts descriptors of:
   1. The human face.
   2. The Physical Gesture (i.e., head, arms, hands and fingers).
   3. The human speech.
4. 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 Emotion and the Expression of the face.
      3. Animate the face of an avatar using face descriptors that are either:
         1. Extracted from a human face.
         2. Synthetically generated.
   2. Physical Gesture Descriptors to:
      1. Extract the Expressions of the Physical Gestures.
      2. Interpret the Meaning conveyed by the Physical Gestures (may be culture and language dependent).
      3. Animate the upper part of an avatar body using Physical Gesture descriptors that are either:
         1. Extracted from a human body.
         2. Synthetically generated.
   3. Speech Descriptors to:
      1. Identify a human as belonging to a group composed of a limited number of humans (closed set identification).
      2. Recognise speech (i.e., extract text from speech).
      3. Extract the emotion in the speech.
      4. Extract the colour in the speech.
      5. Synthesise speech
5. The **machine** holds a domain-specific in-depth conversation with another party which can be a human or an avatar:
   1. In the context of a specific domain (travel by car).
   2. About objects in the machine’s visual space indicated by a human Physical Gesture.

by means of:

1. Analysing and interpreting the other party’s:
   * 1. Text in speech.
     2. Emotions in speech and face.
     3. Expressions in face and Physical Gestures.
2. Uttering speech with Emotion, possibly spatially located on the lips of an avatar.
3. Animating:
   * 1. Lips in sync with an uttered speech.
     2. Eyes, lips and face of an avatar to display Emotions and Expressions.
4. Displaying a sequence of Emotions and Expressions that are congruent with:
   * 1. The other party’s Text, Emotion and Expressions, e.g., showing satisfaction as it has understood the meaning of the question or sentence.
     2. The machine’s response and its associated Emotions and Expressions.
5. Gazing at the other party it is conversing with.
6. The **machine** summarises and refines the speech of other parties by listening to their utterances, understanding text (chats) and monitoring their Emotions and Expressions.

# 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. |
| Audio Object | Audio information with its metadata or a combination of Audio Objects. |
| Audio Scene | The Audio Objects of an AV Scene and their locations. |
| Audio-Visual Object | Audio-Visual information with its metadata or a combination of Audio-Visual Objects. |
| Audio-Visual Scene | (AV Scene) The structured composition of Audio and Visual Objects in an Environment. |
| Avatar | An animated 3D object representing a real or fictitious person in a Virtual Space. |
| Avatar Model | An inanimate avatar exposing handles for animation. |
| Connected Autonomous Vehicle | A vehicle able to autonomously reach an assigned target 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 an audio or visual feature. |
| Emotion | One of the states or processes experienced by humans such as anger or sadness, exemplified by the listings in Table 19 of [3]. An emotion normally combines cognitive and physical experience, but can be distinguished from purely cognitive states or processes such as comprehension or confusion. |
| Emotion (Face) | The Emotion displayed by a face. |
| Emotion (Speech) | The Emotion embedded in a speech segment. |
| Environment | A Physical or Virtual Space containing a Scene. |
| Expression | The attribute indicating that a human conveys a particular Information |
| Expression (Face) | The Expression displayed by a face. |
| Expression (Gesture) | The Expression displayed by the movement of head, arms, hands and fingers. |
| Face | The portion of a 2D or 3D digital representation corresponding to the face of a human. |
| Grade | The intensity of an Emotion. |
| Identity | The label uniquely associated to a human or to an avatar. |
| Intention | Attributes characterising a question, such as purpose or scope. |
| Meaning (Text) | Information extracted from an input text such as syntactic and semantic information. |
| Meaning (Gesture) | Information extracted from head, arms, hands and fingers conveying a coded message. |
| Physical Gesture | A movement of the body or part of it, such as head, arm, hand and finger, often as a complement to a vocal utterance. |
| Scene | An Environment populated by humans and real objects or by avatars and virtual objects. |
| Scene Descriptors | The individual attributes of the coded representation of the objects in a scene, including their location. |
| 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. |
| Subword Lattice | A directed graph containing speech recognition candidates. |
| Text | A series of characters drawn from a finite alphabet. |
| Viewpoint | The point expressed as spatial coordinates and the direction expressed as θ,φ where a user is looking at the Environment. |
| Virtual Twin | An Audio-Visual Object in a Virtual Environment representing an object in a Real Environment. |
| Visual Object | Visual information with its metadata or a combination of Visual Objects. |
| Vocal Gesture | Utterance, such as cough, laugh, hesitation etc. Lexical elements are excluded. |
| Word Lattice | A directed graph containing speech recognition candidates |

# 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: 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.

## Informative References

These references are provided as examples of technologies potentially relevant to this document. MPAI does not endorse any of these technologies as suitable for 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 WD0.10, N604

**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.

**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. https://docs.unity3d.com/Manual/AmbisonicAudio.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

# Use Case Architectures

## Conversation About a Scene

### Scope of Use Case

A human holds a conversation with a 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 Emotion (Speech and Face) and Expression (Face and Gesture).
4. The machine creates a synthetic version of the scene using its internal representation of the audio-visual scene in order to help the human understand how it sees the scene.
5. The human converses with the machine and indicates the object of their interest in his/her Environment.
6. The machine understands which object the human is referring to and generates an avatar
   1. Uttering speech containing emotion that is a pertinent with the human speech and face, and
   2. Showing a face having emotion and expressions that are pertinent to the human speech and face and congruent with the response it is uttering.

### Reference architecture

The Machine operates according to the following workflow:

1. Speech Recognition produces Recognised Text and Emotion-Speech from Input Speech.
2. Language Understanding extracts Emotion-Text from LangUnd-Text.
3. Video Analysis1 extracts Emotion-Face from Input Video.
4. Emotion Fusion fuses Emotion-Text, Emotion-Speech and Emotion-Face.
5. Video Analysis3 creates Visual Scene Descriptors from Input Video.
6. Video Analysis2 creates Gesture Descriptors from Visual Scene Descriptors.
7. Object Identification produces Object ID by combining Visual Scene Descriptors and Gesture Descriptors.
8. Language Understanding Extracts Meaning from Text and ObjectID.
9. Question Analysis produces Intention from Meaning.
10. Scene to Video produces the 2D video as seen from the user-selected Viewpoint from the Scene Descriptors where the objects are labelled with the machine’s understanding of the semantics of the objects.
11. Dialogue Processing produces the machine’s Output Text, Emotion (Speech) and Emotion (Face) from Fused Emotion, Meaning LangUnd Text and Intention.
12. Speech Synthesis (Emotion) produces Output Speech from Output Text and Emotion (Speech).
13. Face Animation (Emotion) produces Output Speech and Emotion Face (Face).

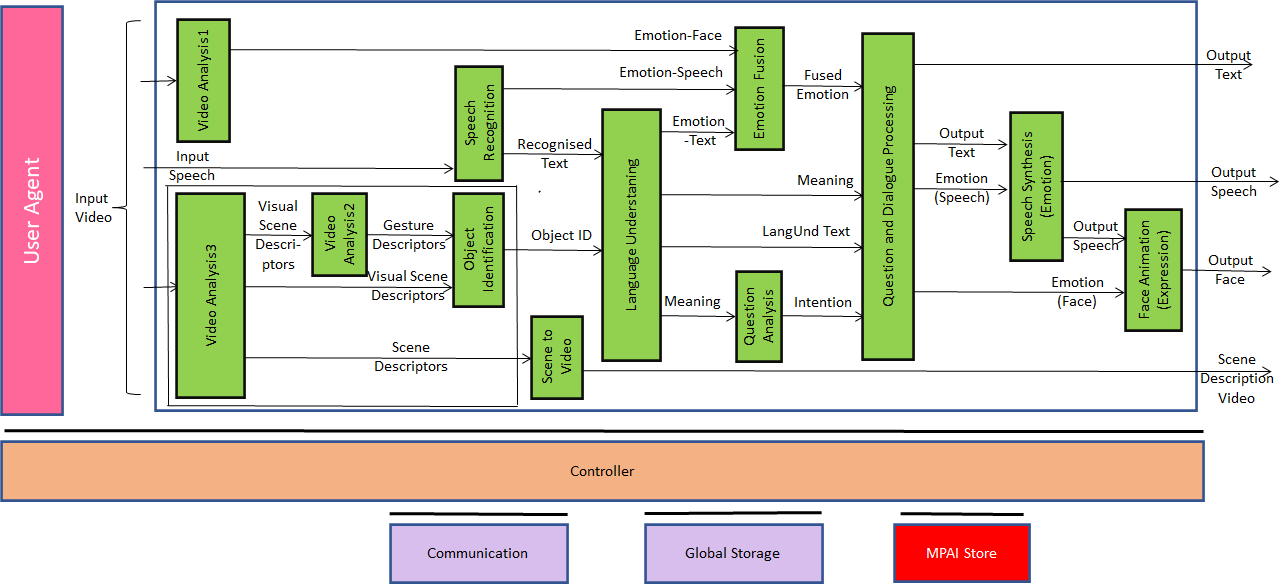


Figure 2 - Reference Model of Conversation About a Scene

### 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 |
| Scene to Video Viewpoint | Human | The point of view of the scene presented by Scene Description Video |
| **Output data** | **To** | **Comments** |
| Output Text | Human | Text version of machine’s speech |
| Output Speech | Human | Machine’s speech |
| Output Face | Human | Machine’s face |
| Scene Description Video | Human | Video reproduction of the scene as seen by machine with labelled objects |

### 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-Face |
| Video Analysis2 | Receives Visual Scene Descriptors  Extracts Gesture Descriptors |
| Video Analysis3 | Receives Input Video  Creates Visual Scene Descriptors |
| Speech Recognition | Receives Input Speech  Extracts Recognised Text from Input Speech  Extracts Emotion-Speech from Input Speech |
| Object Identification | Receives Gesture Descriptors and Scene Descriptors  Produces ID of scene object the human indicates |
| Scene to Video | Receives Visual Scene Descriptors  Received Viewpoint  Converts Scene Descriptors to a Video of the Scene from a the selected Viewpoint compatible with Input Video. |
| Language Understanding | Receives Recognised Text and Object ID  Produces Emotion-Text, Meaning and LangUnd-Text |
| Emotion Fusion | Receives Emotion-Text, Emotion-Speech and Emotion-Face  Produces a single Emotion that fuses the 3 Emotions |
| Question Analysis | Receives Meaning  Produces Intention |
| Dialogue Processing | Receives Meaning, Fused Emotion, LandUnd-Text and Intention  Produces Output Text, Emotion (Speech) and Emotion (Face) |
| Speech Synthesis (Expression) | Receives Output Text  Receives Emotion (Speech)  Produces Output Speech |
| Lips Animation | Receives Output Speech and Emotion (Face)  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 displace itself based on 1) analysis and interpretation of the data sensed by a range of on-board sensors exploring the environment and 2) information transmitted by other sources in range, e.g., other CAVs, traffic lights and roadside units.

Human-CAV Interaction assumes that the CAV is impersonated by an avatar, whose model is selected/produced by the CAV rights-holder (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 Emotion that would be displayed by a human driver in similar conditions.

This use case includes the functions:

1. Outside the CAV, when a group of humans approaches the CAV:
   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 (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 information.
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 radio, call a person, open window or door, search for information etc.

Note: For completeness, Figure 3 includes the conversion of human commands and response from the CAV, even though this document does not address the HCI-AMS (Autonomous Motion Subsystem) interaction.

* 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 senses the humans’ speech, face and gesture and manifests itself as face and shoulder of an avatar able to:
   1. Utter speech with appropriate Emotion.
   2. Display Emotions and Expressions in the face.
   3. Animate lips in sync with the speech and in line with Emotion (Face).
   4. Gazing at the human it is conversing with.
2. Assuming that the passengers allow the HCI to do so, and the processing is carried out inside the CAV, the HCI may follows the conversation on travel matters held by humans in the cabin during the travel.
3. Issue regarding the HCI’s 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.

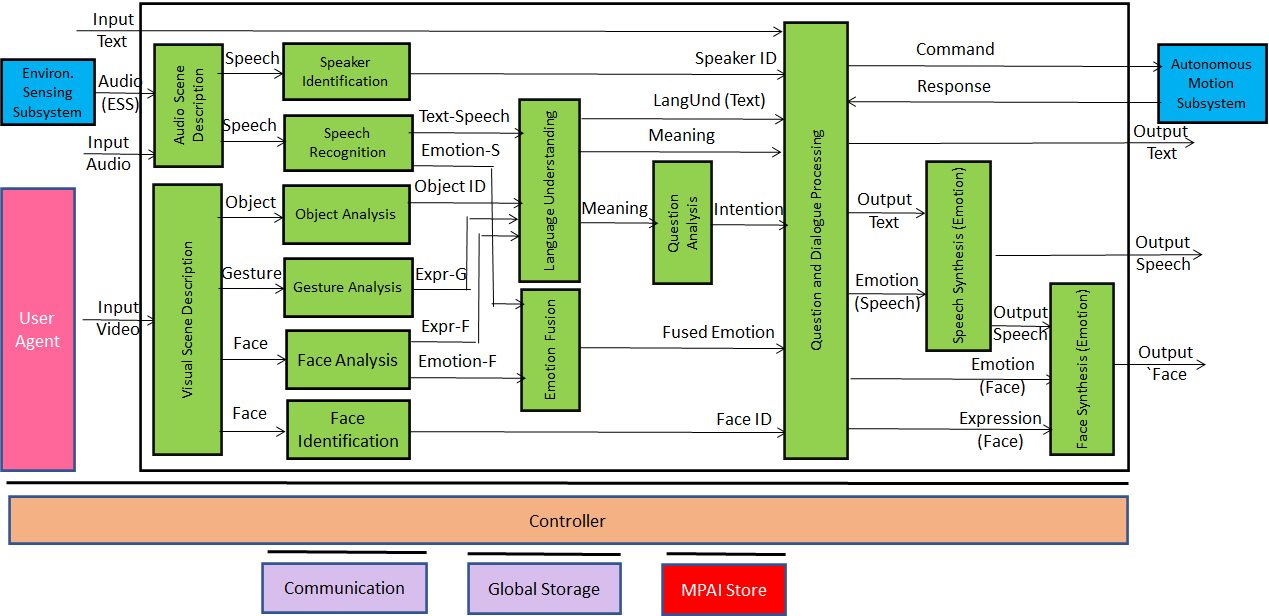


Figure 3 - Human-CAV Interaction Reference Model

HCI operates in two modes:

1. Outdoor: when humans are in the environment and approach the CAV:
   1. HCI separates and locates
      1. The human faces from the environment.
      2. The human speech from the environment sound.
   2. HCI identifies the humans using face and speech.
   3. HCI converses with humans on “travel-by-car”-related matters.
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 identifies the humans.
   4. HCI extracts
      1. Emotion from Speech and Face
      2. Expression from 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 having:

1. Speech containing emotion appropriate to the conversation.
2. Face having
3. Lips moving synchronously with the speech.
4. Emotion and Expression consistent with Text and 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 | Cabin Passengers | User’s social life  Commands/interaction with CAV |
| Audio | Cabin Passengers | User’s social life  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 response to passengers |
| Output Text | Cabin Passengers | 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 | Provides the Audio Scene Descriptors in the outdoor and indoor scenarios. |
| Speaker identification | Recognises speaker. |
| Speech recognition | 1. Analyses the Speech inputs. 2. Produces Text and Emotion-Text. |
| Visual Scene Description | Provides the Visual Scene Descriptors in the outdoor and indoor scenarios. |
| Object analysis | 1. Receives non-human objects. 2. Produces ID of objects in focus. |
| Gesture analysis | 1. Receives human objects. 2. Produces Emotion and Expression in Physical Gesture. |
| Face identification | 1. Receives Face Objects. 2. Recognises the identities of the humans. |
| Face analysis | 1. Receives Face Objects. 2. Extracts Emotion-Face in Face Objects. |
| Language understanding | 1. Analyses text in natural language using an embedded language model. 2. Analyses Expression in Face and Physical Gesture. 3. Produces LangUnd-Text. 4. Produces Meaning using Text, ObjectID, Expression in Face and Gesture. |
| Emotion Fusion | Produces Final Emotion by fusing Emotion-Speech and Emotion-Face. |
| Question analysis | Determines Intention by analysing the Meaning. |
| Question & Dialogue Processing | 1. Receives Speaker ID and Face ID. 2. If speaker ID and face ID match, then:    1. Analyses Text-LangUnd-Text, Fused Emotion, Intention, and Meaning.    2. Produces Emotion (Speech), Emotion (Face), Expression (Face) and Output Text. 3. Else, responds appropriately. |
| Speech synthesis | Converts Output Text and Emotion (Speech) to Output Speech. |
| Face animation | Converts Expression (Face), Emotion (Face) and Output Speech to Output Face. |

## 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 Space 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, answer questions, etc.

The space where the collaboration takes place is called Environment. It 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.

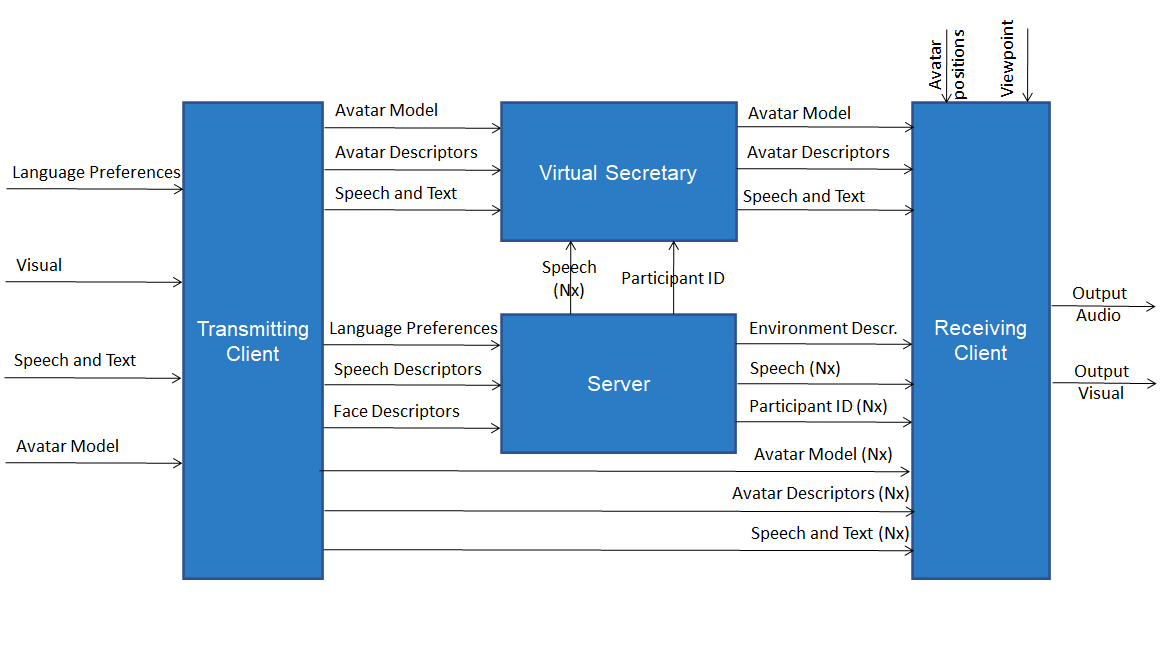


Figure 4 - Avatar-Based Videoconference end-to-end diagram

With reference to Figure 4, these are the elements chracterising the Use Case:

1. A group of geographically dispersed humans intend to participate in a videoconference. Groups of humans may be in the same Physical Environment; however, they have an individual participation in the Virtual Environment where the avatar-based videoconference takes place.
2. The videoconference room is the Virtual Environment os 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 sends to the Server using the transmitting part of their client (Transmitting Client):
   1. The participant’s avatar model, language preference and speech.
   2. The Descriptors of the participant’s face, head, arms, hands and fingers augmented by the participant’s Text, Speech, Meaning, Emotion from from speech and face and Expression from face and gesture.
5. The Server
   1. Authenticates all participants using speech and face information.
   2. Sends the Descriptors of the objects in the Environment, e.g., walls, furniture etc. to all participants.
   3. Uses the language preferences to translate all speech signals to the requested languages.
   4. Sends the original and translated speech signals to participants according to their language preferences.
   5. Forwards the avatar models and descriptors to all participants.
6. The Virtual Secretary
   1. Collects the statements made by participating avatars.
   2. Monitors the avatars’ emotions in their speech and face, and expression in their face and gesture.
   3. Makes a summary by combining information from a. and b.
   4. Displays the summary in the Environment so that avatars can read and
      1. Edit the summary.
      2. Comment on via Text and Speech.
   5. Edits the summary based on avatars’ sentences/questions about the summary expressed with text, speech and face emotion, and face and gesture expression.
   6. Converses with avatars via
      1. Text with embedded emotion.
      2. Speech with embedded emotion.
      3. Face displaying emotion and expression.
      4. Gesture displaying expression.
7. Each participant uses the receiving part of the client (Receiving Client) to:
   1. Place all avatar models at a selected chair.
   2. Animate the avatar models with the received descriptors.
   3. Attach the speech signals to the mouths of the corresponding avatars, so that the speech is properly located in the Environment.
   4. Select a Point of View, possibly different from the position assigned to his/her avatar in the Environment.
   5. Watch the 3D visual space 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).

### Transmitting Client AIW

#### Function

The function of the Transmitting Client AIW is to:

1. Receive:
   1. Input Audio (including Audio field created by Receiving Client) from microphone (array).
   2. Video of participant from camera (array).
   3. Participant’s Avatar Model.
   4. Participant’s spoken language preferences (e.g., EN-US, IT-CH).
2. Send the following data 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 Server:
   1. Language preferences
   2. Avatar model.
2. The following AIMs of Client Transmitting send to Server:
   1. Speaker Identification: Descriptors for Authentication.
   2. Face Identification: Face Descriptors for Authentication.

During the meeting the following AIMs of each participant’s Transmitting Client produce the following data:

1. Audio Scene Description: participant’s speech.
2. Speech Recognition: Recognised Text and Emotion (Text).
3. Visual Scene Description: Object, Gesture and Face.
4. Emotion Analysis: Emotion (Speech).
5. Face Analysis1: Emotion (Video).
6. Face Analysis2: Face Descriptors.
7. Head Analysis: Head Descriptors.
8. Gesture Analysis: Gesture Descriptors.
9. Language Understanding: Meaning.
10. Emotion Fusion: Fused Emotion.
11. Avatar Description: Avatar Descriptors.

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

1. Speech.
2. Avatar Descriptors.

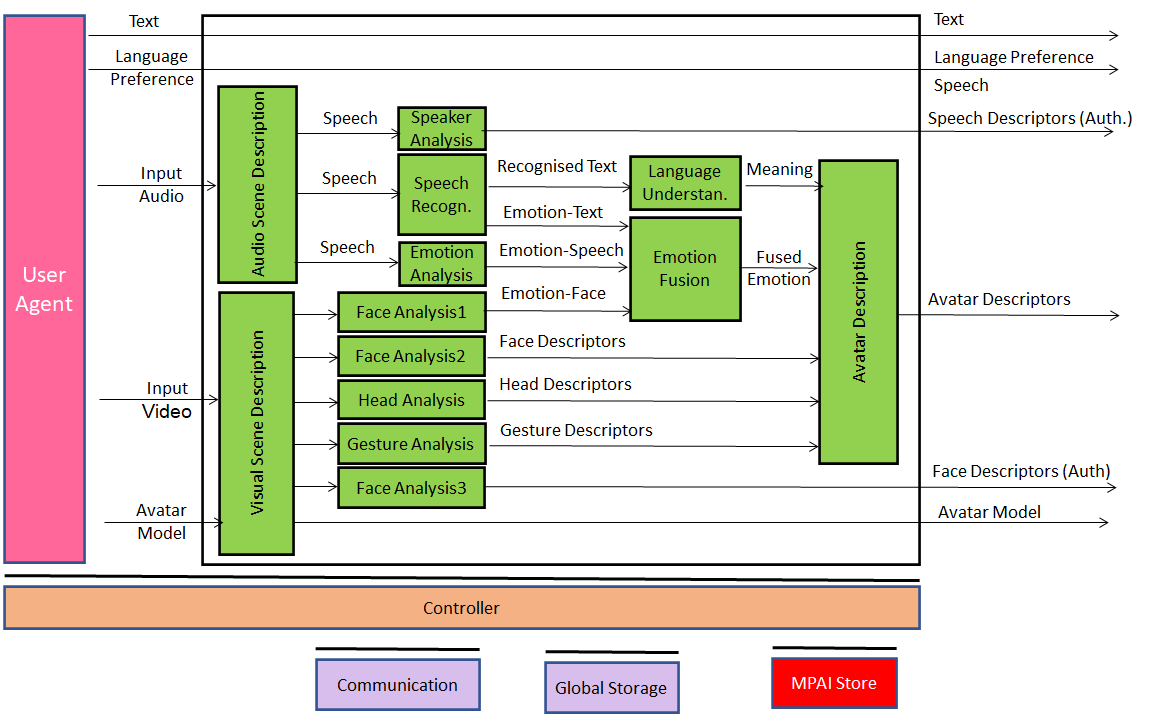


Figure 5 - Reference Model of Avatar Videoconference Transmitting Client

#### Input and output data

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

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

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

#### AI Modules

Table 7 gives the AI Modules of Transmitting Client AIW.

*Table 7 – AI Modules of Transmitting Client AIW*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Audio Scene Description | Provides Speech out of Input Audio. |
| Speaker Analysis | Provides Speaker’s Speech Descriptors for Authentication |
| Speech Recognition | Provides Text and Emotion from Speech. |
| Visual Scene Description | Provides Visual Descriptors of Scene. |
| Face Analysis1 | Provides Emotion from Input Video (face). |
| Face Analysis2 | Provides Face Descriptors for reproduction of participant’s face on avatar. |
| Head Analysis | Provides Head Descriptors for reproduction of participant’s head on avatar. |
| Gesture Analysis | Provides Gesture Descriptors for reproduction of participant’s arms, hands and fingers on 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. To Authenticate Participants.
   2. To Create Environment.
   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. Speaker Descriptors (Auth.).
      2. Face Descriptors (Auth.).
      3. Language Preferences.
      4. Speech.
      5. Avatar Descriptors.
2. Authenticates Participants.
3. Translates Participants’s 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.

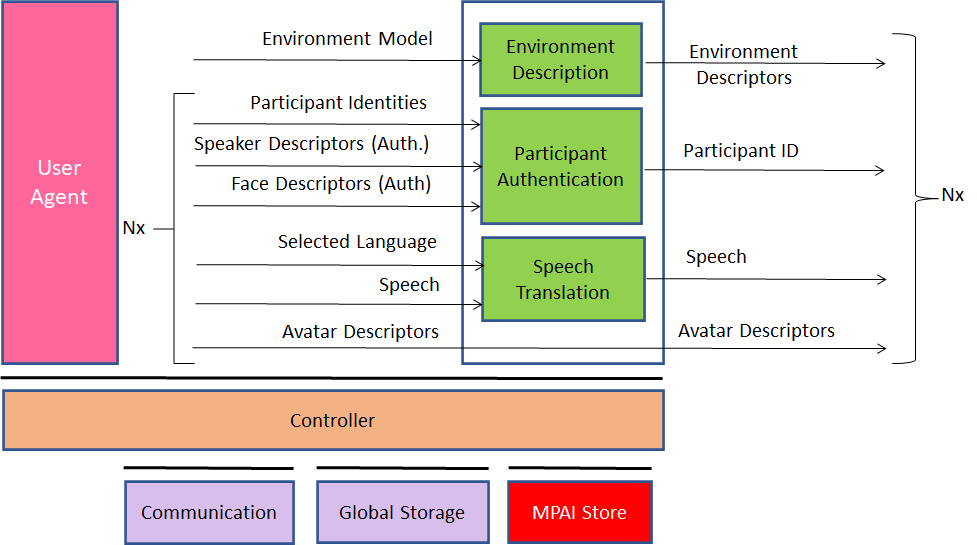


Figure 6 - Reference Model of Avatar-Based Videoconference Server

#### 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** |
| Environment 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** |
| Environment Descriptors (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 9 gives the AI Modules of Server AIW.

*Table 9 – 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

Figure 7 depicts the architecture of the Virtual Secretary AIW.

The primary role of the Virtual Secretary is to listen to the speech of each avatar, monitor their emotion and expression and draft a summary using text and emojis in a common language decided outside of the Use Case.

The summary may be:

1. Transferred to an external application allowing participants to edit the summary or allow avatars to make verbal comments or make offline text comments.
2. Corrected by the Virtual Secretary, who interprets speech and text complemented by the avatars’ emotion and expression and corrects the summary.

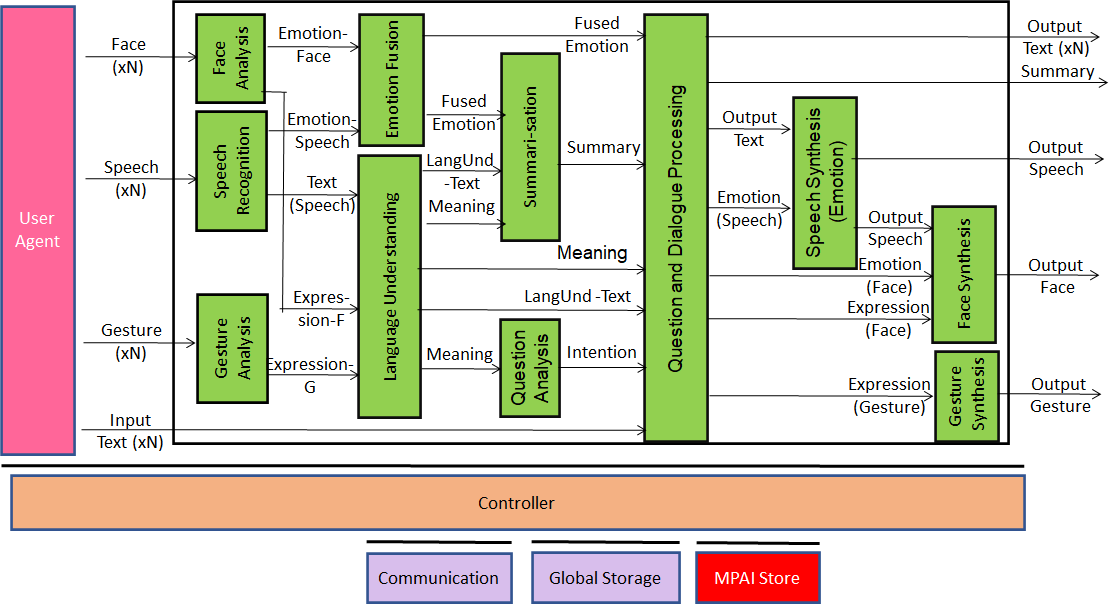


Figure 7 - Reference Model of Virtual Secretary

The operation of the Virtual Secretary workflow is described as follows:

1. The Virtual Secretary recognises the speech of the avatars.
2. The Speech Recognition and Face Analysis provide the emotion in the avatars’ speech and face.
3. Emotion Fusion provides a single emotion from the two emotions.
4. Gesture Analysis extracts the gesture expression.
5. Language Understanding
   1. Receives
      1. Recognised text.
      2. Expression of face.
      3. Expression of gesture.
   2. Creates
      1. Final version of recognised text (LangUnd-Text).
      2. Meaning of the sentence uttered by an avatar.
6. Question analysis extracts the intention from the meaning of the sentence uttered by an avatar.
7. Summarisation
   1. Receives:
      1. LangUnd-Text.
      2. Fused Emotion.
      3. Meaning.
   2. Creates Summary using text and emojis in the meeting’s common language
8. Question and Dialogue Processing (QDP)
   1. Receives
      1. LangUnd-Text.
      2. Text provided by a participant via chat.
      3. Meaning.
      4. Intention
      5. Summary.
   2. Creates
      1. Emotion contained in the text and in the speech to be synthesised.
      2. Emotion to be displayed in the Virtual Secretary’s face.
      3. Expression to be displayed in the Virtual Secretary’s face and gesture.
9. Speech Synthesis (Emotion)
   1. Receives QDP’s text and emotion in speech
   2. Synthesises the Virtual Secretary’s speech with its embedded emotion.
10. Face Synthesis
    1. Receives:
       1. Virtual Secretary’s synthetic speech.
       2. QDP’s face emotion and expression.
    2. Synthesises Virtual Secretary’s face.
11. Gesture Synthesis:
    1. Receives QDP’s gesture expression.
    2. Synthesises Virtual Secretary’s head, arms, hands and fingers.

#### I/O Data

*Table 10 – I/O data of Virtual Secretary*

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

#### AI Modules

Table 11 gives the AI Modules of the Virtual Secretary depicted in Figure 7.

*Table 11 – AI Modules of Virtual Secreatary*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| Speech recognition | 1. Analyses the Speech inputs. 2. Produces text and Emotion-Text. |
| Gesture analysis | 1. Receives avatars as a visual object. 2. Produces Expression-Gesture in Physical Gesture. |
| Face analysis | 1. Receives Face Objects. 2. Extracts Emotion-Face and Expression-Face in Face Objects. |
| Language understanding | 1. Analyses text in natural language using an embedded language model. 2. Analyses Expression-Gesture and Expression-Face. 3. Produces LangUnd-Text. 4. Produces the Meaning of the combination of LangUnd-Text and Expression in Face and Physical Gesture. |
| Emotion Fusion | Produces Final Emotion by fusing Emotion (Speech) and Emotion (Face). |
| Question analysis | Determines Intention by analysing Meaning. |
| Question & Dialogue Processing | 1. Analyses LangUnd-Text, Fused Emotion, Intention, and Meaning. 2. Produces Emotion (Speech), Emotion (Face), Expression (Face), Expression (Gesture) and Output Text. |
| Speech synthesis | Converts Output Text and Emotion (Speech) to Output Speech. |
| Face synthesis | Converts Emotion (Face), Expression (Face) and Output Speech to Output Face. |
| Gesture Synthesis | Converts Expression (Gesture) to Output Gesture |

### Receiving Client AIW

#### Function

The Function of the Receiving Client AIW is to:

1. Create the Environment Using Environment Descriptors.
2. Place Avatars at selected positions based on participant’s input.
   1. Using Avatar Descriptors.
   2. Adding Speech to each Avatar.
3. Present Environment, and animated Avatars with their speech.
4. Select the viewpoint in the Audio-Visual Scene based on participant’s input.

#### Architecture

The Receiving Client AIW:

1. Creates the AV Scene using:
   1. The Environment descriptors.
   2. The Avatars 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 Client Receiving AIW.

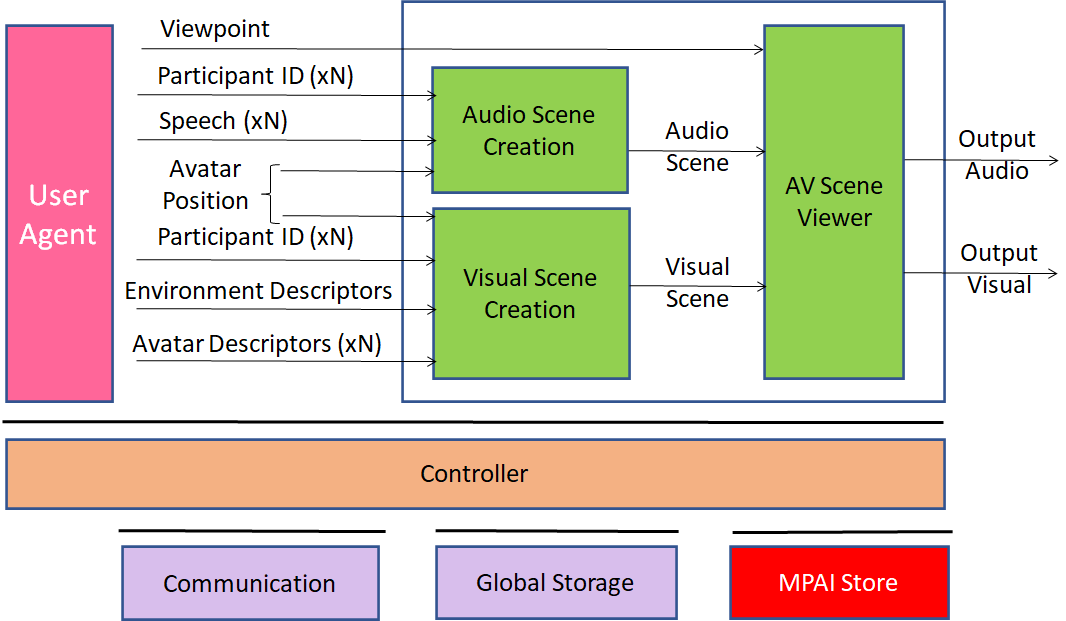


Figure 8 - Reference Model of Avatar-Based Videoconference Receiving Client

#### I/O Data

Table 12 gives the input and output data of Receiving Client AIW.

*Table 12 – 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 Avatars body animation. |
| **Output** | **Comments** |
| Output Audio | Presented using loudspeaker (array). |
| Output Visual | Presented using 2D or 3D display. |

#### AI Modules

Table 13 gives the AI Modules of Receiving Client AIW.

*Table 13 – 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. |

# Functional Requirements

## I/O Data summary

### Conversation About a Scene

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

*Table 14 – 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 Identification | Gesture Descriptors  Scene Descriptors | Object ID |
| Scene2Video | Scene Descriptors | Output Scene  Video |
| Language Understanding | 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 15 – I/O data of Human-CAV Interaction AIMs*

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

### Avatar Videoconference

#### Transmitting 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 Transmitting Client AIW*

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

#### Server

For each AIM (1st column), Table 17 gives the input (2nd column) and the output data (3rd col-umn).

*Table 17 – AIMs and Data of Server AIW*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| Environment Description | Environment Type | Environment Descriptors |
| Participant Authentication | Participant Identities (xN)  Speech Descriptors (Auth.) (xN)  Face Descriptors (Auth.) (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 18 gives the input data (2nd column) and the output data (3rd col-umn).

*Table 18 – AIMs and Data of Server AIW*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| Speech Recognition | Speech | Emotion-Speech  Text-Speech |
| Gesture Analysis | Gesture | Meaning |
| Face Analysis | Face | Emotion-Face |
| Emotion Fusion | Emotion (Speech)  Emotion (Face) | Fused Emotion |
| Language Understanding | Text (Speech) | LangUnd-Text  Meaning |
| Question analysis | Meaning | Intention |
| Question and dialogue processing |  | Output Text  Emotion (Speech)  Emotion (Face)  Summary |
| Speech synthesis | Output Text  Emotion-Speech | Output Speech |
| Face animation | Emotion-Face  Output Speech | Output Face |

#### Receiving Client

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

*Table 19 – 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 |

## Data format requirements

### Environment Descriptors

Environment Descriptors include the Descriptors required to represent the static components of the Virtual Environment where the Use Case takes place:

1. Visual Objects, e.g., table, swivel chair, walls and furniture, and their coordinates.
2. Audio Objects: e.g., audio source at given coordinates in the Virtual Environment.

The Descriptors should describe:

1. The elements bounding the Environment (walls, ceiling, floor, doors., windows).
2. Visual objects in the Environment (e.g., table and chairs).
3. Audio objects (e.g., speech originating from a given point in the Virtual Environment).
4. Integration of static avatars.
5. Association of Audio Objects to Visual Objects, e.g., a Speech to an Avatar.

**To Respondents**

MPAI requests respondents to propose a set of Audio and Visual Descriptors 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.

### Environment Model

The Environment Model is the file format of the Virtual Environment expressed, e.g., as Industry Foundation Classes (IFC)/BIM [12.] files for the visual component and static audio sources for the Audio component.

**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 part of the Environment Model.

### 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

Audio Scene Descripors are the descriptors of the structured composition of Audio Objects in two different modalities:

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 cabin. The CAV separates the different sound sources (speech and other sounds) and spatially locates them.
   2. A group of humans is sitting in a room. A microphone array describes the audio scene in speech and other sounds, and their locations.
2. The Audio Scene is used to create a sound field in a virtual environment, e.g.:
   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 sound from a place of given coordinates.

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 descrived above and other Use Cases of this document.

### Audio-Visual Scene Descriptors

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

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

**To Respondents**

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

### Avatar Descriptors

Avatar Descriptors are Descriptors that can be used to animate an Avatar Model based on the description of the movements of:

1. The surface 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

Avatar Model is the model of an avatar whose lower part (from the waist down) need not be represented. The proposed Avatar Model should allow a user to use the Avatar Descriptors to:

1. Express one of the MPAI standardised Emotions on the face of the avatar.
2. To animate the lips of an avatar in a way that is congruent with the speech it utters, its associated Emotion and the Emotion it expresses on the face.
3. Animate head, arms, hands and fingers, to express one of the Gestures to be standardised by MPAI, e.g., to indicate a particular person or object or the movements required by the sign language.
4. Rotate the upper part of the avatar’s body, e.g., if the avatar turns to watch the avatar next to it.

The model should be extensible to respond to future MPAI Calls for Technology, e.g., if requesting that also the currently static parts of the body are animated (e.g., the avatar walks in the Virtual Environment).

**To Respondents**

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

### Avatar Position

A set of coordinate points corresponding to the positions of the avatars in the Environment selected by a participant.

**To Respondents**

MPAI requests respondents to propose data format for Avatar Positions.

### Emotion Stream

In [3] MPAI has standardised a set of basic Emotions and their semantics to represent emotion conveyed by text, speech and face.

In the Use Cases considered in this document, it is envisaged that an emotion in speech and face at a given time can be followed by another with a transition time of a specified duration as depicted in Figure 9.

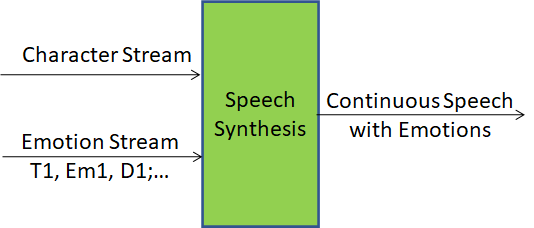


Figure 9 - Emotion sequences

**To Respondents**

Respondents are requested:

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, respondent is requested to motivate their claims, and propose an extension of the MPAI Basic Emotion Set or a new solution.
2. To propose a format able to represent dynamic emotions with the requirements oultlined above.

MPAI requests respondents to propose a format for the Emotion Stream.

### Emotion (Face)

In [3] MPAI has standardised a set of basic Emotions and their semantics to represent emotion conveyed by text, speech, and face.

The Use Cases considered in this document envisage that an emotion in speech and face at a given time can be followed by another emotion or lack of it with a transition time of a specified duration as depicted in Figure 9.

The dynamic emotion representation should allow for two different emotions to happen at the same time, possibly with different durations. In other terms, (T1, Em1, D1), (T1, Em2, D2) should be a syntactically correct expression. However, this may not correspond to a combination of emotions that a human being might display on their face.

Figure 10 depicts a possible implementation of a system able to synthesise speech and face from a stream of characters and two stream of speech and face emotions.

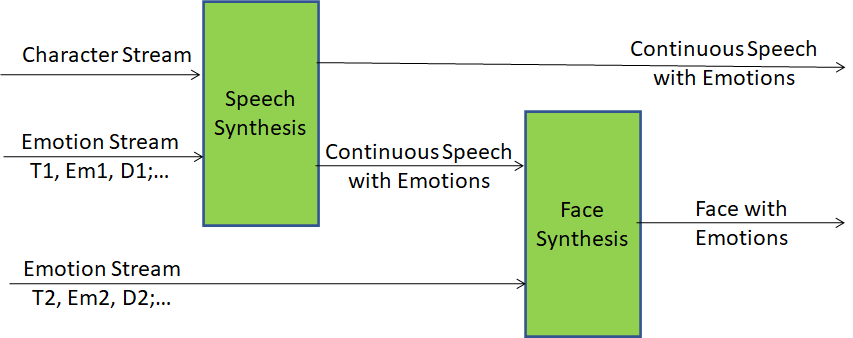


Figure 10 - Synthetic speech and face with emotions

The representation should allow for two separate streams and for one multiplexed streams.

**To Respondents**

Respondents are requested:

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, respondent is requested to motivate their claims, and propose an extension of the MPAI Basic Emotion Set or a new solution.
2. To propose a format able to represent dynamic emotions with the requirements oultlined above.

### Emotion (Speech)

See Emotion above.

### Emotion (Text)

See Emotion above.

### Face Descriptors

MPAI requests 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.
2. To describe the features of a face for the purpose of extracting:
   1. The physical features of a face.
   2. The facial Emotion and Expression.
3. To describe the movement of lips and eyes.
4. To animate the face of an avatar model using any of 2, 3, 4.

Note: movement of face, head, and gesture may be correlated.

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

**To Respondents**

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

### Face Objects

Face object is the 2D or 3D image of a face that can be used for different purposes such as:

1. To identify a human.
2. To extract Emotion and Expression from the face.
3. To enable an avatar to gaze at the human whose face object has been provided

The Face Object should contain the spatial coordinates of the face.

**To respondents**

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

### 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 Descriptors

MPAI requests Head Descriptors to describe:

1. The features of a head for the purpose of extracting:
   1. The physical features of a head.
   2. The Expression of the head.
2. The coordinates of the point representing the head.
3. The trajectory of the point representing a moving head
4. 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.

Note: movement of face, head, and gesture may be correlated.

**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 the speech signals captured by 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 to Translation

In [3] MPAI has standardised Translation of a spoken sentence from a language to another language preserving the colour (speech features) of the speaker, and the addition of emotion to the synthetic speech.

MPAI is now requesting a data format to be used as input format to a speech synthesiser satisfying the following requiremenst:

1. The text to be translated is represented as Time-stamped (Sub-)Word Lattices. As in the speech translation Use Case of MPAI-MMC-V1, the input to the translator is the output of a speech recogniser.
2. The text has an associated sequence of time-stamped Emotions each with duration of transitions between Emotions.

**To Respondents**

Respondents are requested to

1. To comment on the use of [3] to represent emotion.
2. If a respondent claims that the technology in [3] is unsuitable, to motivate their claims, and propose an extension of the MPAI Basic Emotion Set or a new solution.
3. Propose a format satisying the requirements above.

### 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 and physical gesture expression 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 a displayed object.

**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 defined 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.

### Speech Features

In all Version 2 Use Cases the user may need to select an avatar with a particular visual and speech model. Being able to do the latter, implies that the user can select the colour, i.e., the speech features of the speech the avatar will utter.

**To Respondents**

MPAI requests respondents to propose technology allowing a user to specify the colour of the synthetic speech generated by a machine and to communicate their choice to a third-party machine that will understand the selection unambiguously.

### Summary

Formatted characters with emojis suitable for human editing.

**To respondents**

Respondents are invited to propose a format.

### 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 the point expressed as coordinates from where a user looks at the space around them in a given direction (θ,φ).

**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 20 – 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 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 11 – The MPAI ecosystem operation*