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

MPAI Technical Specification

Human and Machine Communication
MPAI-HMC

V1.0

WARNING

Use of the technologies described in this Technical Specification may infringe patents, copyrights or intellectual property rights of MPAI Members or non-members.

MPAI and its Members accept no responsibility whatsoever for damages or liability, direct or consequential, which may result from the use of this Technical Specification.

Readers are invited to review Annex 2 - Notices and Disclaimers.

Technical Specification
Human and Machine Communication
(MPAI-HMC) V1.0

Contents

1 Introduction (Informative)

Artificial Intelligence (AI) has recently made great strides in offering more efficient ways to implement processes formerly carried out with Data Processing (DP) technologies. However, AI has often been used in an ad hoc way. Many machines using AI perform extremely complex functions, but the value of the result is known to depend on the training data sets, which are typically known only to the implementer. In certain applications – information services, for instance – this data issue may have potentially devastating social impacts due to biased training. In other applications – such as in autonomous vehicles – the issue is the lack of explainability: the inability to trace processes leading to a particular decision may likewise be unacceptable.

Data Processing standards have played a major role in promoting the wide use of digital technologies for products, services, and applications. However, few if any examples are known of AI standards with an approach comparable to that of DP standards. The MPAI organisation (Moving Picture, Audio, and Data Coding by Artificial Intelligence [13]) has taken on the mission of developing AI-based data coding standards. The group has already developed several Technical Specifications using AI Modules (AIMs) that attempt to break monolithic applications into components with known functions and interfaces and implementable using AI or DP technologies. By incorporating these modules, applications can be implemented as AI Workflows (AIWs), themselves with known functions and external interfaces, composed of AIMs interconnected according to a specified topology.

MPAI Technical Specifications offer two main advantages. The first is the ability to implement AI applications whose operation is more traceable and explainable. The second is the ability to create a competitive market of components – AIMs – with standardised functions and interfaces and potentially providing competitive performance.

MPAI has been pursuing this mission for several years. The group has developed *Technical Specification: Governance of the MPAI Ecosystem (MPAI-GME)* [1]. The MPAI Ecosystem is defined by the following elements:

1. The collections of Technical Specifications, Reference Software Specifications, Conformance Testing Specifications, and Performance Assessment Specifications jointly called Standard.
2. The MPAI Store in charge of making AIMs and AIWs available and providing Implementer Identifiers through its Implementer ID Registration Authority.
3. Implementers of MPAI Technical Specifications who have obtained an Implementer Identifier.
4. Performance Assessors, i.e., independent entities appointed by MPAI who assess the performance of implementations in terms of Reliability, Replicability, Robustness, and Fairness.

Another foundational Technical Specification is *Technical Specification: AI Framework (MPAI-AIF)* [2] enabling dynamic configuration, initialisation, and control of AIWs in a standard environment (AI Framework) depicted in Figure 1.

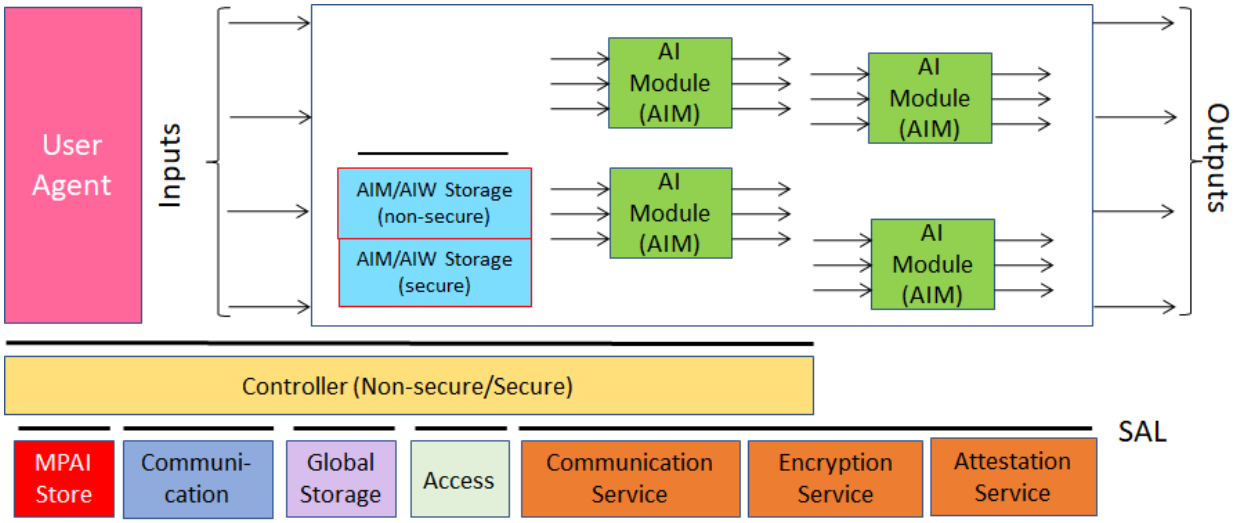

Figure 1 - The AI Framework (MPAI-AIF) V2 Reference Model

An Implementation of MPAI-AIF enables the secure execution of AIWs constituted by AIMs. AIMs can execute Data Processing (DP) or Artificial Intelligence (AI) algorithms and can be implemented in hardware, software, or hybrid hardware/software. They can be *Composite*, i.e., include interconnected AIMs.

Thus, MPAI specifications enable the implementation of applications whose internal operation end-users can understand to some degree, rather than machines that are just "black boxes" resulting from unknown training with unknown datasets. The developers of AIMs used in the AIWs can compete providing components with standard interfaces that can have improved performance compared to other implementations.

So far, MPAI has developed eight application-specific Technical Specifications on a wide range of application domains: context-enhanced audio [3], connected autonomous vehicles [13], audio enhancement [3], prediction of company performance [17], multimodal human-machine conversation [4], metaverse architecture [5], neural network watermarking [19], object and scene description [6], and portable avatars [7].

MPAI Technical Specifications are developed in compliance with a rigorous process [14] in service of the following policies:

1. While closely accommodating a given AI use case, so far as possible, remain agnostic to the technology – AI or DP – used in an implementation.
2. Facilitate the practical exploitation of Technical Specifications once adopted by MPAI.
3. Attempt to attract various industries, end users, and regulators.
4. Address three levels of standardisation, any of which an implementer can freely decide to adopt: the data exchanged by AIMs ("Data Types"), AIMs, and AIWs.
5. Specify the Data Types with clear, humanly understandable semantics, so far as possible.

This *Technical Specification: Human and Machine Communication (MPAI-HMC)* leverages five MPAI Technical Specifications: Context-based Audio Enhancement [3], MPAI Metaverse Model – Architecture [5], Multimodal Conversation [4], Object and Scene Description [6], and Portable Avatar Format [7], all of which deal with technologies enabling communication of real and digital humans in real or virtual environments. MPAI-HMC reproduces the normative elements from the five Technical Specifications that are relevant to this Technical Specification.

A Term beginning with a capital letter is defined in Table 1 if it is MPAI-HMC-specific or in Table 59 if its use extends across MPAI Technical Specifications. A term beginning with a small letter has the commonly intended meaning.

MPAI may extend this Version of MPAI-HMC with new technologies drawing from existing of new Technical Specifications.

Chapters, Sections, and Annexes are Normative unless they are explicitly identified as Informative.

2 Scope

Technical Specification: Human and Machine Communication (MPAI-HMC) – referred to in the following as MPAI-HMC – enables new forms of communication. The communicating participants are Entities, that is, either humans present in a real space or represented in a Virtual Space, or Machines represented in a Virtual Space or rendered in the real space as speaking avatars. The communicating participants act in a Context using text, speech, face, gesture, and the audio-visual scene in which they are embedded.

MPAI-HMC specifies the *Communicating Entities in Context* Use Case.

MPAI-HMC includes the following Chapters:
1. Scope
2. Definitions
3. References
4. Use Case
5. Functions
6. Reference Model
7. I/O Data
8. SubAIMs
9. JSON Metadata
10. Data Types.

Note that:
1. The SubAIMs Chapter of point 8. specifies Functions, Reference Model, and I/O Data of all AIMs.
2. If an AIM is a Composite AIM, its SubAIMs are specified in a hierarchical fashion.
3. All JSON Metadata are provided in a single Chapter.

3 Definitions

Terms beginning with a capital letter have the meaning defined in Table 1 or Table 59. Terms beginning with a small letter have the meaning commonly defined for the context in which they are used. For instance, Table 1 defines *Object* and *Scene* but does not define *object* and *scene*.

A dash "-" preceding a Term in Table 1 indicates the following readings according to the font:
1. Normal font: the Term in the table without a dash and preceding the one with a dash should be read before that Term. For example, "Avatar" and "- Model" will yield "Avatar Model."
2. *Italic* font: the Term in the table without a dash and preceding the one with a dash should be read after that Term. For example, "Avatar" and "- Portable" will yield "Portable Avatar."

Table 1 - General MPAI-HMC terms

Terms	Definitions
Attitude	
- *Social*	The coded representation of the internal state related to the way a human or avatar intends to position vis-à-vis the Environment or subsets of it, e.g., "Respectful", "Confrontational", "Soothing".
- *Spatial*	Position and Orientation and their velocities and accelerations of an Object in a Real or Virtual Environment.
Audio	Digital representation of an analogue audio signal sampled at a frequency between 8-192 kHz with a number of bits/sample between 8 and 32, and non-linear and linear quantisation. Data with characteristics of Audio may be synthetically produced.
Audio Block	A set of consecutive Audio samples.
Audio Channel	A sequence of Audio Blocks.
Avatar	An Object rendered to represent a Human of a Machine in a virtual space.
- Model	An inanimate Avatar exposing animation interfaces.
- *Portable*	A Data Type including Avatar ID, Time, Visual Environment, Spatial Attitude, Avatar Model, Body Descriptors, Face Descriptors, Language Preference, Speech Coding, Speech Data, Text, and Personal Status [6].
Body	A digital representation of a human body, head included, face excluded.
Centre Point	The point of an Object selected to have Local Coordinates (0,0,0).
Cognitive State	The coded representation of the internal state reflecting the way a human or avatar understands the Environment, such as "Confused", "Dubious", "Convinced".
Communication Item	An element generated by a Machine communicating with an Entity expressed with a Portable Avatar.
Context	The semantics of the information emitted by an Entity or included in its surrounding Scene.
Coordinate System	A coordinate system where the position of a point is specified by three numbers.
- *Cartesian*	A coordinate system where the three numbers are the signed distances from the point to three mutually perpendicular planes.
- *Spherical*	A coordinate system where the three numbers are: - the radial distance of that point from a fixed origin. - the polar angle measured from a fixed zenith direction. - the azimuthal angle of its orthogonal projection on a reference plane.
Culture	The collection of language and customs governing the way a human, or a group of humans employ to express their internal statuses.
Data	Information in digital form.
- Format	The standard digital representation of Data.
- Type	An instance of Data with a specific Data Format.
Descriptor	The Digital Representation of a feature of an Object.
- *Body*	A Data Type including the digital representation of the features of the body of a real or digital human.
- *Face*	A Data Type including the digital representation of a feature of the face of a real or digital human.
- *Speech*	A Data Type representing a variety of information elements incorporated in a Speech Segment, e.g., personal identity, Personal Status, additional factors such as vocal tension, creakiness, whispery quality, etc.

- *Text*	A Data Type including the digital representation of a feature of text.
Digital Representation	Data corresponding to and representing a physical entity.
Emotion	The coded representation of the internal state resulting from the interaction of a human or avatar with the Environment or subsets of it, such as "Angry", "Sad", "Determined".
Entity	A Human of a Machine communicating with a Machine.
Environment	A Virtual Space that may be null or may include an Audio-Visual Scene.
Experience	The state of an Entity whose senses/sensors are continuously affected for a meaningful period.
Face	A digital representation of a human face.
Factor	One of Emotion, Cognitive State, and Attitude.
Gesture	A movement of a Digital Human or part of it, such as the head, arm, hand, and finger, often a complement to a vocal utterance.
Human	A human being in a real space.
- *Digital*	A Digitised or a Virtual Human in a Virtual Space.
- *Digitised*	An Object in a Virtual Space that has the appearance of a specific human when rendered.
- *Virtual*	An Object in a Virtual Space created by a computer that has a human appearance when rendered but is not a Digitised Human.
Identifier	The label uniquely associated with a human or an Object.
Instance	An element of a set of entities – Objects, Digital Humans etc. – belonging to some levels in a hierarchical classification (taxonomy).
- *Audio*	The instance of an Audio Object.
- *Visual*	The instance of a Visual Object.
Machine	An Implementation of MPAI-MMC.
Meaning	Information extracted from Text such as syntactic and semantic information, Personal Status, and other information, such as an Object Identifier.
Microphone Array Geometry	A Data Type representing the position of each microphone comprising a microphone array and characteristics such as microphone type, look directions, and array type.
Modality	One of Text, Speech, Face, or Gesture.
Object	A data structure that can be rendered to cause an Experience.
- *Audio*	An Object described by Audio Descriptors.
- *Audio-Visual*	An Object described by Audio-Visual Descriptors.
- *Body*	A digital representation of the body of a Human or a Machine.
- Descriptor	The digital representation of the feature of an Object.
- *Digital*	A Digitised or a Virtual Object.
- *Digitised*	The digital representation of a real object.
- *Face*	The digital representation of the face of a Human or a Machine.
- *Speech*	An Object described by Speech Descriptors.
- *Text*	A string of Text.
- *Virtual*	An Object not representing an object in the real environment.
- *Visual*	An Object described by Visual Descriptors.
Orientation	The 3 Euler angles of an Object in a Virtual Space.
Personal Status	A Data Type including three Factors – Cognitive State, Emotion and Social Attitude – conveyed by four Modalities – Text, Speech, Face, and Gesture and providing standard extensible labels for the three Factors [4].

- *Face*	The Cognitive State, Emotion, and Social Attitude conveyed by a Face Object.
- *Gesture*	The Cognitive State, Emotion, and Social Attitude conveyed by the Gesture of a Body Object.
- *Speech*	The Cognitive State, Emotion, and Social Attitude conveyed by a Speech Object.
- *Text*	The Cognitive State, Emotion, and Social Attitude conveyed by a Text Object.
Position	The coordinates of a representative point for an object in a Virtual Space with respect to a set of coordinate axes.
Principal Axis	The x axis of an Object.
Rendering	The process of instantiating a Virtual Space as a human-perceptible entity.
Scene	A composition of Objects located according to a Scene Geometry.
- *Audio*	A Scene composed of Audio Objects.
- *Audio-Visual*	A Scene composed of Audio Objects, Visual Objects and co-located Audio-Visual Objects.
- *Multichannel*	A data structure containing at least 2 time-aligned interleaved Audio Channels.
- *Visual*	A Scene composed of Visual Objects.
Scene Descriptors	The digital representation of a feature of a scene.
- *Audio*	A Data Type including the digital representation of the audio features of a real or digital scene.
- *Audio-Visual*	A Data Type combining the Audio or Visual Scene Descriptors.
- *Visual*	A Data Type including the digital representation of the visual features of a real or digital scene.
Scene Geometry	The digital representation of the Object arrangement of a Scene.
- *Audio*	A Data Type describing the spatial arrangement of the Visual Objects of a Scene.
- *Audio-Visual*	A Data Type describing the spatial arrangement of the Audio, Visual, and Audio-Visual Objects of a Scene.
- *Visual*	A Data Type describing the spatial arrangement of the Visual Objects of a Scene.
Selector	Input Data having the goal to set a parameter (e.g., use of Text vs Speech or Language Preference) or an operating mode of a Machine.
Virtual Space	A space generated and maintained by a computing platform that can be rendered.
Speech	Digital representation of analogue speech sampled at a frequency between 8 kHz and 96 kHz with a number of bits/sample of 8, 16 or 24, and non-linear and linear quantisation or compressed. Data with characteristics of Speech may be synthetically produced.
Text	A sequence of characters represented according to [10].
- *Recognised*	The Text at the output of an Automatic Speech Recognition AIM.
- *Refined Text*	The Text at the output of a Natural Language Understanding AIM.
- *Translated Text*	The Text at the output of a Natural Language Translation AIM.

4 References

4.1 Normative References

1. Technical Specification; MPAI Ecosystem Governance (MPAI-GME) V1.1; https://mpai.community/standards/mpai-gme/.
2. Technical Specification; AI Framework (MPAI-AIF) V2; https://mpai.community/standards/mpai-aif/.
3. Technical Specification: Context-based Audio Enhancement (MPAI-CAE) V2; https://mpai.community/standards/mpai-cae/.
4. Technical Specification: Multimodal Conversation (MPAI-MMC) V2; https://mpai.community/standards/mpai-mmc/.
5. Technical Specification: MPAI Metaverse Model (MPAI-MMM) – Architecture V1; https://mpai.community/standards/mpai-mmm/.
6. Technical Specification: Object and Scene Description (MPAI-OSD) V1; https://mpai.community/standards/mpai-osd/.
7. Technical Specification: Portable Avatar Format (MPAI-PAF) V1; https://mpai.community/standards/mpai-paf/.
8. ITU-R; Long-form file format for the international exchange of audio programme materials with metadata; BS.2088-1 (10/2019) https://www.loc.gov/preservation/digital/formats/fdd/fdd000001.shtml.
9. ISO 639; Codes for the Representation of Names of Languages – Part 1: Alpha-2 Code.
10. ISO/IEC 10646; Information technology – Universal Coded Character Set.
11. ISO/IEC 19774-1:2019 Information technology – Computer graphics, image processing and environmental data representation – Part 1: Humanoid animation (HAnim) architecture; see https://www.web3d.org/documents/specifications/19774-1/V2.0/index.html
12. Khronos; Graphics Language Transmission Format (glTF); October 2021; https://registry.khronos.org/glTF/specs/2.0/glTF-2.0.html

4.2 Informative References

13. MPAI; The MPAI Statutes; https://mpai.community/statutes/.
14. MPAI; The MPAI Patent Policy; https://mpai.community/about/the-mpai-patent-policy/.
15. MPAI; Framework Licence: Human and Machine Communication (MPAI-HMC).
16. Technical Specification: Connected Autonomous Vehicle (MPAI-CAV) – Architecture V1; https://mpai.community/standards/mpai-cav/.
17. Technical Specification: Compression and Understanding of Industrial Data (MPAI-CUI) V1; https://mpai.community/standards/mpai-cav/.
18. Technical Specification: MPAI Metaverse Model (MPAI-MMM) – Architecture V1; https://mpai.community/standards/mpai-mmm/.
19. Technical Specification: Neural Network Watermarking (MPAI-NNW) V1; https://mpai.community/standards/mpai-nnw/.
20. Ekman, Paul (1999), "Basic Emotions", in Dalgleish, T; Power, M (eds.), Handbook of Cognition and Emotion (PDF), Sussex, UK: John Wiley & Sons.

5 Use Case

5.1 Communicating Entities in Context

The *Communicating Entities in Context* (HMC-CEC) Use Case involves 1) a human in a real audio-visual scene or the Digital Human representation of a machine in an Audio-Visual Scene and

2) another human in a real audio-visual scene or the Digital Human representation of a Machine in an Audio-Visual Scene.

A Machine
1. Receives Communication Items from other Machines.
2. Captures
    a. Audio-visual scenes containing communicating humans.
    b. Audio-Visual Scenes containing Digital Humans.
3. Understands the information emitted by the Entity including its Context.
4. Produces a response based on the understood information.
5. Produces
    a. Communication Item for use by other Machines.
    b. Audio-Visual Scenes containing a representation of itself.
6. Renders the Audio-Visual Scene.

MPAI-HMC assumes that:
1. Input Audio is Multichannel Audio and Input Visual is visual information in a format suitable for processing by the Visual Scene Description AIM.
2. Output Audio and Output Visual convey audio and visual information for rendering by the Audio-Visual Rendering AIM.
3. The real space where the human is located is digitally represented as an Audio-Visual Scene that may include other humans and generic objects.
4. The Virtual Space containing a Digitised or Virtual Human (collectively "Digital Humans") and/or its Audio components is represented as an Audio, Visual, or Audio-Visual Scene that may include other Digital Humans and generic Objects.
5. The Machine can:

5.1. Understand the semantics of the communicated information at different layers of depth.
5.2. Produce a multimodal response expected to be congruent with the received information.
5.3. Represent itself as a speaking humanoid immersed in an Audio-Visual Scene.

6. A Machine can convert the semantics of the Text, Speech, Face, and Gesture issued by an Entity in a Context to a form that is compatible with the Context of another Entity.
7. An AI Module is specified only by its Functions and Interfaces. Implementers are free to use their preferred technologies to achieve the Functions providing the features while respecting the constraints of the Interfaces.

The following Informative Section provides communication examples relevant to the HMC-CEC Use Case.

5.2 Usage Scenarios (Informative)

This Chapter includes five usage scenarios mostly described as particular cases of the combined usage scenarios of Figure 2 that combines some of the communication settings between Humans and Machines targeted by HMC-CEC. In Figure 2, the term Machine followed by a number indicates an HMC-CEC Implementation. A Machine can be an application, a device, or a function of a larger system. For the sake of simplicity, the Text component is not included in Figure 2, but is supported by HMC-CEC.

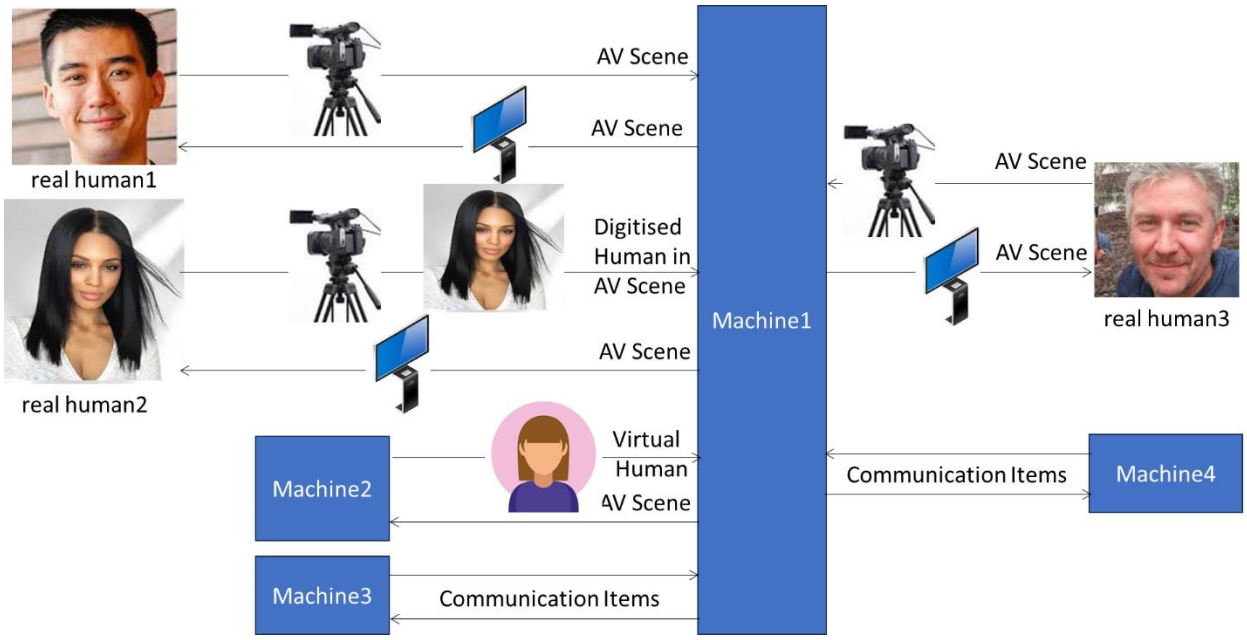

Figure 2 – Combined usage scenarios of HMC-CEC communication.

Figure 2 describes the following usage scenarios in which:

1. real human1 in his real environment and Machine1 communicate if:
1.1. real human1 emits audio-visual signals in an audio-visual scene that the sensors of Machine1 convert to Audio-Visual Scenes.
1.2. Machine1 generates Audio-Visual Scenes that its actuators convert to audio-visual signals.
2. real human1 and real human3 – belonging to different cultural environments – communicate if:
2.1. Both real humans emit audio-visual signals in audio-visual scenes that the sensors of Machine1 convert to Audio-Visual Scenes.
2.2. Machine1 converts (e.g., translates) the semantics of the Audio-Visual Scenes sensed from the audio-visual scenes where real human1 or real human 3 reside to those of the cultural environment of real human3 or real human 1, and generates Audio-Visual Scenes that its actuators convert to audio-visual signals.
3. real human1 and Machine4 communicate if:
3.1. real human1 emits audio-visual signals that the Sensors of Machine1 convert to Audio-Visual Scenes.
3.2. Machine1 converts the semantics of the Audio-Visual Scenes to Machine4's cultural environment and generates either Audio-Visual Scenes or Communication Items – called Communication Items – formatted according to the Portable Avatar Format [6].
3.3. Machine4 generates and emits either Visual Scenes or Communication Items in its own cultural environment.
3.4. Machine1 converts (e.g., translates) the semantics of Audio-Visual Scenes or Communication Items to the semantics of Audio-Visual Scenes in real human1's cultural environment, and emits Audio-Visual Scenes that real human1 can perceive.
4. real human2 in her real environment communicates with Machine 1 if:
4.1. real human2 locates her Digitised Human in a Virtual Environment, such as the one specified by the MPAI Metaverse Model – Architecture [5].
4.2. Machine1 perceives the Digitised Human in the Virtual Environment and generates a Virtual Human that real human2 can perceive. The Virtual Environment may use various means to enable real human2 to perceive the Virtual Environment.

5. Machine2 communicates with Machine1 if both Machines generate Virtual Humans in a Virtual Environment. Both Machines may communicate with the Digitised Human of point 2. above if all participants are in the same Virtual Environment.
6. Machine3 communicates with:

6.1. real human3 by using the same process as in point 2. above.
6.2. Machine4 by exchanging Audio-Visual Scenes or Communication Items.

Note that Communication Items may include a multimodal message (Text, Speech, Face, and Gesture), an associated Personal Status specifying Emotion, Cognitive State, and Social Attitude [4], language, and information about a Virtual Space [6].

5.2.1 Information Service

A human in a public space wants to access an information service implemented as a kiosk equipped with audio-visual sensors able to capture the space containing the human and processing functions to extract the human as an audio (speech) and visual object, while ignoring other humans and other audio and visual objects. The human may request information on an object present in the real space that the human indicates with a forefinger (see Conversation About a Scene in MPAI-MMC V2 [4]). The kiosk responds with a speaking avatar displayed by its actuators.
Figure 3 depicts the usage scenario using an appropriate subset of Figure 2.

Figure 3 - Information Service

5.2.2 Cross-Cultural Information Service

A frequently travelling human uses his portable HMC-enabled device (Machine1) designed and trained to capture the subtleties of that human's Culture. The human interacts with a local Information Service (Machine4) using Machine1 that acts as an interpreter between the human and Machine4 by exchanging Communication Items that may include the human's avatar.
Figure 4 depicts the usage scenario using the appropriate subset of Figure 2.

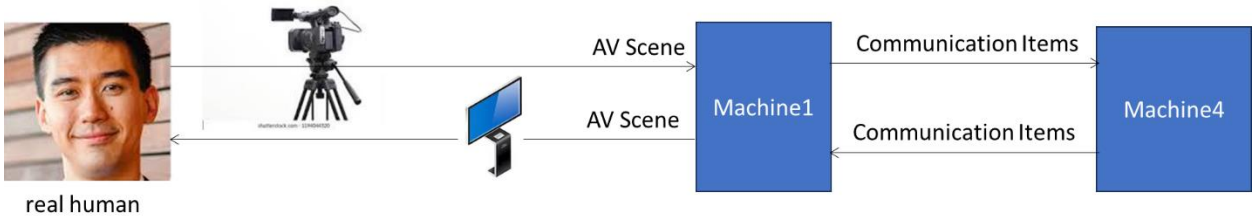

Figure 4 – Cross-Cultural Information Service

5.2.3 Virtual Assistant

This usage scenario has already been developed by MPAI-MMC V2 [4] and is used by the MPAI-PAF Avatar-Based Videoconference (PAF-ABV) [6], a videoconference whose participants are speaking avatars realistically impersonating the human participants. A speaking avatar not representing a human participant is the Virtual Secretary (generated by Machine2) which plays the role of note-taker and summariser by:

1. Listening to all Avatars' Speech.

2. Monitoring their Personal Statuses [4].
3. Drafting a Summary using the Avatars' Personal Status and Text, which may be obtained via Face and Body analysis, Speech Recognition, or Text input.

The Portable Avatar of the Virtual Secretary is distributed to all participants, who can then place it around the meeting table.

Figure 5 depicts two Digitised Human participants and one Virtual Human participant (Virtual Secretary). Machine 1 acts as cultural mediator between real human2 and real human3.

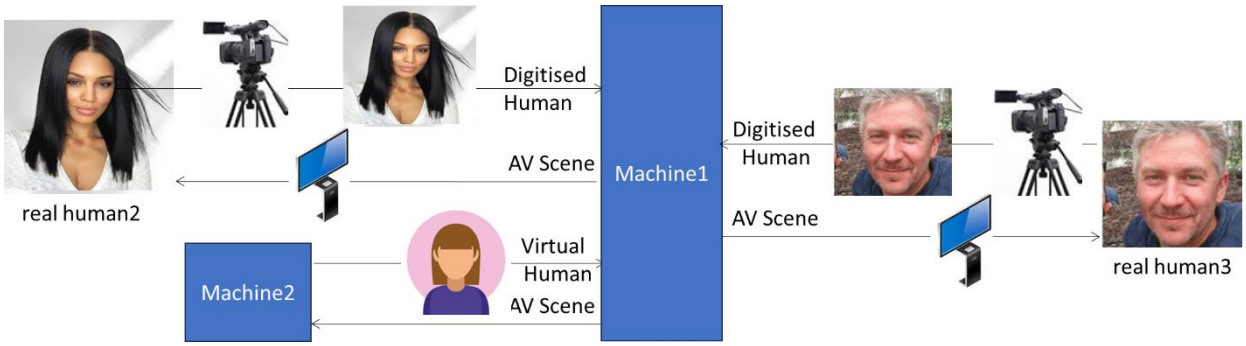

Figure 5 - Virtual Assistant

5.2.4 Conversation companion

A human is sitting in her living room wishing to converse about a topic with a Machine, represented and displayed as a speaking avatar. The human asks questions, and the Machine responds. The human displays pleasure, dissatisfaction, or other indications of Personal Status (including Emotion, Cognitive State, and Social Attitude). The Machine responds appropriately, with appropriate vocal and facial expressions.

Figure 6 illustrates the usage scenario.

Figure 6 - Conversation Companion

5.2.5 Strolling in the metaverse

$User_A$ – a Process representing a human in an M-Instance (a metaverse instantiation) rendered as a speaking Avatar – is in a public area in the M-Instance. She is approached by $User_B$, a Process rendered as an animated speaking Avatar representing personnel of a company promoting a particular product. $User_A$ does not reject the encounter. $User_B$ captures all relevant information from the speech, face, and body of $User_A$'s Avatar, and expresses itself by uttering relevant speech and appropriately moving its face and body. Eventually, $User_A$ gets annoyed and calls a security entity (Machine3 in Figure 7) that deals with the complaints of $User_A$, using audio and visual information as if it were representing a real human.

Figure 7 illustrates the usage scenario. Note that Machine1 includes the function that enables hosting of Digitised and Virtual Humans.

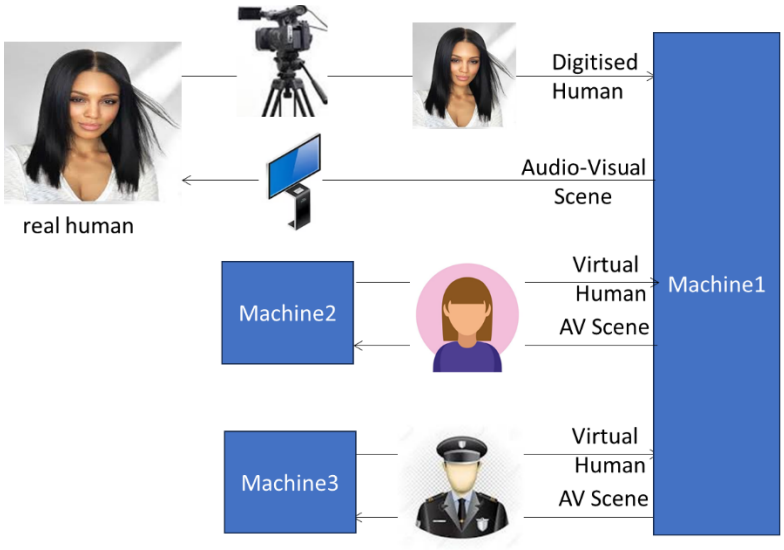

Figure 7 - Strolling in the metaverse

5.2.6 Travelling in a Connected Autonomous Vehicle

Two humans travel in a Connected Autonomous Vehicle (CAV) conversing with a Machine that performs some of the functions of the CAV's Human-CAV Interaction Subsystem [13]. The Machine is aware of the position of the human it is talking to at a particular time and directs its Avatar's gaze accordingly.

6 Functions

A Machine communicates with an Entity by performing the following high-level functions:

1. Receive a sequence of either:
   1.1. Audio-visual scenes or Audio-Visual Scenes that include the communicating Entity represented as Audio-Visual Scene Descriptors.
   1.2. Communication Items containing an Avatar representing the Entity communicating with the Machine and Audio-Visual Scenes represented as Audio-Visual Scene Descriptors.
2. Locate the Entity in the audio-visual scene or Audio-Visual Scene that it should communicate with and understand the information issued by the Entity and the Context where the Entity is embedded.
3. Produce and direct multimodal responses to the communicating Entity either by generating a Communication Item or an Audio-Visual Scene both of which may include itself.

7 Reference Model

The Reference Model of Communicating Entities in Context (HMC-CEC) depicted in Figure 8 implements an AI Workflow (AIW) with six AI Modules (AIM) conforming with *Technical Specification: AI Framework (MPAI-AIF)* [2]. *Annex 1 - MPAI Basics* provides an informative introduction to MPAI-AIF. The AIW receives input data processed by its AIMs and provides output data. Three AIMs in Figure 8 – Audio-Visual Scene Description, Entity and Context Understanding, and Personal Status Display – are Composite AIMs.

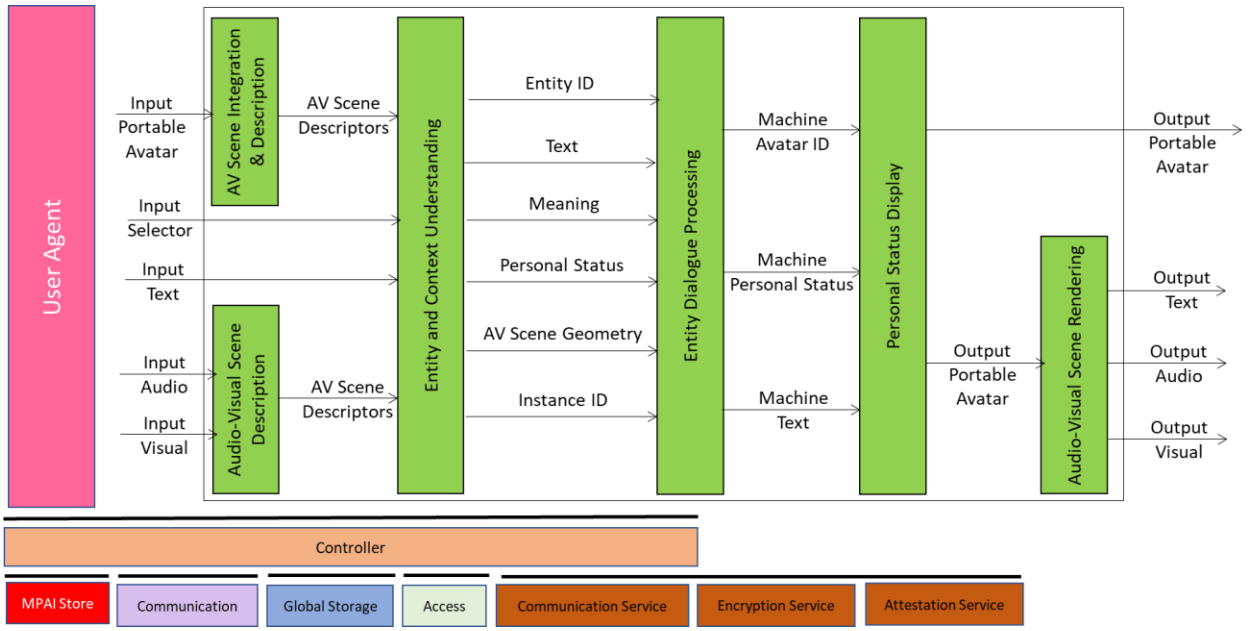

Figure 8 – Communicating Entities in Context AIW

Note that:

1. Input Selector enables the Entity to inform the Machine through the Entity and Context Understanding AIM about use of Text vs. Speech, Language Preferences, and Selected Language in translation.
2. Input Text, Input Speech, and Input Visual convey the information emitted by the Entity.
3. The Input Portable Avatar is the Communication Item received from a communicating Machine.
4. The Audio-Visual Scene Descriptors produced by the Audio-Visual Scene Description and Audio-Visual Scene Integration and Description AIMs are digital representations of a real audio-visual scene or a Virtual Audio-Visual Scene.

8 I/O Data of AIW

Table 2 specifies the Input and Output Data of the HMC-ECC AIW.

Table 2 – I/O Data of HMC-CEC AIW

Input	Description
Portable Avatar	A Communication Item emitted by the Machine.
Input Selector	Selector containing data that determines: 1. Whether an Entity uses Speech or Text as input. 2. Which language is used as input. 3. The target Language in translation.
Input Text	Text Object generated by Entity as information additional to or in lieu of Speech Object.
Input Audio	The audio scene captured by the Machine.
Input Visual	The visual scene captured by the Machine.
Output	**Description**
Portable Avatar	The Communication Item produced by the Machine.

9 SubAIMs

9.1 AV Scene Integration and Description (HMC-SID)

9.1.1 Functions

AV Scene Integration and Description (HMC-SID) performs the following functions:

1. Receives a Portable Avatar.
2. Adds the Avatar in the Input Portable Avatar to the Audio-Visual Scene conveyed by the Input Portable Avatar. If the Input Portable Avatar does not include a Scene, a generic Scene is used.
3. Provides the Descriptors of the resulting Audio-Visual Scene.

9.1.2 Reference Model

Figure 9 depicts the HMC-SID Reference Model.

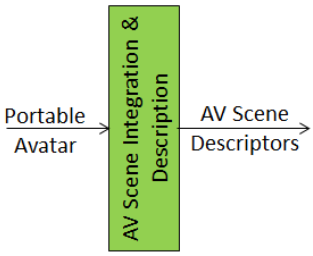

Figure 9 – The AV Scene Integration and Description AIM

9.1.3 I/O Data

Table 3 specifies the Input and Output Data of the Audio-Visual Scene Integration and Description AIM.

Table 3 – I/O Data of the Audio-Visual Scene Integration and Description AIM

Input	Description
Portable Avatar	A Communication Item from a Machine Entity.
Output	**Description**
Audio-Visual Scene Descriptors	The Descriptors of the AV Scene where the Avatar conveyed by the Input Portable Avatar has been added to the Scene with the appropriate Spatial Attitude.

9.2 Audio-Visual Scene Description (OSD-AVS)

9.2.1 Functions

The Audio-Visual Scene Description (OSD-AVS):

1. Receives the Audio-Visual Scene composed of:
1.1 Text.
1.2 Audio Objects – Speech Objects or generic Audio Objects whose source is a point.
1.3 Visual Objects that are either Entities or Generic Objects.
2. Produces the Audio-Visual Scene Descriptors.

9.2.2 Reference Model

Figure 10 depicts the Reference Model of Audio-Visual Scene Description Composite AIM.

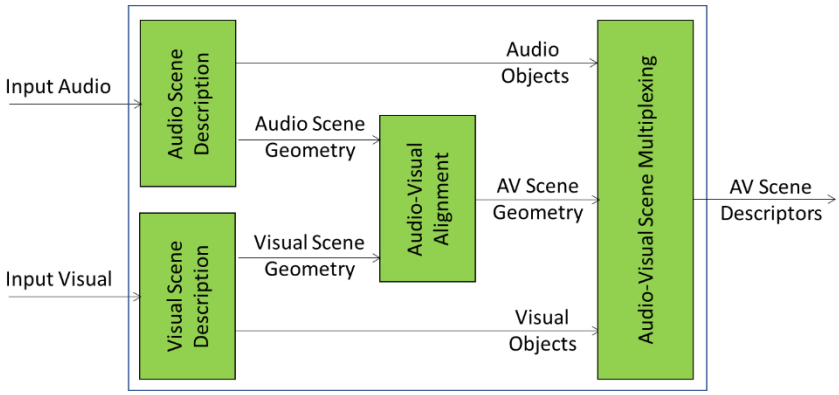

Figure 10 – The Audio-Visual Scene Description Composite AIM

9.2.3 I/O Data

Table 4 specifies the Input and Output Data of the Audio-Visual Description.

Table 4 – I/O Data of the Audio-Visual Description Composite AIM

Input	Description
Input Audio	The audio scene captured by Machine.
Input Visual	The visual scene captured by Machine.
Output	**Description**
Audio-Visual Scene Descriptors	The digital representation of the Audio-Visual Scene Geometry and of the Audio, Visual and Audio-Visual Objects of the Scene.

9.2.4 SubAIMs

9.2.4.1 Audio Scene Description (CAE-ASD)

9.2.4.1.1 Functions

Audio Scene Description:

1. Receives the Audio Scene composed of:
1.1. Microphone Array Geometry.
1.2. Multichannel Audio, i.e., the output of the Microphone Array.
2. Separates Audio Objects.
3. Produces Audio Scene Descriptors.

9.2.4.1.2 Reference Model

Figure 11 depicts the Reference Model of the Audio Scene Description AIM.

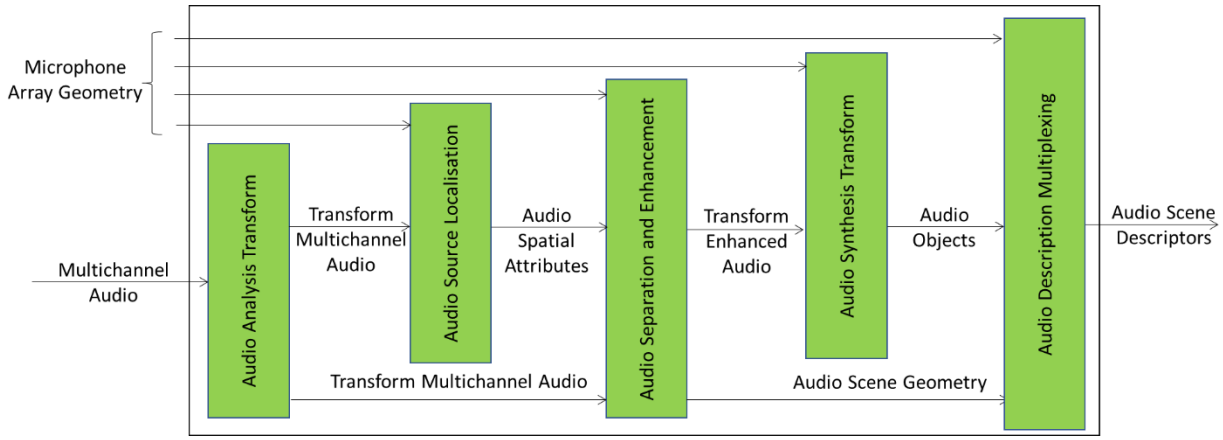

Figure 11 – The Audio Scene Description Composite AIM

9.2.4.1.3 I/O Data

Table 5 specifies the Input and Output Data of the Audio Scene Description AIM.

Table 5 – I/O Data of Audio Scene Description

Input	Description
Microphone Array Geometry	The spatial description of microphone arrangement.
Multichannel Audio	The Audio output of the Microphone Array.
Output	**Description**
Audio Scene Descriptors	The combination of Audio Scene Geometry and Audio Objects.

9.2.4.1.4 SubAIMs

9.2.4.1.4.1 Audio Analysis Transform (CAE-AAT)

9.2.4.1.4.1.1 Function

Audio Analysis Transform:

1. Receives Multichannel Audio.
2. Transforms Multichannel Audio into frequency bands via Fast Fourier Transform (FFT). The operations of the subsequent AIMs are carried out in discrete frequency bands. When such a configuration is used, a 50% overlap between subsequent Audio Blocks must be employed.
3. Outputs a data structure comprising complex valued audio samples in the frequency domain.

9.2.4.1.4.1.2 Reference Model

Figure 12 depicts the Reference Model of the Audio Analysis Transform AIM.

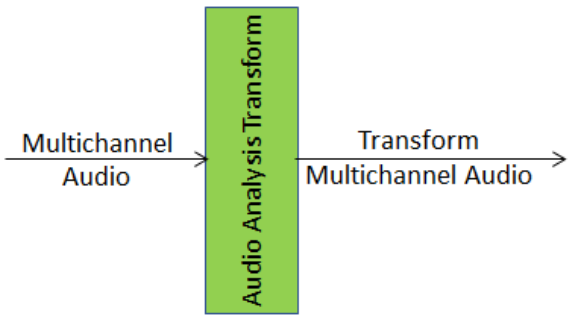

Figure 12 – The Audio Analysis Transform AIM

9.2.4.1.4.1.3 I/O Data

Table 6 specifies the Input and Output Data of the Audio Analysis Transform AIM.

Table 6 - I/O Data of the Audio Analysis Transform AIM

Input	Description
Multichannel Audio	The Audio output of the Microphone Array.
Output	**Description**
Transform Multichannel Audio	The result of the application of the Fast Fourier Transform to Multichannel Audio.

9.2.4.1.4.2 Audio Source Localisation (CAE-ASL)

9.2.4.1.4.2.1 Function

Audio Source Localisation:

1. Receives
1.1. Microphone Array Geometry.
1.2. Transform Multichannel Audio
2. Produces Audio Spatial Attributes (Orientation and Direction of the Audio Objects).

9.2.4.1.4.2.2 Reference Model

Figure 13 depicts the Reference Model of the Audio Source Localisation AIM.

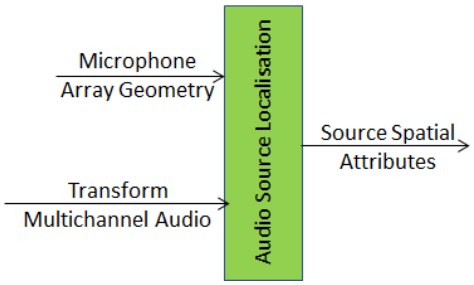

Figure 13 – Audio Source Localisation AIM

9.2.4.1.4.2.3 I/O Data

Table 7 specifies the Input and Output Data of the Audio Source Localisation AIM.

Table 7 – Audio Source Localisation AIM

Input	Description
Microphone Array Geometry	The spatial description of microphone arrangement.
Transform Multichannel Audio	The result of the application of the Fast Fourier Transform to the Multichannel Audio.
Output	**Description**
Audio Spatial Attitude	The Orientations and Directions of Audio Objects.

9.2.4.1.4.3 Audio Separation and Enhancement (CAE-ASE)

9.2.4.1.4.3.1 Function

Audio Separation and Enhancement:

1. Receives the Transform Multichannel Audio and the Microphone Array Geometry.
2. Separates the Audio Objects by using their spatial attributes.
3. Outputs the individual Audio Objects.

9.2.4.1.4.3.2 Reference Model

Figure 14 depicts the Reference Model of the Audio Separation and Enhancement AIM.

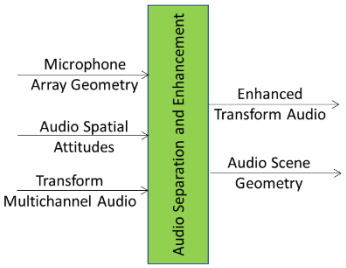

Figure 14 - Audio Separation and Enhancement AIM

9.2.4.1.4.3.3 I/O Data

Table 8 specifies the Input and Output Data of the Audio Separation and Enhancement AIM.

Table 8 - I/O Data of Audio Separation and Enhancement

Input	Description
Transform Multichannel Audio	The result of the application of the Fast Fourier Transform to the Multichannel Audio.
Audio Spatial Attitude	The Orientations and Directions of Audio Objects.
Microphone Array Geometry	The spatial description of microphone arrangement.
Output	**Description**
Enhanced Transform Audio	Multichannel Audio in the transform domain.
Audio Scene Geometry	The spatial arrangement of the Audio Objects.

9.2.4.1.4.4 Audio Synthesis Transform (CAE-AST)

9.2.4.1.4.4.1 Function

Audio Synthesis Transform:

1. Receives Transform Multichannel Audio, Source Spatial Attributes, and Microphone Array Geometry.
2. Transforms the Enhanced Transform Source from the frequency domain to the time domain via an Inverse Fast Fourier Transform.
3. Outputs Enhanced Audio Objects.

9.2.4.1.4.4.2 Reference Model

Figure 15 depicts the Reference Model of the Audio Synthesis Transform AIM.

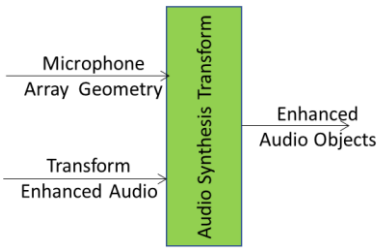

Figure 15 – The Audio Synthesis Transform AIM

9.2.4.1.4.4.3 I/O Data

Table 9 specifies the Input and Output Data of the Audio Synthesis Transform AIM.

Table 9 – I/O Data of Synthesis Transform

Input	Description
Microphone Array Geometry	The spatial description of microphone arrangement.
Enhanced Transform Audio	Audio Objects without noise in the time-frequency domain.
Output	**Description**
Enhanced Audio Objects	Time-domain Audio Objects without noise.

9.2.4.1.4.5 Audio Descriptor Multiplexing (CAE-ADM)

9.2.4.1.4.5.1 Function

Audio Descriptor Multiplexing:

1. Receives Microphone Array Geometry, Enhanced Audio Objects, and the Audio Scene Geometry.
2. Multiplexes into one stream:

2.1. Microphone Array Geometry
2.2. Enhanced Audio
2.3. Audio Scene Geometry.

9.2.4.1.4.5.2 Reference Model

Figure 16 depicts the Reference Model of the Audio Descriptor Multiplexing AIM.

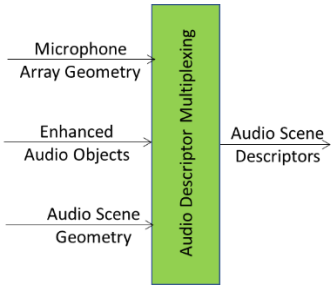

Figure 16 – The Audio Descriptor Multiplexing AIM

9.2.4.1.4.5.3 I/O Data

Table 10 specifies the Input and Output Data of the Audio Descriptor Multiplexing AIM.

Table 10 – I/O Data of Audio Descriptor Multiplexing

Input	Description
Microphone Array Geometry	The spatial description of microphone arrangement.
Enhanced Audio Objects	Time-domain Audio Objects without noise.
Audio Scene Geometry	The spatial arrangement of the Audio Objects
Output	**Description**
Audio Scene Descriptors	The combination of Audio Scene Geometry and Audio Objects.

9.2.4.2 Visual Scene Description

9.2.4.2.1 Functions

Visual Scene Description (OSD-VSD)

1. Receives a Visual Scene.
2. Produces the Visual Scene Descriptors.

9.2.4.2.2 Reference Model

The Reference Model is depicted in Figure 17.

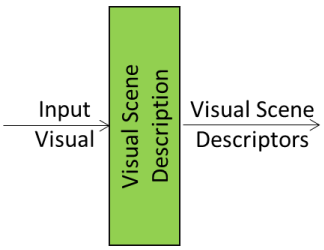

Figure 17 – The Visual Scene Description AIM

9.2.4.2.3 I/O Data

Table 11 specifies the Input and Output Data of the Visual Scene Description AIM.

Table 11 – I/O Data of the Visual Scene Description AIM

Input	Description
Input Visual	Visual Scene captured by Machine.
Output	**Description**
Visual Scene Descriptors	The Visual Descriptors of the Visual Scene.

9.2.4.3 Audio-Visual Alignment (OSD-AVA)

9.2.4.3.1 Functions

Audio-Visual Alignment:

1. Receives the Audio Scene Geometry and the Visual Scene Geometry.
2. Produces the Identifiers of the Audio Objects and Visual Objects that share the same Spatial Attitude.

9.2.4.3.2 Reference Model

Figure 18 depicts the Reference Model of the Audio-Visual Alignment AIM.

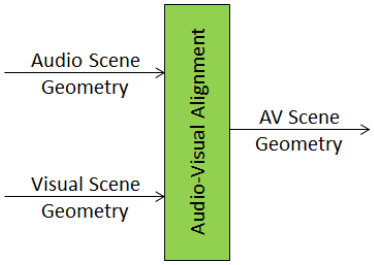

Figure 18 - Audio-Visual Alignment AIM

9.2.4.3.3 I/O Data

Table 12 specifies the Input and Output Data of the Audio-Visual Alignment AIM.

Table 12 – I/O Data of the Audio-Visual Alignment AIM

Input	Description
Audio Scene Geometry	The digital representation of the Spatial arrangement of the Audio Objects of the Scene.

Visual Scene Geometry	The digital representation of the spatial arrangement of the Visual Objects of the Scene.
Output	**Description**
Audio-Visual Scene Geometry	The digital representation of the Spatial arrangement of the Audio, Visual and Audio-Visual Objects of the Scene.

9.2.4.4 Audio-Visual Scene Multiplexing (OSD-SMX)

9.2.4.4.1 Functions

Audio-Visual Scene Multiplexing:
1. Receives Audio and Visual Objects and Audio-Visual Scene Geometry
2. Produces the Descriptors of the Audio-Visual Scene.

9.2.4.4.2 Reference Model

Figure 19 depicts the Reference Model of the Audio-Visual Scene Multiplexing AIM.

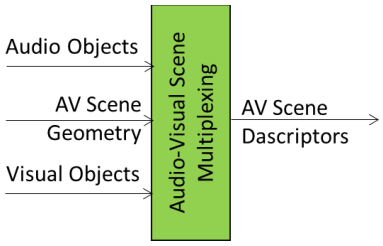

Figure 19 – The Audio-Visual Scene Multiplexing AIM

9.2.4.4.3 I/O Data

Table 12 specifies the Input and Output Data of the Audio-Visual Scene Multiplexing AIM.

Table 13 – I/O Data of the Audio-Visual Alignment AIM

Input	**Description**
Audio Objects	The Audio Objects of the Scene.
Audio-Visual Scene Geometry	The arrangement of the Audio, Visual, Audio-Visual Objects of the Scene.
Visual Objects	The Visual Objects of the Scene.
Output	**Description**
Audio-Visual Scene Descriptors	The combination of Audio, Visual, and Audio-Visual Objects, and Audio-Visual Scene Geometry.

9.3 Entity and Context Understanding (HMC-ECU)

9.3.1 Functions

Entity and Context Understanding (HMC-ECU):
1. Receives Audio-Visual Scene Descriptors.
2. Demultiplexes Audio-Visual Scene Descriptors components.
3. Performs:

3.1. Recognition of Entity's Speech.
3.2. Recognition of Audio Object and Visual Object.
3.3. Understanding of Entity's Natural Language expressed as Text.
3.4. Extraction of Entity's Personal Status.
3.5. Translation of Entity's Text.

4. Produces:
4.1. Audio-Visual Scene Geometry
4.2. Entity ID
4.3. Audio Instance ID
4.4. Visual Instance ID
4.5. Personal Status
4.6. Translated and Refined Text
4.7. Meaning.

Figure 20 depicts the Reference Model of the Entity and Context Understanding Composite AIM.

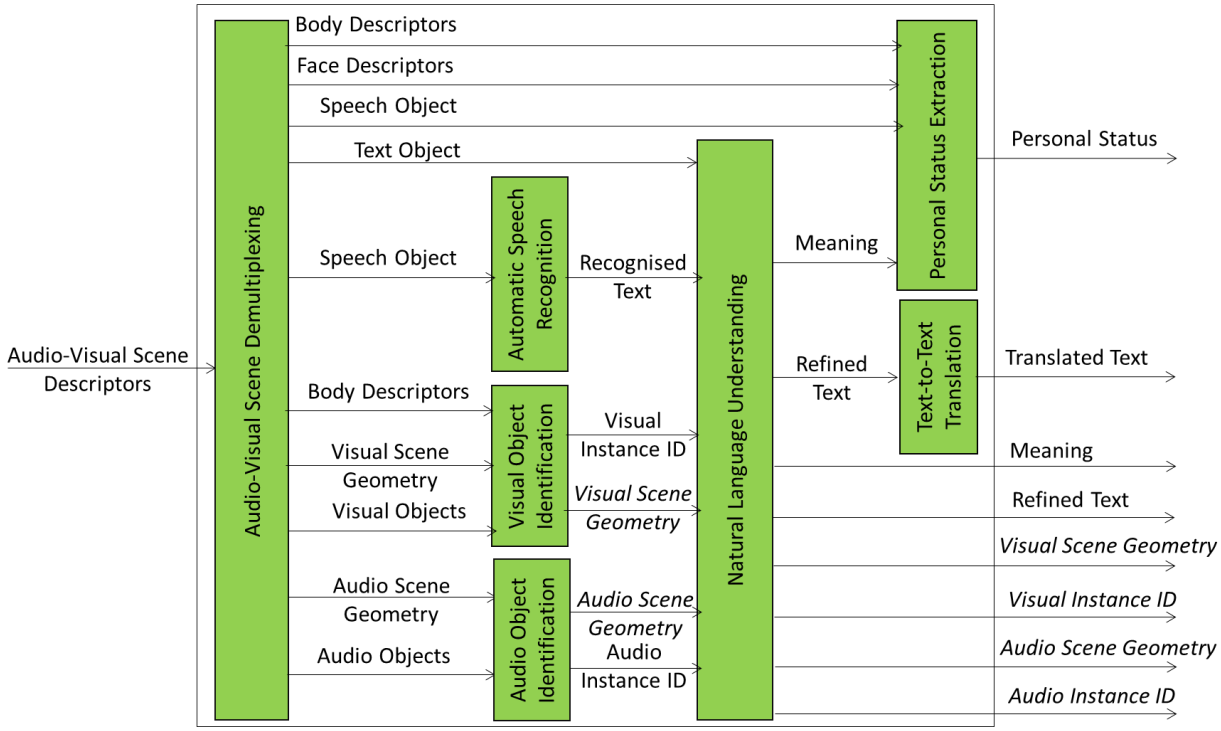

Figure 20 – The Entity and Context Understanding Composite AIM

Note that Output Data in italic are passed directly from the homonymous Input Data.

9.3.2 I/O Data

Table 14 specifies the Input and Output Data of the of the Entity and Context Understanding AIM.

Table 14 – I/O Data of the Entity and Context Understanding Composite AIM

Input	Description
Audio-Visual Scene Descriptors	The combination of Audio, Visual, and Audio-Visual Objects, and Audio-Visual Scene Geometry.
Output	**Description**
Personal Status	Personal Status of Entity having the Entity ID.
Translated Text	Translated Text of Text Object or of Text conveyed by Speech Object.
Refined Text	Refined Text of Speech Object.
Meaning	Other name for Refined Text Descriptors.

Visual Instance ID	The Identifier of the specific Visual Object belonging to a level in the taxonomy.
Audio Scene Geometry	The arrangement of the Audio Objects in the Audio Scene.
Visual Scene Geometry	The arrangement of the Visual Objects in the Visual Scene.
Audio Instance ID	The Identifier of the specific Audio Object belonging to a level in the taxonomy.

9.3.3 SubAIMs

9.3.3.1 Audio-Visual Scene Demultiplexing

9.3.3.1.1 Functions

The Audio-Visual Scene Demultiplexing AIM (OSD-SDX):

1. Receives Audio-Visual Scene Descriptors.
2. Produces:
    a. Audio Scene Geometry
    b. Audio Objects
    c. Visual Objects
    d. Visual Scene Geometry.

9.3.3.1.2 Reference Model

Figure 21 depicts the Reference Model of the Entity Context Understanding Composite AIM.

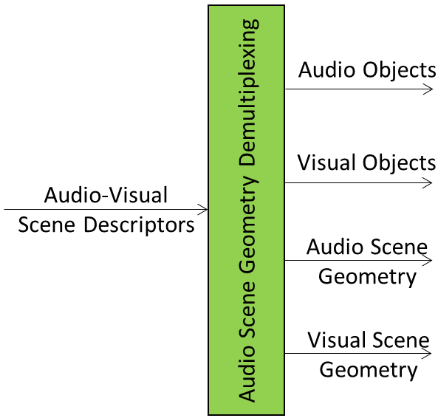

Figure 21 - Audio Scene Geometry Demultiplexing

9.3.3.1.3 I/O Data

Table 15 specifies the Input and Output Data of the of the Audio Scene Geometry Demultiplexing AIM.

Table 15 - Audio Scene Geometry Demultiplexing

Input	**Description**
Visual Scene Descriptors	The Descriptors of the Audio-Visual Scene.
Output	**Description**
Audio Scene Geometry	The arrangement of the Audio Objects in the Audio Scene.
Audio Object	The Audio Objects in the Scene.
Visual Object	The Visual Objects in the Scene.
Visual Scene Geometry	The arrangement of the Visual Objects in the Visual Scene.

9.3.3.2 Automatic Speech Recognition (MMC-ASR)

9.3.3.2.1 Functions

Automatic Speech Recognition:

1. Receives Input Speech.
2. Extracts the Text conveyed by the Input Speech.

9.3.3.2.2 Reference Model

Figure 22 depicts the Reference Model of the Automatic Speech Recognition AIM.

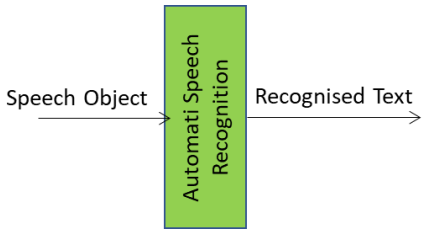

Figure 22 – The Automatic Speech Recognition AIM

9.3.3.2.3 I/O Data

Table 16 specifies the Input and Output Data of the Automatic Speech Recognition AIM.

Table 16 – I/O Data of the Automatic Speech Recognition AIM

Input	**Description**
Speech Object	Speech Object emitted by Entity
Output	**Description**
Recognised Text	Output of the Automatic Speech Recognition AIM

9.3.3.3 Visual Object Identification (OSD-VOI)

9.3.3.3.1 Functions

Visual Object Identification:

1. Receives the Visual Scene Geometry, the Visual Objects, and the Body Descriptors.
2. Produces a Visual Instance ID identifying a Visual Object in the Scene that belongs to some level in a taxonomy.

9.3.3.3.2 Reference Model

Figure 23 depicts the Reference Model of the Visual Object Identification AIM.

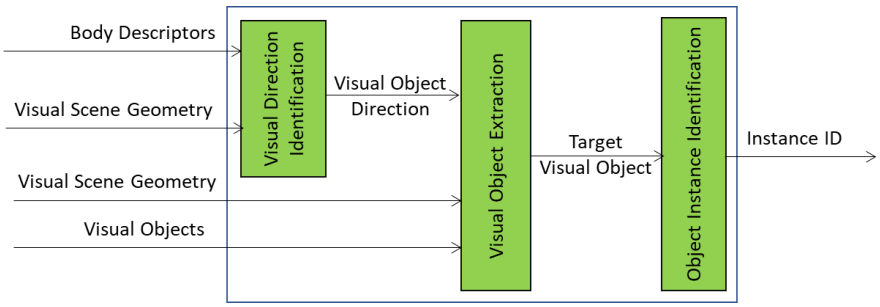

Figure 23 – The Visual Object Identification Composite AIM

Note that the Visual Direction Identification AIM can parse either an AV Scene Geometry or its Visual Scene Geometry subset.

9.3.3.3.3 I/O Data

Table 17 specifies the Input and Output Data of the Visual Object Identification AIM.

Table 17 – I/O Data of the Visual Object Identification AIM

Input	Description
Body Descriptors	The Descriptors of the Body Objects of Entities in the Visual Scene.
Visual Scene Geometry	The digital representation of the spatial arrangement of the Visual Objects of the Scene.
Visual Objects	The Visual Objects in the Visual Scene.
Output	**Description**
Visual Instance ID	The Identifier of the specific Visual Object belonging to a level in the taxonomy.

9.3.3.3.4 SubAIMs

9.3.3.3.4.1 Visual Direction Identification (VOI-VDI)

9.3.3.3.4.1.1 Function

Visual Direction Identification:

1. Receives Visual Scene Geometry and Body Descriptors.
2. Produces the direction of a line traversing the forefinger of the Entity.

9.3.3.3.4.1.2 Reference Model

Figure 24 depicts the Reference Model of the Visual Direction Identification AIM.

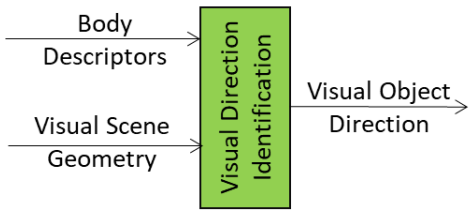

Figure 24 – The Visual Direction Identification AIM

9.3.3.3.4.1.3 I/O Data

Table 18 specifies the Input and Output Data of the Visual Direction Identification AIM.

Table 18 – I/O Data of the Visual Direction Identification AIM

Input	Description
Body Descriptors	The Descriptors of the Body Objects of Entities in the Visual Scene.
Visual Scene Geometry	The digital representation of the spatial arrangement of the Visual Objects of the Scene.
Output	**Description**
Visual Object Direction	The direction of the line traversing the forefinger of the target Entity.

9.3.3.3.4.2 Visual Object Extraction (VOI-VOE)

9.3.3.3.4.2.1 Function

Visual Object Extraction:

1. Receives Visual Scene Geometry, Visual Objects, and Anchored Direction.

2. Singles out the Visual Object indicated by the Entity.

9.3.3.3.4.2.2 Reference Model

Figure 25 depicts the Reference Model of the Visual Object Extraction AIM.

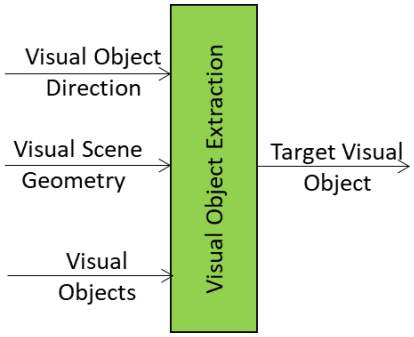

Figure 25 – The Visual Object Extraction AIM

9.3.3.3.4.2.3 I/O Data

Table 19 specifies the Input and Output Data of the Visual Object Extraction AIM.

Table 19 – I/O Data of the Visual Object Extraction AIM

Input	Description
Anchored Direction	The direction of the line traversing the forefinger of the Entity.
Visual Scene Geometry	The digital representation of the spatial arrangement of the Visual Objects of the Scene.
Visual Objects	The Visual Objects identified in the Visual Scene Geometry.
Output	**Description**
Target Visual Object	The Visual Object crossed by the line traversing the forefinger of the Entity.

9.3.3.3.4.3 Visual Instance Identification (VOI-OII)

9.3.3.3.4.3.1 Function

Visual Instance Identification:

1. Receives a Visual Object.
2. Produces an Instance ID identifying an element of a set of Visual Objects belonging to a level in a taxonomy.

9.3.3.3.4.3.2 Reference Model

Figure 26 depicts the Reference Model of the Visual Instance Identification AIM.

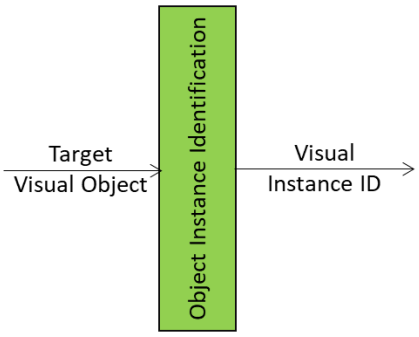

Figure 26 – The Object Instance Identification AIM

9.3.3.3.4.3.3 I/O Data

Table 20 specifies the Input and Output Data of the Visual Instance Identification AIM.

Table 20 – I/O Data of Visual Instance Identification

Input	Description
Target Visual Object	The Visual Object crossed by the line traversing the forefinger of the Entity.
Output	**Description**
Visual Instance ID	The Identifier of the specific Visual Object belonging to a level in the taxonomy.

9.3.3.4 Audio Object Identification (CAE-AOI)

9.3.3.4.1 Functions

Audio Object Identification:

1. Receives the Audio Scene Geometry and the Audio Objects.
2. Produces an Audio Instance ID identifying an Audio Object in the Scene that belongs to some level in a taxonomy.

9.3.3.4.2 Reference Model

Figure 27 depicts the Reference Model of the Audio Object Identification AIM.

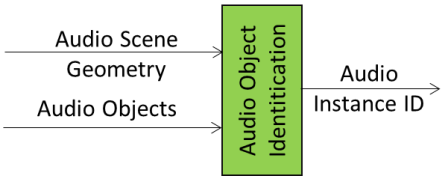

Figure 27 - Audio Object Identification AIM

Note that the Audio Object Identification AIM can parse either an AV Scene Geometry or its Audio Scene Geometry subset.

9.3.3.4.3 I/O Data

Table 17 specifies the Input and Output Data of the Audio Object Identification AIM.

Table 21 – I/O Data of the Audio Object Identification AIM

Input	Description
Audio Scene Geometry	The digital representation of the spatial arrangement of the Audio Objects of the Scene.
Audio Objects	The Audio Objects in the Audio Scene Geometry subject to identification.
Output	**Description**
Audio Instance ID	The Identifier of the specific Audio Object belonging to a level in the taxonomy.

9.3.3.5 Natural Language Understanding

9.3.3.5.1 Functions

Natural Language Understanding (MMC-NLU):

1. Receives Text Object, Recognised Text, Visual Instance ID, AV Scene Geometry, and Audio Instance ID.
2. Refines Recognised Text and extracts Meaning considering the spatial position of the selected Audio Instance and Visual Instance and the semantics of the two Instances obtained from Audio Instance ID and Visual Instance ID.
3. Produces Refined Text and Text Descriptors (Meaning).

9.3.3.5.2 Reference Model

Figure 28 depicts the Reference Model of the Natural Language Understanding AIM.

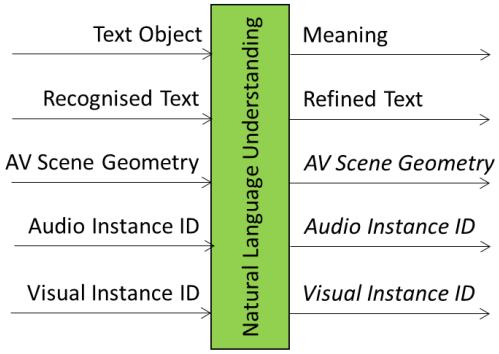

Figure 28 – The Natural Language Understanding AIM

Note that Output Data in italic are passed directly from the homonymous Input Data.

9.3.3.5.3 I/O Data

Table 22 specifies the Input and Output Data of the Natural Language Understanding AIM.

Table 22 – I/O Data of the Natural Language Understanding AIM

Input	Description
Text Object	ID of the Entity emitting an Audio-Visual Scene or a Communication Item.
Recognised Text	Input from the Automatic Speech Recognition AIM.
Audio-Visual Scene Geometry	The digital representation of the spatial arrangement of the Audio-Visual Objects of the Scene.
Audio Instance ID	The Identifier of the specific Audio Object belonging to a level in the taxonomy.
Visual Instance ID	The Identifier of the specific Visual Object belonging to a level in the taxonomy.
Output	**Description**
Meaning	Descriptors of the Refined Text.
Refined Text	The refined version of the Recognised Text.
Audio-Visual Scene Geometry	As in Input
Audio Instance ID	As in Input
Visual Instance ID	As in Input

9.3.3.6 Personal Status Extraction (MMC-PSE)

9.3.3.6.1 Functions

Personal Status Extraction:

1. Receives:
1.1. Text Object or Text Descriptors
1.2. Speech Object or Speech Descriptors
1.3. Face Object or Face Descriptors
1.4. Body Object or Gesture Descriptors
2. Computes and then Interprets, depending on whether the Descriptors of a Modality (Text, Speech, or Face) have been received:
2.1. Text Descriptors: alternatively, Interprets the received Descriptors and produces Personal Status of the Text Object (PS-Text).
2.2. Speech Descriptors: alternatively, Interprets the received Descriptors and produces Personal Status of the Speech Object (PS-Speech).
2.3. Face Descriptors: alternatively, Interprets the received Descriptors and produces Personal Status of the Face Object (PS-Face).
2.4. Gesture Descriptors; alternatively, Interprets the received Gesture Descriptors of the Body Object.
3. Produces the Personal Status of the Entity by multiplexing the results of the interpretations.

9.3.3.6.2 Reference Model

Figure 29 depicts the Reference Model of the Personal Status Extraction AIM.

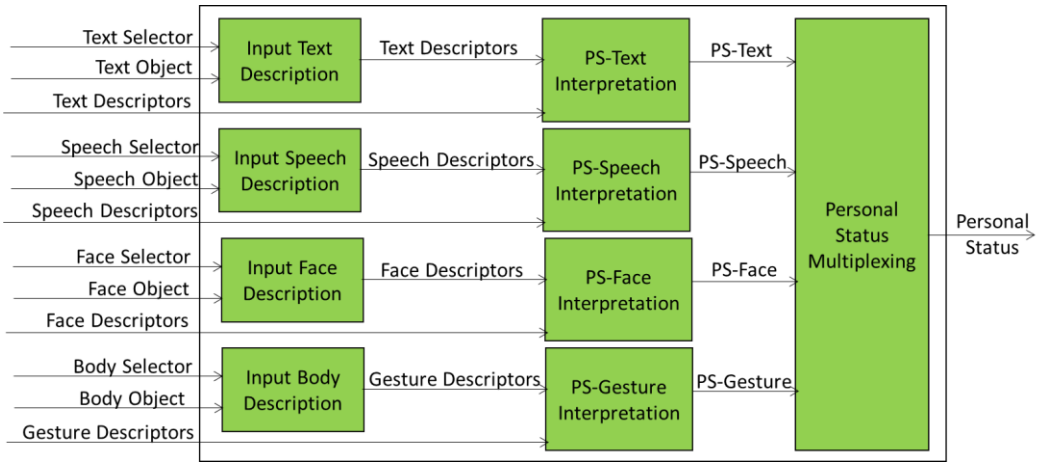

Figure 29 – The Personal Status Extraction Composite AIM

9.3.3.6.3 I/O Data

Table 23 specifies the Input and Output Data of the Personal Status Extraction AIM.

Table 23 – I/O Data of the Personal Status Extraction AIM

Input data	From	Comment
Text Selector	An external signal	Text/Descriptors Selector
Text Object	Keyboard or AIM	Text or Recognised Text.
Text Descriptors	An upstream AIM	Functionally equivalent to Text Description.
Speech Selector	An external signal	Speech/Descriptors Selector.
Speech Object	Microphone/upstream AIM	Speech of Entity.
Speech Descriptors	An upstream AIM	Functionally equivalent to Meaning.
Face Selector	An external signal	Face/Descriptors Selector.
Face Object	Visual Scene Description	The face of the Entity.

Face Descriptors	An upstream AIM	Functionally equivalent to Face Description.
Gesture Selector	An external signal	Body/Descriptors Selector
Body Object	Visual Scene Description	The body of the Entity.
Gesture Descriptors	An upstream AIM	Functionally equivalent to Body Description.
Output data	**To**	**Description**
Personal Status	A downstream AIM	For further processing

9.3.3.6.4 SubAIMs

9.3.3.6.4.1 Input Text Description (MMC-ITD)

9.3.3.6.4.1.1 Functions

Input Text Description:
1. Receives Text Selector and Text.
2. Produces Text Descriptors.

9.3.3.6.4.1.2 Reference Model

Figure 30 depicts the Reference Model of the Input Text Description AIM.

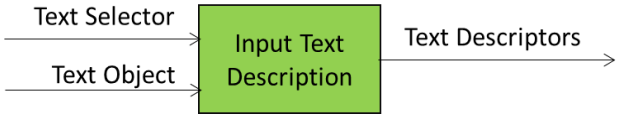

Figure 30 – The Input Text Description AIM

9.3.3.6.4.1.3 I/O Data

Table 24 specifies the Input and Output Data of the Input Text Description AIM.

Table 24 – I/O Data of the Input Text Description AIM

Input	**Description**
Text Selector	Text/Descriptors Selector
Text Object	Text or Recognised Text
Output	**Description**
Text Descriptors	Descriptors of Text

9.3.3.6.4.2 Input Speech Description (MMC-ISD)

9.3.3.6.4.2.1 Functions

Input Speech Description:
1. Receives Speech Selector and Speech
2. Produces Speech Description

9.3.3.6.4.2.2 Reference Model

Figure 31 depicts the Reference Model of the Input Speech Description AIM.

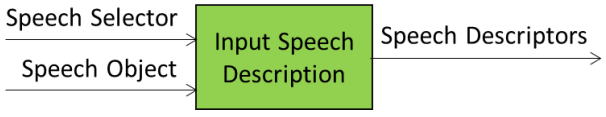

Figure 31 Input Speech Description AIM

9.3.3.6.4.2.3 I/O Data

Table 25 specifies the Input and Output Data of the Input Speech Description AIM.

Table 25 – I/O Data of the Input Speech Description AIM

Input	Description
Speech Selector	Speech/Descriptors Selector
Speech Object	Speech of Entity
Output	**Description**
Speech Descriptors	Descriptors of Speech

9.3.3.6.4.3 Input Face Description (PAF-IFD)

9.3.3.6.4.3.1 Functions

Input Face Description:

1. Receives Face Selector and Face
2. Produces Face Description.

9.3.3.6.4.3.2 Reference Model

Figure 32 depicts the Reference Model of the Input Face Description AIM.

Figure 32 – The Input Face Description AIM

9.3.3.6.4.3.3 I/O Data

Table 26 specifies the Input and Output Data of the Input Face Description AIM.

Table 26 – I/O Data of the Input Face Description AIM

Input	Description
Face Selector	Face/Descriptors Selector
Face Object	Face of Entity
Output	**Description**
Face Descriptors	Descriptors of Face

9.3.3.6.4.4 Input Body Description (PAF-IBD)

9.3.3.6.4.4.1 Functions

Input Body Description:

1. Receives Body Selector and Body
2. Produces Gesture Description

9.3.3.6.4.4.2 Reference Model

Figure 33 depicts the Reference Model of the Input Body Description AIM.

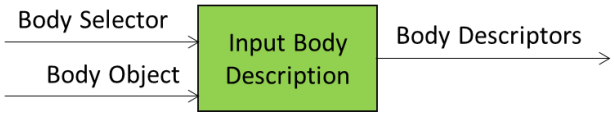

Figure 33 Input Body Description AIM

9.3.3.6.4.4.3 I/O Data

Table 27 specifies the Input and Output Data of Input Body Description AIM.

Table 27 – I/O Data of the Input Body Description

Input	Description
Body Selector	Body/Descriptors Selector
Body Objects	Body of Entity.
Output	**Description**
Gesture Descriptors	Descriptors of Body

9.3.3.6.4.5 PS-Text Interpretation (MMC-PTI)

9.3.3.6.4.5.1 Functions

PS-Text Interpretation:

1. Receives Text Descriptors, either from Text Description or as an input to PS-Text Interpretation.
2. Produces PS-Text, the Personal Status of the Text Modality.

9.3.3.6.4.5.2 Reference Model

Figure 34 depicts the Reference Model of the PS-Text Interpretation AIM.

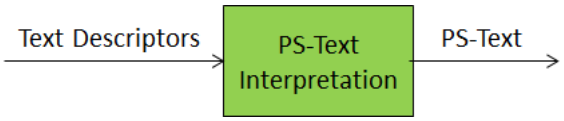

Figure 34 – The PS-Text Interpretation AIM

9.3.3.6.4.5.3 I/O Data

Table 28 specifies the Input and Output Data of the PS-Text Interpretation AIM.

Table 28 – I/O Data of the PS-Text Interpretation AIM

Input	Description
Text Descriptors	Descriptors of Text
Output	**Description**
PS-Text	Personal Status of Text

9.3.3.6.4.6 PS-Speech Interpretation (MMC-PSI)

9.3.3.6.4.6.1 Functions

PS-Speech Interpretation:

1. Receives PS-Speech Descriptors, either from Speech Description or as an input to PS-Speech Interpretation
2. Produces PS-Speech, the Personal Status of the Speech Modality.

9.3.3.6.4.6.2 Reference Model

Figure 35 depicts the Reference Model of the PS-Speech Interpretation AIM.

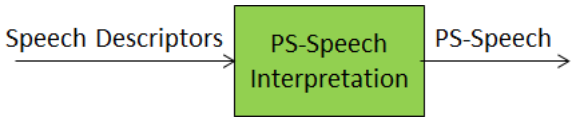

Figure 35 The PS-Speech Interpretation AIM

9.3.3.6.4.6.3 I/O Data

Table 29 specifies the Input and Output Data of the PS-Speech Interpretation AIM.

Table 29 – I/O Data of the PS-Speech Interpretation AIM

Input	Description
Speech Descriptors	Descriptors of Speech
Output	**Description**
PS-Speech	Personal Status of Speech

9.3.3.6.4.7 PS-Face Interpretation (PAF-PFI)

9.3.3.6.4.7.1 Functions

PS-Face Interpretation:

1. Receives PS-Face Descriptors, either from Face Description or as an input to PS-Face Interpretation
2. Produces PS-Face, the Personal Status of the Face Modality.

9.3.3.6.4.7.2 Reference Model

Figure 36 depicts the Reference Model of the PS-Face Interpretation AIM.

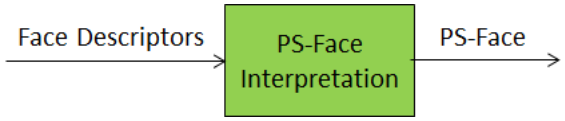

Figure 36 – The PS-Face Interpretation AIM

9.3.3.6.4.7.3 I/O Data

Table 30 specifies the Input and Output Data of the PS-Face Interpretation AIM.

Table 30 – I/O Data of the PS-Face Interpretation AIM

Input	Description
Face Descriptors	Descriptors of Face
Output	**Description**
PS-Face	Personal Status of Face

9.3.3.6.4.8 PS-Gesture Interpretation (PAF-PGI)

9.3.3.6.4.8.1 Functions

PS-Gesture Interpretation:

1. Receives PS-Gesture Descriptors, either from Gesture Description or as an input to PS-Gesture Interpretation
2. Produces PS-Gesture, the Personal Status of the Gesture Modality.

9.3.3.6.4.8.2 Reference Model

The Reference Model is depicted in Figure 37.

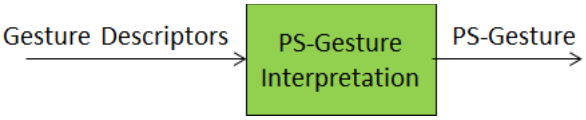

Figure 37 PS-Gesture Interpretation AIM

9.3.3.6.4.8.3 I/O Data

Table 31 specifies the Input and Output Data of PS-Gesture Interpretation AIM.

Table 31 – I/O Data of the PS-Gesture Interpretation AIM

Input	Description
Gesture Descriptors	Descriptors of Body
Output	**Description**
PS-Gesture	Personal Status of Body

9.3.3.6.4.9 Personal Status Multiplexing (MMC-PSM)

9.3.3.6.4.9.1 Functions

Personal Status Multiplexing:

1. Receives any of PS-Text, PS-Speech, PS-Face, and PS-Gesture.
2. Multiplexes the input data.
3. Produces Personal Status.

9.3.3.6.4.9.2 Reference Model

Figure 38 depicts the Reference Model of the Personal Status Multiplexing AIM.

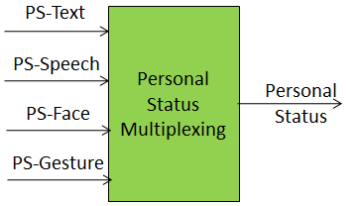

Figure 38 – The Personal Status Multiplexing AIM

9.3.3.6.4.9.3 I/O Data

Table 32 specifies the Input and Output Data of the Personal Status Multiplexing AIM.

Table 32 – I/O Data of the Personal Status Multiplexing

Input	Description
PS-Text	Personal Status of Text Object.
PS-Speech	Personal Status of Speech Object.
PS-Face	Personal Status of Face Object.
PS-Gesture	Personal Status of Gesture conveyed by Body Object.
Output	**Description**
Personal Status	Personal Status of Machine.

9.3.3.7 Text-to-Text Translation (MMC-TTT)

9.3.3.7.1 Functions

Text-to-Text Translation:
1. Receives:
1.1. Selector determining the input and target language.
1.2. Refined Text.
2. Produces Translated Text.

9.3.3.7.2 Reference Model

Figure 39 depicts the Reference Model of the Text-to-Text Translation AIM.

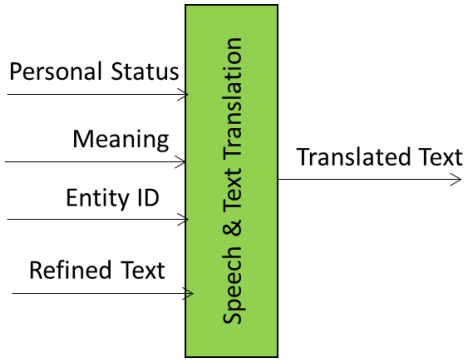

Figure 39 - Text-to-Text Translation Composite AIM

9.3.3.7.3 I/O Data

Table 33 specifies the Input and Output Data of Text-to-Text Translation AIM.

Table 33 – I/O Data of the Text-to-Text Translation AIM

Input	Description
Personal Status	Personal Status of Machine.
Meaning	Descriptors of Text.
Entity ID	ID of the Entity.
Refined Text	Text Object emitted by Entity.
Output	**Description**
Translated Text	Translation of Text emitted by Entity.

9.4 Entity Dialogue Processing

9.4.1 Functions

Entity Dialogue Processing (MMC-EDP):
1. Receives:
1.1. ID of Visual Instance and ID of Audio Instance the Entity refers to.
1.2. ID, Personal Status, Text and/or Translated Text, and Meaning of the Entity the Machine is communicating with.
2. Produces its Machine ID and Text and Personal Status based on Input Data.

9.4.2 Reference Model

Figure 40 depicts the Reference Model of the Entity Dialogue Processing AIM.

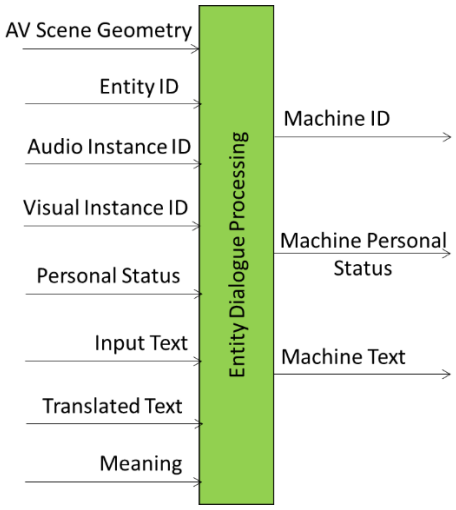

Figure 40 – The Entity Dialogue Processing AIM

9.4.3 I/O Data

Table 34 specifies the Input and Output Data of the Entity Dialogue Processing AIM.

Table 34 – I/O Data of the Entity Dialogue Processing AIM

Input	Description
Audio-Visual Scene Geometry	The digital representation of the spatial arrangement of Audio-Visual Objects in the Audio-Visual Scene.
Entity ID	The ID of the Entity the Machine is communicating with.
Audio Instance ID	ID of the Audio Object the Entity refers to.
Visual Instance ID	ID of the Visual Object indicated by the Entity.
Personal Status	Personal Status of the Entity the Machine is communicating with.
Input or Refined Text	Text or Refined Text from the Entity the Machine is communicating with.
Translated Text	Translated Text of the Entity the Machine is communicating with.
Meaning	Descriptors of Text and/or Translated Text of the Entity the Machine is communicating with.
Output	**Description**
Machine ID	ID of the Avatar the Machine gives as input to Personal Status Display.
Machine Text	Text produced by the Machine in response to the Communication Item emitted by the Entity and its Context.
Machine Personal Status	The Personal Status the Machine intends to add to its Modalities.

9.5 Personal Status Display (PAF-PSD)

9.5.1 Functions

Personal Status Display (PAF-PSD):

1. Receives Avatar ID, Text, Avatar Model, and Personal Status.
2. Generates a Portable Avatar containing:

2.1. ID of Machine.
2.2. Text of Machine.
2.3. Avatar of Machine.
2.4. Speech of Machine conveying the intended Personal Status.

2.5. Face and Gesture conveying the intended Personal Status.

Personal Status Display may add other elements of the Portable Avatar Format, such as an Audio-Visual Scene containing components not included in the list above.

9.5.2 Reference Model

Figure 41 depicts the Reference Model of the Personal Status Display AIM.

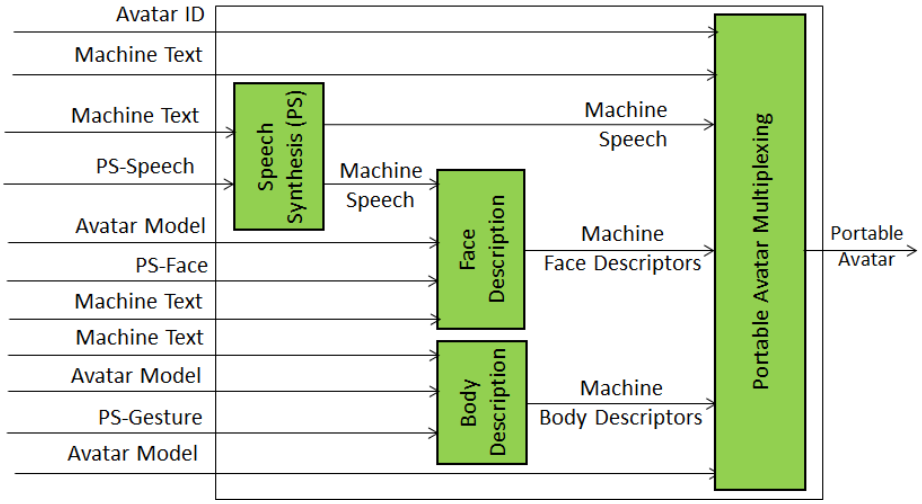

Figure 41 – The Personal Status Display Composite AIM

9.5.3 I/O Data

Table 35 specifies the Input and Output Data of the Personal Status Display AIM.

Table 35 – I/O Data of the Personal Status Display AIM

Input data	From	Comment
Text	Entity Dialogue Processing	Text produced by Machine
PS-Speech	Entity Dialogue Processing	Personal Status of Speech
Avatar Model	Entity Dialogue Processing	Model used to display Machine
PS-Face	Entity Dialogue Processing	Personal Status of Face
PS-Gesture	Entity Dialogue Processing	Personal Status of Gesture
Output data	**To**	**Description**
Portable Avatar	Downstream AIM	e.g., for actual rendering

9.5.4 SubAIMs

9.5.4.1 Text-to-Speech (MMC-TTS)

9.5.4.1.1 Functions

Text-To-Speech:

1. Receives Text, Speech Descriptors, and Personal Status.
2. Produces utterances that convey the input Text with a type of speech specified by Speech Descriptors and with a Personal Status specified by the input Personal Status.

9.5.4.1.2 Reference Model

Figure 42 depicts the Reference Model of the Text-To-Speech AIM.

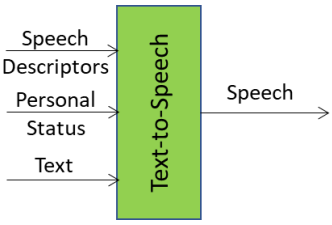

Figure 42 - The Text-To-Speech AIM

9.5.4.1.3 I/O Data

Table 36 specifies the Input and Output Data of the Personal Status Display AIM.

Table 36 - I/O Data of the Text-To-Speech AIM

Input data	From	Comment
Speech Descriptors	E.g., a local storage of a Personal Status Display	To personalise Machine's utterances.
Text	Entity Dialogue Processing	Text produced by Machine
PS-Speech	Entity Dialogue Processing	Personal Status of Speech
Output data	**To**	**Description**
Speech	Downstream AIM	e.g., for actual rendering via a Portable Avatar

9.5.4.2 Input Face Description (PAF-IFD)

See 9.3.3.6.4.3.

9.5.4.3 Input Body Description (PAF-IBD)

See 9.3.3.6.4.4.

9.5.4.4 Portable Avatar Multiplexing (PAF-PMX)

9.5.4.4.1 Functions

A standard Portable Avatar Multiplexing (PSD-PAM) AIM:

1. Receives any number – including none – of the following elements: Avatar ID, Time, Audio-Visual Scene Description, Spatial Attitude, Avatar Model, Body Descriptors, Face Descriptors, Language Preference, Speech Type, Speech, Text, and Personal Status.
2. Produces Portable Avatar.

9.5.4.4.2 Reference Model

Figure 43 depicts the Reference Model of the Portable Avatar Multiplexing AIM.

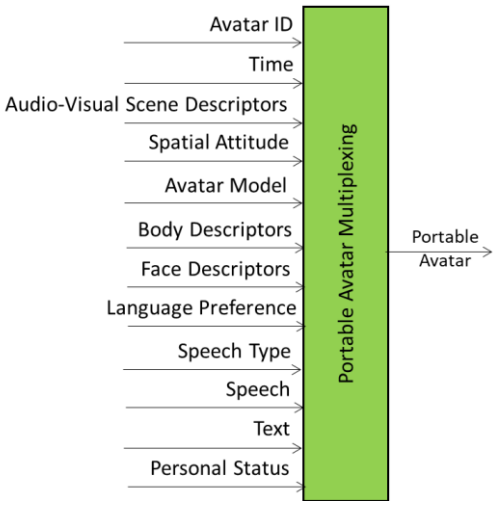

Figure 43 – The Portable Avatar Multiplexing AIM

9.5.4.4.3 I/O Data

Table 37 specifies the Input and Output Data of the Portable Avatar Multiplexing AIM.

Table 37 - Data in and out of the Portable Avatar Multiplexing (PMX)

Input	Description
Avatar ID	ID of Machine
Time	Time the Communication Items refers to (system time).
Audio-Visual Scene Descriptors	Description of the Scene of which the Avatar is part.
Spatial Attitude	Spatial Attitude of the Avatar in the Environment.
Avatar Model	Avatar model used.
Body Descriptors	Body Descriptors of Avatar.
Face Descriptors	Face Descriptors of Avatar.
Language Preference	Language used by Machine.
Speech Type	Speech representation type.
Speech	Speech segment relevant to Time.
Machine Text	Text of Machine.
Personal Status	Personal Status of Machine.
Output	**Description**
Portable Avatar	Communication Item emitted by the Machine

9.6 Audio-Visual Scene Rendering (HMC-AVR)

9.6.1 Functions

Audio-Visual Scene Rendering (HMC-AVR):

1. Receives:
1.1. Portable Avatar
1.2. Receiving Entity's Point of View
2. Produces the Audio-Visual Scene Described by the Portable Avatar.
3. Renders the Audio-Visual Scene and outputs the Text included in the Portable Avatar as seen and heard from the Point of View.

9.6.2 Reference Model

Figure 44 depicts the Reference Model of the Audio-Visual Scene Rendering AIM.

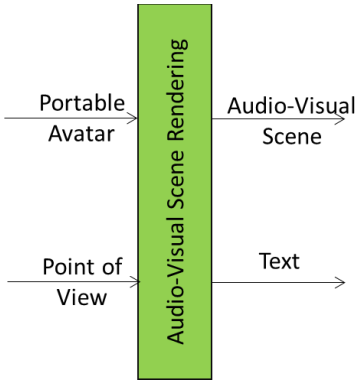

Figure 44 – The Audio-Visual Scene Rendering AIM

9.6.3 I/O Data

Table 38 specifies the Input and Output Data of the Speech & Text Translation AIM.

Table 38 – I/O Data of the Speech & Text Translation AIM

Input	Description
Portable Avatar	Communication Item emitted by Personal Status Display.
Point of View	Point from where an Entity perceives the Audio-Visual Scene
Output	**Description**
Audio-Visual Scene	The rendered Audio-Visual Scene Described by the Portable Avatar.
Text	The Text included in the Portable Avatar.

10 AIW, AIMs, and JSON Metadata

Table 39 provides links to the online specification (column 1) and to the JSON Metadata (column 2) of the HMC-CEC AI Workflow and AI Modules. Columns 3, 4, 5, and 6 provide the acronyms of the AIW and AIMs.

Table 39 - AIW, AIM, and JSON Metadata

Name	JS	AIW	AIMs		
Human and Machine Communication	X	HMC-HMC			
AV Scene Integration and Description	X		HMC-SID		
Audio-Visual Scene Description	X		OSD-AVS		
Audio Scene Description	X			CAE-ASD	
Audio Analysis Transform	X				CAE-AAT
Audio Source Localisation	X				CAE-ASL
Audio Separation and Enhancement	X				CAE-ASE
Audio Synthesis Transform	X				CAE-AST
Audio Description Multiplexing	X				CAE-ADM
Visual Scene Description	X			OSD-VSD	
Audio-Visual Alignment	X			OSD-AVA	
Audio-Visual Scene Multiplexing	X			OSD-AMX	
Entity and Context Understanding	X		HMC-ECU		
Audio-Visual Scene Demultiplexing	X			OSD-SDX	
Automatic Speech Recognition	X			MMC-ASR	
Visual Object Identification	X			OSD-VOI	
Visual Direction Identification	X				OSD-VDI

Visual Object Extraction	X				OSD-VOE
Visual Instance Identification	X				OSD-VII
Audio Object Identification	X			CAE-AOI	
Natural Language Understanding	X			MMC-NLU	
Personal Status Extraction	X			MMC-PSE	
Input Text Description	X				MMC-ITD
Input Speech Description	X				MMC-ISD
Input Face Description	X				PAF-IFD
Input Body Description	X				PAF-IBD
PS-Text Interpretation	X				MMC-PTI
PS-Speech Interpretation	X				MMC-PSI
PS-Face Interpretation	X				PAF-PFI
PS-Gesture Interpretation	X				PAF-PGI
Personal Status Multiplexing	X				MMC-PMX
Text-to-Text Translation	X			MMC-TTT	
Entity Dialogue Processing	X		MMC-EDP		
Personal Status Display	X		PAF-PSD		
Text-to-Speech	X			MMC-TTS	
Input Face Description	X			PAF-IFD	
Input Body Description	X			PAF-IBD	
Portable Avatar Multiplexing	X			PAF-PMX	
Audio-Visual Scene Rendering	X		PAF-AVR		

11 Data Types

11.1 Media

11.1.1 Text

Text is represented according to ISO/IEC 10646; Information technology – Universal Coded Character Set [10].

11.1.2 Audio

Audio is a Data Type representing an analogue audio signal sampled at a frequency between 8-192 kHz with a bits/sample number between 8 and 32.
Input Audio is Multichannel Audio as provided by a Microphone Array.
Output Audio is Audio information such as provided by the Audio-Visual Rendering AIM.

11.1.3 Speech

Speech is a Data Type representing an analogue audio signal sampled at a frequency between 8 kHz and 96 kHz with a number of bits/sample of 8, 16 and 24, and uniform and non-uniform quantisation.

11.1.4 Multichannel Audio

Multichannel Audio is a Data Type whose structure contains between 4 and 256 time-aligned interleaved Audio Channels organised in blocks as depicted in *Figure 45*.

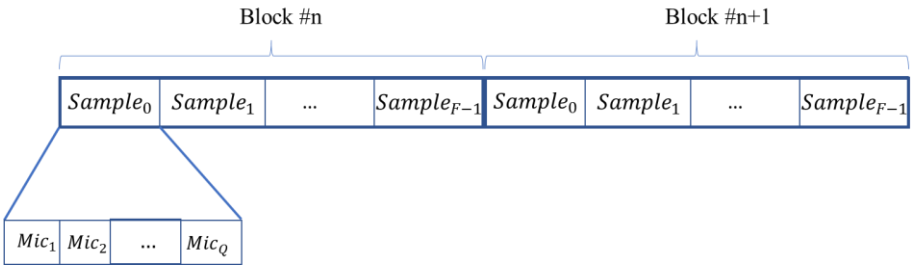

Figure 45 – Ordering of Samples in Multichannel Audio

11.1.5 Visual

Input Visual is digital representation of visual information in a format suitable for processing by, e.g., the Visual Scene Description AIM.
Output Visual is visual information as rendered, e.g., by the Audio-Visual Rendering AIM.

11.1.6 Face

Face is a digital 2D or 3D representation of a human face.

11.1.7 Body

Body is a 2D or 3D digital representation of a human body, head included, face excluded.

11.1.8 Avatar

Avatar Model is a Data Type that combines the Body and Face Models.
Avatar Descriptors is a Data Type that combines Body and Face Descriptors.

11.1.9 Enhanced Transform Audio

Transform Multichannel Audio whose samples are samples of Enhanced Transform Audio.

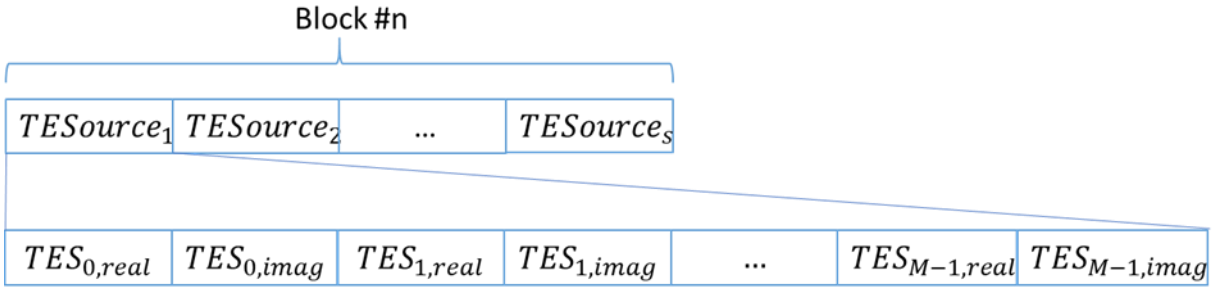

Figure 46 – Enhanced Transform Audio

11.1.10Enhanced Audio

Multichannel Audio whose samples are Enhanced Audio samples.

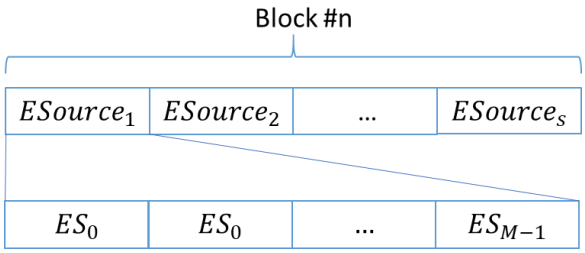

Figure 47 –Enhanced Audio

11.1.11Transform Multichannel Audio

A data structure obtained from the transformation of Multichannel Audio.

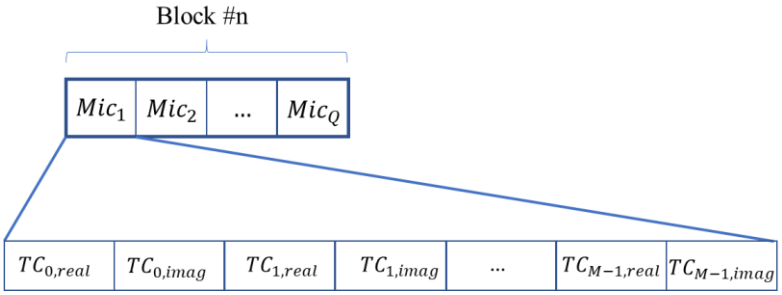

Figure 48 – Transform Multichannel Audio

11.2 Descriptors

11.2.1 Text Descriptors

See Meaning.

11.2.2 Body Descriptors

Body Descriptors conform with the Humanoid animation (HAnim) specification [11].

11.2.3 Gesture Descriptors

Gesture Descriptors are the Body Descriptors selected by an application to convey Gesture information.

11.2.4 Face Descriptors

11.2.4.1 Definition

A Data Type representing the features of an Entity's Face.
Face Descriptors are based on the Actions Units of the Facial Action Coding System (FACS) [20].

11.2.4.2 Syntax

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "$id": "https://schemas.mpai.community/PAF/V1.1/PortableAvatarFormat.json",
  "title": "FaceDescriptors",
  "type": "array",
  "items": {
    "type": "number",
    "enum": [ 1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 22, 23, 24, 25, 26,
27, 28, 41, 42, 43, 44, 45, 46, 61, 62, 63, 64 ]
  }
```

}

11.2.4.3 Semantic

Table 40 gives the semantics of the Face Descriptors.

Table 40 - Semantics of the Face Descriptors

AU	Description	Facial muscle generating the Action
1	Inner Brow Raiser	Frontalis, pars medialis
2	Outer Brow Raiser	Frontalis, pars lateralis
4	Brow Lowerer	Corrugator supercilii, Depressor supercilii
5	Upper Lid Raiser	Levator palpebrae superioris
6	Cheek Raiser	Orbicularis oculi, pars orbitalis
7	Lid Tightener	Orbicularis oculi, pars palpebralis
9	Nose Wrinkler	Levator labii superioris alaquae nasi
10	Upper Lip Raiser	Levator labii superioris
11	Nasolabial Deepener	Zygomaticus minor
12	Lip Corner Puller	Zygomaticus major
13	Cheek Puffer	Levator anguli oris (a.k.a. Caninus)
14	Dimpler	Buccinator
15	Lip Corner Depressor	Depressor anguli oris (a.k.a. Triangularis)
16	Lower Lip Depressor	Depressor labii inferioris
17	Chin Raiser	Mentalis
18	Lip Puckerer	Incisivii labii superioris and Incisivii labii inferioris
20	Lip stretcher	Risorius with platysma
22	Lip Funneler	Orbicularis oris
23	Lip Tightener	Orbicularis oris
24	Lip Pressor	Orbicularis oris
25	Lips part	Depressor labii inferioris or relaxation of Mentalis, or Orbicularis oris
26	Jaw Drop	Masseter, relaxed Temporalis and internal Pterygoid
27	Mouth Stretch	Pterygoids, Digastric
28	Lip Suck	Orbicularis oris
41	Lid droop	Relaxation of Levator palpebrae superioris
42	Slit	Orbicularis oculi
43	Eyes Closed	Relaxation of Levator palpebrae superioris; Orbicularis oculi, pars palpebralis
44	Squint	Orbicularis oculi, pars palpebralis
45	Blink	Relaxation of Levator palpebrae superioris; Orbicularis oculi, pars palpebralis
46	Wink	Relaxation of Levator palpebrae superioris; Orbicularis oculi, pars palpebralis
61	Eyes turn left	Lateral rectus, medial rectus
62	Eyes turn right	Lateral rectus, medial rectus
63	Eyes up	Superior rectus, Inferior oblique
64	Eyes down	Inferior rectus, Superior oblique

11.2.5 Prosodic Speech Descriptors

11.2.5.1 Definition

A Data Type representing the prosody of a Speech Segment in terms of pitch, duration, and intensity per phoneme.

11.2.5.2 Syntax

```
{
   "$schema":"http://json-schema.org/draft-07/schema",
   "definitions":{
      "SpeechFeatures":{
         "type":"object",
         "properties":{
            "pitch":{
               "type":"real"
            },
            "tone":{
               "type":"ToneType"
            },
            "intonation":[
               {
                  "type_p":"pitch",
                  "type_s":"speed",
                  "type_i":"intensity"
               }
            ],
            "intensity":{
               "type":"real"
            },
            "speed":{
                "type":"real" ,
            },
            "emotion":{
               "type":"EmotionType"
            },
            "NNSpeechFeatures":{
               "type":"vector of floating point"
            }
         }
      }
   },
   "type":"object",
   "properties":{
      "primary":{
         "$ref":"#/definitions/SpeechFeatureType"
      },
      "secondary":{
         "$ref":"#/definitions/SpeechFeatureType"
      }
   }
}
```

11.2.5.2.1 Neural Speech Descriptors

```
{
   "$schema":"http://json-schema.org/draft-07/schema",
   "definitions":{
      "ToneType":{
         "type":"object",
         "properties":{
            "toneName":{
               "type":"string"
            },
            "toneSetName":{
               "type":"string"
            }
         }
      },
      "type":"object",
      "properties":{
         "primary":{
```

```
            "$ref":"#/definitions/ToneType"
         },
         "secondary":{
            "$ref":"#/definitions/ToneType"
         }
      }
   }
}
```

11.2.5.3 Semantics

Name	*Definition*
SpeechFeatures	Indicates characteristic elements extracted from the input speech, specifically pitch, tone, intonation, intensity, speed, emotion, and NNspeechFeatures.
NNSpeechFeatures	Indicates specifically neural-network-based characteristic elements extracted from the input speech by Neural Network
pitch	Indicates the fundamental frequency of Speech, expressed as a real number indicating frequency as Hz (Hertz).
tone	Tone is a variation in the pitch of the voice while speaking, expressed as human readable words.
ToneType	Indicates the Tone that the input speech carries.
intonation	A variation of the pitch, intensity, and speed within a time period measured in seconds.
intensity	Energy of Speech, expressed as a real number indicating dBs (deci-bel).
speed	Indicates the Speech Rate as a real number indicating specified lin-guistic units (e.g., Phonemes, Syllables, or Words) per second.
emotion	Indicates the Emotion that the input speech carries.
EmotionType	Indicates the Emotion that the input speech carries.
toneName	Specifies the name of a Tone.
toneSetName	Name of the Tone set which contains the Tone. Tone set is used as a baseline, but other sets are possible.

Note: The semantics of "tone" defines a basic set of elements characterising tone. Elements can be added to the basic set or new sets defined using the registration procedure defined in 11.3.

Table 41 – Basic Tones

TONE CATEGORIES	**ADJEC-TIVAL**	**Semantics**
FORMALITY	formal informal	serious, official, polite everyday, relaxed, casual
ASSERTIVENESS	assertive	certain about content

	factual hesitant	neutral about content uncertain about content
REGISTER (per situation or use case)	conversational directive	appropriate to an informal speaking related to commands or requests for action

11.3 Space information

Coordinate Systems enable the specification of the position of a point by three numbers.
In a Cartesian Coordinate System, the three numbers are the signed distances from the point to three mutually perpendicular planes.
In a Spherical coordinate system, the three numbers are:

- The radial distance of that point from a fixed origin.
- The polar angle measured from a fixed zenith direction.
- The azimuthal angle of its orthogonal projection on a reference plane.

Coordinate Systems can be global or local. An Object in a Global Coordinate System may have a Local Coordinate Systems – Cartesian or Spherical. A rigid Object in a Virtual Space has a Spatial Attitude defined as the Position and Orientation and their velocities and accelerations. The Position of an Object composed of rigid Objects is that of a representative point in the Object. The notion of Spatial Attitude can also be applied to Audio Objects.

11.3.1 Spatial Attitude

11.3.1.1 Definition

A Data Type representing an Object's Position, Orientation and their velocities and accelerations.

11.3.1.2 Syntax

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Object Spatial Attitude",
  "type": "object",
  "properties": {
    "Header": {
      "type": "object",
      "properties": {
        "Standard": {
          "type": "string"
        },
        "Version": {
          "type": "integer"
        },
        "Subversion": {
          "type": "integer"
        }
      }
    },
    "OSAID": {
      "type": "string"
    },
    "General": {
      "type": "object",
      "properties": {
        "CoordType": {
          "type": "number"
        },
        "ObjectType": {
          "type": "number"
        },
        "Precision": {
```

```
      "type": "number"
    },
    "MediaType": {
      "type": "number"
    }
  }
},
"CartPosition": {
  "type": "array",
  "minItems": 3,
  "maxItems": 3,
  "items": {
    "type": "number"
  }
},
"SpherPosition": {
  "type": "array",
  "minItems": 3,
  "maxItems": 3,
  "items": {
    "type": "number"
  }
},
"Orientation": {
  "type": "array",
  "minItems": 3,
  "maxItems": 3,
  "items": {
    "type": "number"
  }
},
"CartVelociry": {
  "type": "array",
  "minItems": 3,
  "maxItems": 3,
  "items": {
    "type": "number"
  }
},
"SpherVelocity": {
  "type": "array",
  "minItems": 3,
  "maxItems": 3,
  "items": {
    "type": "number"
  }
},
"OrientVelocity": {
  "type": "array",
  "minItems": 3,
  "maxItems": 3,
  "items": {
    "type": "number"
  }
},
"CartAccel": {
  "type": "array",
  "minItems": 3,
  "maxItems": 3,
  "items": {
    "type": "number"
  }
},
"SpherAccel": {
  "type": "array",
  "minItems": 3,
  "maxItems": 3,
  "items": {
    "type": "number"
  }
},
"OrientAccel": {
  "type": "array",
  "minItems": 3,
```

```
      "maxItems": 3,
      "items": {
        "type": "number"
      }
    }
  }
}
```

11.3.1.3 Semantics

Table 42 provides the semantics of the components of the Spatial Attitude. Note that:

1. Each of Position, Velocity, and Acceleration can be expressed in Cartesian (X,Y,Z) or Spherical (r,φ,θ) Coordinates.
2. The Euler angles are indicated by (α,β,γ).

Table 42 – Components of the Object Spatial Attitude

<table>
<tr><td>HEADER</td><td>9 Bytes</td><td></td></tr>
<tr><td>• Standard</td><td>7 Bytes</td><td>The OSD-OSA string</td></tr>
<tr><td>• Version</td><td>1 Byte</td><td>Major version</td></tr>
<tr><td>• Subversion</td><td>1 Byte</td><td>Minor version</td></tr>
<tr><td>OSAID</td><td>16 Bytes</td><td>UUID Identifier of the set of Object Spatial Attitudes.</td></tr>
<tr><td>General</td><td></td><td></td></tr>
<tr><td>• CoordType</td><td>bit 0</td><td>0: Cartesian, 1: Spherical</td></tr>
<tr><td>• ObjectType</td><td>bit 1-2</td><td>00: Digital Human
01: Generic
10 and 11: reserved</td></tr>
<tr><td>• Precision</td><td>bit 3</td><td>0: single precision; 1: double precision</td></tr>
<tr><td>• MediaType</td><td>bit 4-6</td><td>000: Audio; 001: Visual; 010: Haptic; 011: Smell; 100: RADAR; 101: LiDAR; 110: Ultrasound; 111: reserved</td></tr>
<tr><td>• Reserved</td><td>bit 6-7</td><td>reserved</td></tr>
<tr><td>• SpatialAttitudeMask</td><td>2 Bytes</td><td>3*3 matrix of booleans (by rows)
<table>
<tr><td></td><td>Position</td><td>Velocity</td><td>Acceleration</td></tr>
<tr><td>Cartesian</td><td></td><td></td><td></td></tr>
<tr><td>Spherical</td><td></td><td></td><td></td></tr>
<tr><td>Orientat.</td><td></td><td></td><td></td></tr>
</table></td></tr>
<tr><td colspan="3">Position and Orientation</td></tr>
<tr><td>• CartPosition (X,Y,Z)</td><td>12/24 Bytes</td><td>Array (in metres)</td></tr>
<tr><td>• SpherPosition (r,φ,θ)</td><td>12/24 Bytes</td><td>Array (in metres and degrees)</td></tr>
<tr><td>• Orient (α,β,γ)</td><td>12/24 Bytes</td><td>Array (in degrees)</td></tr>
<tr><td colspan="3">Velocity of Position and Orientation</td></tr>
<tr><td>• CartVelocity (X,Y,Z)</td><td>12/24 Bytes</td><td>Array (in metres)</td></tr>
<tr><td>• SpherVelocity (r,φ,θ)</td><td>12/24 Bytes</td><td>Array (in metres and degrees)</td></tr>
<tr><td>• OrientVelocity (α,β,γ)</td><td>12/24 Bytes</td><td>Array (in degrees)</td></tr>
<tr><td colspan="3">Acceleration of Position and Orientation</td></tr>
<tr><td>• CartAccel (X,Y,Z)</td><td>12/24 Bytes</td><td>Array (in metres)</td></tr>
<tr><td>• SpherAccel (r,φ,θ)</td><td>12/24 Bytes</td><td>Array (in metres and degrees)</td></tr>
<tr><td>• OrientAccel (α,β,γ)</td><td>12/24 Bytes</td><td>Array (in degrees)</td></tr>
</table>

11.3.2 Microphone Array Geometry

11.3.2.1 Definition

A Data Type representing the position of each microphone comprising a microphone array and characteristics such as microphone type, look directions, and array type.

11.3.2.2 Syntax

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Microphone Array Geometry",
  "type": "object",
  "properties": {
    "Header": {
      "type": "object",
      "properties": {
        "Standard": {
          "type": "string"
        },
        "Version": {
          "type": "integer"
        },
        "Subversion": {
          "type": "integer"
        }
      }
    },
    "MAGID": {
      "type": "string"
    },
    "MicrophoneFeatures": {
      "type": "object",
      "properties": {
        "ArrayType": {
          "type": "integer"
        },
        "ArrayScat": {
          "type": "integer"
        },
        "ArrayFilterURI": {
          "type": "string",
          "format": "uri"
        }
      }
    },
    "SamplingFeatures": {
      "type": "object",
      "properties": {
        "SamplingRate": {
          "type": "integer"
        },
        "SampleType": {
          "type": "integer"
        }
      }
    },
    "BlockSize": {
      "type": "integer"
    },
    "NumberofMicrophones": {
      "type": "integer"
    },
    "Microphoneattributes": {
      "type": "array",
      "items": {
        "type": "object",
        "properties": {
          "xCoord": {
            "type": "number"
          },
          "yCoord": {
```

```
            "type": "number"
          },
          "zCoord": {
            "type": "number"
          },
            "directivity": {
              "type": "integer"
            },
            "micxLookCoord": {
              "type": "number"
            },
            "micyLookCoord": {
              "type": "number"
            },
            "miczLookCoord": {
              "type": "number"
            }
          }
        },
        "minItems": 4,
        "uniqueItems": true,
        "required": [
          "xCoord",
          "yCoord",
          "zCoord",
          "directivity",
          "micxLookCoord",
          "micyLookCoord",
          "miczLookCoord"
        ]
      },
      "MicrophoneArrayLookCoord": {
        "type": "object",
        "properties": {
          "xLookCoord": {
            "type": "number"
          },
          "yLookCoord": {
            "type": "number"
          },
          "zLookCoord": {
            "type": "number"
          }
        },
        "uniqueItems": true,
        "required": [
          "xLookCoord",
          "yLookCoord",
          "zLookCoord"
        ]
      }
    },
    "required": [
      "MicrophoneArrayType",
      "MicrophoneArrayScat",
      "MicrophoneArrayFilterURI",
      "SamplingRate",
      "SampleType",
      "BlockSize",
      "NumberofMicrophones",
      "MicrophoneList",
      "MicrophoneArrayLookCoord"
    ]
  }
```

11.3.2.3 Semantics

Table 43 gives the Semantics of Microphone Array Geometry.

Table 43 – Semantics of Microphone Array Geometry

Label	Size	Description
HEADER	9 Bytes	
• Standard	7 Bytes	The CAE-MAG string
• Version	1 Byte	Major version
• Subversion	1 Byte	Minor version
MAGID	16 Bytes	UUID Identifier of the Microphone Array Geometry.
Microphone features		
• ArrayType	bit 0-1	Indicates the type of microphone array positioning such as 00: Spherical, 01: Circular, 10: Planar, 11: Linear.
• ArrayScat	bit 2	Indicates the type of the microphone array (0: Rigid, 1: Open).
• Reserved	bit 2-7	
• ArrayFilterURI	N Bytes	A uniform resource identifier (URI) string identifying the path to a local or remote file containing specific filter coefficients of the microphone array to be used for equalisation.
Sampling features		
• SamplingRate	0-3 bits	0:8, 1:16, 2:24, 3:32, 4:44.1, 5:48, 6: 64, 7: 96, 8: 192 (all kHz)
• SampleType	4-5 bits	0:16, 1:24, 2:32, 3:64 (all bits/sample)
• Reserved	bit 6-7	
BlockSize	4 Bytes	Minimum BlockSize: ≥ 256.
NumberOfMicrophones		
MicrophoneAttributes		A list containing Microphone attributes.
• MicrophoneID	1 Byte	
• xCoord	4 Bytes	x position of the microphone in m. (number)
• yCoord	4 Bytes	y position of the microphone in m.(number)
• zCoord	4 Bytes	z position of the microphone in m. (number)
• directivity	bit 0-2	The directivity pattern of the specific microphone, 000: omnidirectional, 001: figure of eight, 010: cardioid, 011: supercardioid, 100: hypercardioid (uint8)
• Reserved	Bit 3-7	
• micxLookCoord	4 Bytes	x component of the vector representing the look direction of the microphone in m. (number)
• micyLookCoord	4 Bytes	y component of the vector representing the look direction of the microphone in m. (number)
• miczLookCoord	4 Bytes	z component of the vector representing the look direction of the microphone. (number)
MicrophoneArrayLookCoord		
xLookCoord	4 Bytes	x component of the vector representing the look direction of the microphone array. (number)
yLookCoord	4 Bytes	y component of the vector representing the look direction of the microphone array. (number)
zLookCoord	4 Bytes	z component of the vector representing the look direction of the microphone array. (number)

11.3.3 Audio Scene Geometry

11.3.3.1 Definition

A Data Type representing the spatial arrangement of the Audio Objects of a Scene.

11.3.3.2 Syntax

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Audio Scene Geometry",
  "type": "object",
  "properties": {
    "Header": {
      "type": "object",
      "properties": {
        "Standard": {
          "type": "string"
        },
        "Version": {
          "type": "integer"
        },
        "Subversion": {
          "type": "integer"
        }
      }
    },
    "ASDID": {
      "type": "string"
    },
    "Time": {
      "type": "object",
      "properties": {
        "TimeType": {
          "type": "boolean"
        },
        "StartTime": {
          "type": "number"
        },
        "EndTime": {
          "type": "number"
        }
      }
    },
    "BlockSize": {
      "type": "integer"
    },
    "AudioObjectCount": {
      "type": "integer"
    },
    "AudioObjectsData": {
      "type": "object",
      "properties": {
        "AudioObjectID": {
          "type": "string"
        },
        "SpatialAttitude": {
          "$ref": "https://schemas.mpai.community/OSD/V1.0/data/SpatialAttitude.json"
        }
      }
    }
  }
}
```

11.3.3.3 Semantics

Table 44 provides the semantics of the Audio Scene Geometry.

Table 44 – Audio Scene Geometry Semantics

Label	Size	Description

HEADER	9 Bytes	
• Standard	7 Bytes	The string CAE-ASD
• Version	1 Byte	Major version
• Subversion	1 Byte	Minor version
ASDID	16 Bytes	UUID Identifier of Audio Scene Descriptors set.
Time	17 Bytes	Collects various data expressed with bits
• TimeType	0 bit	0=Relative: time starts at 0000/00/00T00:00 1=Absolute: time starts at 1970/01/01T00:00.
• Reserved	1-7 bits	reserved
• StartTime	8 Bytes	Start of current Audio Scene Descriptors (in µs).
• EndTime	8 Bytes	End of current Audio Scene Descriptors (in µs).
BlockSize	4 Bytes	Minimum BlockSize: ≥ 256.
AudioObjectCount	1 Byte	Number of Audio Objects in the Audio Scene.
AudioObjectsData	N1 Bytes	Data associated to each Audio Object.
• AudioObjectID	1 Byte	ID of a specific Audio Object in the Audio Scene.
• SamplingRate	0-3 bits	0:8, 1:16, 2:24, 3:32, 4:44.1, 5:48, 6: 64, 7: 96, 8: 192 (all kHz)
• SampleType	4-5 bits	0:16, 1:24, 2:32, 3:64 (all bits/sample)
• Reserved	6-7 bits	
• Spatial Attitude	N2 Bytes	

11.3.4 Audio Object

11.3.4.1 Definition

A Data Type representing an Object that can be rendered to and perceived by a human ear.

11.3.4.2 Syntax

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "AudioObject",
  "type": "object",
  "properties": {
    "Header": {
      "type": "object",
      "properties": {
        "Standard": {
          "type": "string"
        },
        "Version": {
          "type": "integer"
        },
        "Subversion": {
          "type": "integer"
        }
      }
    },
    "AOBID": {
      "type": "string"
    },
    "AudioObjectsData": {
      "type": "object",
      "properties": {
        "AudioObject": {
          "type": "object",
          "properties": {
            "FormatID": {
              "type": "integer"
            },
            "ObjectLength": {
```

```
            "type": "integer"
          },
          "DataInObject": {
            "$ref": "https://schemas.mpai.community/CAE/V2.1/data/AudioObject.json"
          }
        }
      }
    }
  }
}
```

11.3.4.3 Semantics

Label	Size	Description
HEADER	9 Bytes	
• Standard	7 Bytes	The string CAE-ASD
• Version	1 Byte	Major version
• Subversion	1 Byte	Minor version
AOBID	16 Bytes	UUID Identifier of the Audio Object.
AudioObjectData	N1 Bytes	Data associated to each Audio Object.
• Reserved	6-7 bits	
• AudioObject	N2 Bytes	
○ FormatID	1 Byte	Audio Object Format Identifier
○ ObjectLength	4 Bytes	Number of Bytes in Audio Object
○ DataInObject	N3 Bytes	Data of Audio Object

11.3.5 Audio Scene Descriptors

11.3.5.1 Definition

A Data Type that includes the Geometry and the Objects of an Audio Scene.

11.3.5.2 Syntax

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Audio Scene Descriptors",
  "type": "object",
  "properties": {
    "Header": {
      "type": "object",
      "properties": {
        "Standard": {
          "type": "string"
        },
        "Version": {
          "type": "integer"
        },
        "Subversion": {
          "type": "integer"
        }
      }
    },
    "ASDID": {
      "type": "string"
    },
    "Time": {
      "type": "object",
      "properties": {
        "TimeType": {
          "type": "boolean"
        },
        "StartTime": {
```

```
          "type": "number"
        },
        "EndTime": {
          "type": "number"
        }
      }
    },
    "BlockSize": {
      "type": "integer"
    },
    "AudioObjectCount": {
      "type": "integer"
    },
    "AudioObjectsData": {
      "type": "object",
      "properties": {
        "AudioObjectID": {
          "type": "string"
        },
        "SamplingRate": {
          "type": "number"
        },
        "SamplingType": {
          "type": "number"
        },
        "SpatialAttitude": {
          "$ref": "https://schemas.mpai.community/OSD/V1.0/data/SpatialAttitude.json"
        },
        "AudioObject": {
          "type": "object",
          "properties": {
            "FormatID": {
              "type": "integer"
            },
            "ObjectLength": {
              "type": "integer"
            },
            "DataInObject": {
              "$ref": "https://schemas.mpai.community/CAE/V2.1/data/AudioObject.json"
            }
          }
        }
      }
    }
  }
}
```

11.3.5.3 Semantics

Table 45 provides the semantics of Audio Scene Descriptors.

Table 45 – Audio Scene Descriptors

Label	Size	Description
HEADER	9 Bytes	
• Standard	7 Bytes	The string CAE-ASD
• Version	1 Byte	Major version
• Subversion	1 Byte	Minor version
ASDID	16 Bytes	UUID Identifier of Audio Scene Descriptors set.
Time	17 Bytes	Collects various data expressed with bits
• TimeType	0 bit	0=Relative: time starts at 0000/00/00T00:00 1=Absolute: time starts at 1970/01/01T00:00.
• Reserved	1-7 bits	reserved
• StartTime	8 Bytes	Start of current Audio Scene Descriptors (in µs).
• EndTime	8 Bytes	End of current Audio Scene Descriptors (in µs).

BlockSize	4 Bytes	Minimum BlockSize: ≥ 256.
AudioObjectCount	1 Byte	Number of Audio Objects in the Audio Scene.
AudioObjectsData	N1 Bytes	Data associated to each Audio Object.
• AudioObjectID	1 Byte	ID of a specific Audio Object in the Audio Scene.
• SamplingRate	0-3 bits	0:8, 1:16, 2:24, 3:32, 4:44.1, 5:48, 6: 64, 7: 96, 8: 192 (all kHz)
• SampleType	4-5 bits	0:16, 1:24, 2:32, 3:64 (all bits/sample)
• Reserved	6-7 bits	
• Spatial Attitude	N2 Bytes	
• AudioObject	N3 Bytes	
○ FormatID	1 Byte	Audio Object Format Identifier
○ ObjectLength	4 Bytes	Number of Bytes in Audio Object
○ DataInObject	N4 Bytes	Data of Audio Object

11.3.6 Visual Scene Geometry

11.3.6.1 Definition

A Data Type representing the spatial arrangement of the Visual Objects of a Scene.

11.3.6.2 Syntax

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Visual Scene Geometry",
  "type": "object",
  "properties": {
    "Header": {
      "type": "object",
      "properties": {
        "Standard": {
          "type": "string"
        },
        "Version": {
          "type": "integer"
        },
        "Subversion": {
          "type": "integer"
        }
      }
    },
    "VSGID": {
      "type": "string"
    },
    "Time": {
      "type": "object",
      "properties": {
        "TimeType": {
          "type": "boolean"
        },
        "StartTime": {
          "type": "number"
        },
        "EndTime": {
          "type": "number"
        }
      }
    },
    "VisualObjectCount": {
      "type": "integer"
    },
    "VisualObjectsData": {
      "type": "object",
      "properties": {
        "VisualObjectID": {
```

```
          "type": "string"
        },
        "SpatialAttitude": {
          "$ref": "https://schemas.mpai.community/OSD/V1.0/data/SpatialAttitude.json"
        }
      }
    }
  }
}
```

11.3.6.3 Semantics

Table 46 provides the semantics of Visual Scene Geometry.

Table 46 – Semantics of Visual Scene Geometry

Label	Size	Description
HEADER	9 Bytes	
• Standard	7 Bytes	The string OSD-VSD
• Version	1 Byte	Major version
• Subversion	1 Byte	Minor
VSGID	16 Bytes	UUID Identifier of the total set of Visual Scene Geometries (uuid).
Time	17 Bytes	Collects various data expressed with bits
• TimeType	0 bit	0=Relative: time starts at 0000/00/00T00:00 1=Absolute: time starts at 1970/01/01T00:00.
• Reserved	1-7 bits	reserved
• StartTime	8 Bytes	Start time of current Visual Scene Descriptors (in microseconds).
• EndTime	8 Bytes	End time of current Visual Scene Descriptors (in microseconds).
VisualObjectCount	1 Byte	Number of Visual Objects in Visual Scene.
VisualObjectsData	N1 Bytes	Data associated to each Visual Object.
• VisualObjectID	1 Byte	ID of a specific Visual Object in a Visual Scene.
• Reserved	1 Byte	
• SpatialAttitude	N2 Bytes	Spatial Attitude of each Object

11.3.7 Visual Object

11.3.7.1 Definition

The digital representation of an object captured from an electromagnetic or high-frequency audio signal or computer-generated that can be rendered to and perceived by a human eye.

11.3.7.2 Syntax

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "VisualObject",
  "type": "object",
  "properties": {
    "Header": {
      "type": "object",
      "properties": {
        "Standard": {
          "type": "string"
        },
        "Version": {
          "type": "integer"
        },
```

```
        "Subversion": {
          "type": "integer"
        }
      }
    },
    "VOBID": {
      "type": "string"
    },
    "VisualObjectsData": {
      "type": "object",
      "properties": {
        "FormatID": {
          "type": "integer"
        },
        "ObjectLength": {
          "type": "integer"
        },
        "DataInObject": {
          "$ref": "https://schemas.mpai.community/OSD/V1.0/data/VisualObject.json"
        }
      }
    }
  }
}
```

11.3.7.3 Semantics

Label	Size	Description
HEADER	9 Bytes	
• Standard	7 Bytes	The string CAE-ASD
• Version	1 Byte	Major version
• Subversion	1 Byte	Minor version
VOBID	16 Bytes	UUID Identifier of the Visual Object.
VisualObjectData	N1 Bytes	Data associated to each Visual Object.
• VisualObject	N2 Bytes	
○ FormatID	1 Byte	Audio Object Format Identifier
○ ObjectLength	4 Bytes	Number of Bytes in Audio Object
○ DataInObject	N3 Bytes	Data of Audio Object

11.3.8 Visual Scene Descriptors

11.3.8.1 Definition

A Data Type representing the Audio-Visual Objects and their spatial arrangement in an Audio-Visual Scene.

11.3.8.2 Syntax

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Visual Scene Descriptors",
  "type": "object",
  "properties": {
    "Header": {
      "type": "object",
      "properties": {
        "Standard": {
          "type": "string"
        },
        "Version": {
          "type": "integer"
        },
        "Subversion": {
          "type": "number"
```

```
        }
      }
    },
    "VSDID": {
      "type": "string"
    },
    "Time": {
      "type": "object",
      "properties": {
        "TimeType": {
          "type": "boolean"
        },
        "StartTime": {
          "type": "number"
        },
        "EndTime": {
          "type": "number"
        }
      }
    },
    "VisualObjectCount": {
      "type": "integer"
    },
    "VisualObjectsData": {
      "type": "object",
      "properties": {
        "VisualObjectID": {
          "type": "string"
        },
        "SpatialAttitude": {
          "$ref": "https://schemas.mpai.community/OSD/V1.0/data/SpatialAttitude.json"
        },
        "VisualObject": {
          "type": "object",
          "properties": {
            "FormatID": {
              "type": "integer"
            },
            "ObjectLength": {
              "type": "integer"
            },
            "DataInObject": {
              "$ref": "https://schemas.mpai.community/OSD/V1.0/data/VisualObject.json"
            }
          }
        }
      }
    }
  }
}
```

11.3.8.3 Semantics

Table 47 provides the semantics of Visual Scene Descriptors.

Table 47 – Visual Scene Descriptors Semantics

Label	Size	Description
HEADER	9 Bytes	
• Standard	7 Bytes	The string OSD-VSD
• Version	1 Byte	Major version
• Subversion	1 Byte	Minor
VSDID	16 Bytes	UUID Identifier of the total set of Visual Scene Descriptors (uuid).
Time	17 Bytes	Collects various data expressed with bits
• TimeType	0 bit	0=Relative: time starts at 0000/00/00T00:00 1=Absolute: time starts at 1970/01/01T00:00.

• Reserved	1-7 bits	reserved
• StartTime	8 Bytes	Start time of current Visual Scene Descriptors (in microseconds).
• EndTime	8 Bytes	End time of current Visual Scene Descriptors (in microseconds).
VisualObjectCount	1 Byte	Number of Visual Objects in Visual Scene.
VisualObjectsData	N1 Bytes	Data associated to each Visual Object.
• VisualObjectID	1 Byte	ID of a specific Visual Object in a Visual Scene.
• Reserved	1 Byte	
• SpatialAttitude	N2 Bytes	According to MPAI-OSD V1
• VisualObject	N3 Bytes	
○ FormatID	1 Byte	Visual Object Format Identifier
○ Length	4 Bytes	Number of Bytes in Visual Object
○ DataInObject	N4 Bytes	Data of Visual Object

11.3.9 Audio-Visual Scene Geometry

11.3.9.1 Definition

A Data Type representing the spatial arrangement of the Audio, Visual, and Audio-Visual Objects of a Scene.

11.3.9.2 Syntax

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Audio-Visual Scene Geometry",
  "type": "object",
  "properties": {
    "Header": {
      "type": "object",
      "properties": {
        "Standard": {
          "type": "string"
        },
        "Version": {
          "type": "integer"
        },
        "Subversion": {
          "type": "integer"
        }
      }
    },
    "AVGID": {
      "type": "string"
    },
    "Time": {
      "type": "object",
      "properties": {
        "TimeType": {
          "type": "boolean"
        },
        "StartTime": {
          "type": "number"
        },
        "EndTime": {
          "type": "number"
        }
      }
    },
    "AVObjectCount": {
      "type": "integer"
    },
    "AVObjectsData": {
```

```
      "type": "object",
      "properties": {
        "AVObjectID": {
          "type": "string"
        },
        "SpatialAttitude": {
          "$ref": "https://schemas.mpai.community/OSD/V1.0/data/SpatialAttitude.json"
        }
      }
    }
  }
}
```

11.3.9.3 Semantics

Table 48 provides the semantics of the Audio-Visual Scene Geometry.

Table 48 – Audio-Visual Scene Geometry

Label	Size	Description
HEADER	9 Bytes	
• Standard	7 Bytes	The string OSD-AVG
• Version	1 Byte	Major version
• Subversion	1 Byte	Minor
AVGID	16 Bytes	UUID Identifier of the total set of Audio-Visual Scene Geometries.
Time	17 Bytes	Collects various data expressed with bits
• TimeType	0 bit	0=Relative: time starts at 0000/00/00T00:00 1=Absolute: time starts at 1970/01/01T00:00.
• Reserved	1-7 bits	reserved
• StartTime	8 Bytes	Start time of current Audio-Visual Scene Descriptors (in microseconds).
• EndTime	8 Bytes	End time of current Audio-Visual Scene Descriptors (in microseconds).
AVObjectCount	1 Byte	Number of Objects in Scene.
AVObjectData	N1 Bytes	Data associated to each Object.
• AVObjectID	1 Byte	ID of a specific Object in the Scene.
• SpatialAttitude	N2 Bytes	

11.3.10Audio-Visual Scene Descriptors

11.3.10.1 Definition

A Data Type representing the Audio-Visual Objects and their spatial arrangement in an Audio-Visual Scene.

11.3.10.2 Syntax

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Audio-Visual Scene Descriptors",
  "type": "object",
  "properties": {
    "Header": {
      "type": "object",
      "properties": {
        "Standard": {
          "type": "string"
        },
        "Version": {
```

```
          "type": "integer"
        },
        "Subversion": {
          "type": "integer"
        }
      }
    },
    "AVSID": {
      "type": "string"
    },
    "Time": {
      "type": "object",
      "properties": {
        "TimeType": {
          "type": "boolean"
        },
        "StartTime": {
          "type": "number"
        },
        "EndTime": {
          "type": "number"
        }
      }
    },
    "AVObjectCount": {
      "type": "integer"
    },
    "AVObjectsData": {
      "type": "object",
      "properties": {
        "AVObjectID": {
          "type": "string"
        },
        "SamplingRate": {
          "type": "number"
        },
        "SamplingType": {
          "type": "number"
        },
        "SpatialAttitude": {
          "$ref": "https://schemas.mpai.community/OSD/V1.0/data/SpatialAttitude.json"
        },
        "AVObject": {
          "type": "object",
          "properties": {
            "FormatID": {
              "type": "integer"
            },
            "ObjectLength": {
              "type": "integer"
            },
            "DataInAObject": {
              "$ref": "https://schemas.mpai.community/CAE/V2.1/data/AudioObject.json"
            },
            "DataInVObject": {
              "$ref": "https://schemas.mpai.community/OSD/V1.0/data/VisualObject.json"
            }
          }
        }
      }
    }
  }
}
```

11.3.10.3 Semantics

Table 49 provides the semantics of the Audio-Visual Scene Descriptors.

Table 49 – Audio-Visual Scene Descriptors

Label	Size	Description

HEADER	9 Bytes	
• Standard	7 Bytes	The string OSD-AVS
• Version	1 Byte	Major version
• Subversion	1 Byte	Minor
AVDID	16 Bytes	UUID Identifier of the total set of Audio-Visual Scene Descriptors.
Time	17 Bytes	Collects various data expressed with bits
• TimeType	0 bit	0=Relative: time starts at 0000/00/00T00:00 1=Absolute: time starts at 1970/01/01T00:00.
• Reserved	1-7 bits	reserved
• StartTime	8 Bytes	Start time of current Audio-Visual Scene Descriptors (in microseconds).
• EndTime	8 Bytes	End time of current Audio-Visual Scene Descriptors (in microseconds).
AVObjectCount	1 Byte	Number of Objects in Scene.
AVObjectData	N1 Bytes	Data associated to each Object.
• AVObjectID	1 Byte	ID of a specific Object in the Scene.
• SamplingRate	0-3 bits	0: 8kHz, 1: 16kHz, 2: 24kHz, 3: 32kHz, 4: 44.1kHz, 5: 48kHz, 6: 64kHz, 7: 96kHz, 8: 192kHz
• SampleType	4-5 bits	0:16bit, 1:24bit, 2:32bit, 3:64bit)
• Reserved	6-7 bits	
• SpatialAttitude	N2 Bytes	According to MPAI-OSD V1
• *AudioObject*	N3 Bytes	
○ FormatID	1 Byte	Audio Object Format Identifier
○ Length	4 Bytes	Number of Bytes in Audio Object
○ DataInObject	N4 Bytes	Data of Audio Object
• *VisualObject*	N5 Bytes	
○ FormatID	1 Byte	Visual Object Format Identifier
○ Length	4 Bytes	Number of Bytes in Audio Object
○ DataInObject	N6 Bytes	Data of Visual Object

11.4 Personal Status

11.4.1 Definitions

Personal Status is a data structure composed of three Personal Status *Factors*:

1. Emotion (such as “angry” or “sad”).
2. Cognitive State (such as “surprised” or “interested”).
3. Social Attitude (such as “polite” or “arrogant”).

Factors can be expressed via several Personal Status *Modalities*: Text, Speech, Face, and Gestures. Other Modalities, such as body posture, are currently not supported and may be added to future Versions of MPAI Technical Specifications.

Within a given Modality, the Factors can be analysed and interpreted via various *Descriptors*. For example, when expressed via Speech, the elements may be expressed through combinations of such features as prosody (pitch, rhythm, and volume variations); separable speech effects (such as degrees of voice tension, breathiness, etc.); and vocal gestures (laughs, sobs, etc.).

Each of the three Factors (Emotion, Cognitive State, and Social Attitude) is represented by a standard set of labels and associated semantics. For each of these Factors, two tables are provided:

- A *Label Set Table* containing descriptive labels relevant to the Factor in a three-level format:
  - The CATEGORIES column specifies the relevant categories using nouns (e.g., "ANGER").
  - The GENERAL ADJECTIVAL column gives adjectival labels for general or basic labels within a category (e.g., "angry").
  - The SPECIFIC ADJECTIVAL column gives more specific (sub-categorised) labels in the relevant category (e.g., "furious").
- A *Label Semantics Table* providing the semantics for each label in the GENERAL ADJECTIVAL and SPECIFIC ADJECTIVAL columns of the Label Set Table. For example, for "angry" the semantic gloss is "emotion due to perception of physical or emotional damage or threat."

These sets have been compiled in the interests of basic cooperation and coordination among AIM submitters and vendors, complemented by a procedure whereby AIM submitters may propose extended or alternate sets for their purposes.

An Implementer wishing to extend or replace a *Label Set Table* for one of the three Factors is requested to do the following:

1. Create a new *Label Set Table* where:
   a. Proposed additions are clearly marked (in case of extension).
   b. All the elements of the target Factor and levels (up to 3) are listed (in case of replacement).
2. Create a new *Label Semantics Table* where the semantics of elements of the target Factor is:
   a. Added to the semantics of the existing target Factor (in case of extension).
   b. Provided (in case of replacement).

   The submitted semantics should have a level of detail comparable to the semantics given in the current *Label Semantics Table*.
3. Submit both tables to the MPAI Secretariat (secretariat@mpai.community).

The appropriate MPAI Development Committee will examine the proposed extension or replacement. Only the adequacy of the proposed new tables in terms of clarity and completeness will be considered. In case the new tables are not clear or complete, a revision of the tables will be requested.

The accepted External Factor Set will be identified as proposed by the submitter, reviewed by the appropriate MPAI Committee, and posted to the MPAI web site.

The versioning system is based on a name – MPAI for MPAI-generated versions or "organisation name" for the proposing organisation – with a suffix m.n where m indicates the version and n indicates the subversion.

11.4.2 Syntax

```
{
    "$id": "https://schemas.mpai.community/MMC/V2.0/PersonalStatus.json",
    "$schema": "http://json-schema.org/draft-07/schema#",
    "title": "Personal Status",
    "type": "object",
    "properties": {
        "Timestamp": {
            "type": "object",
            "properties": {
                "Timestamp type": {
                    "type": "string"
```

```
             },
            "Timestamp value": {
                "type": "string",
                "oneOf": [
                    { "format" : "date-time" },
                    { "const" : "0" }
                ]
            }
        },
        "required": ["Timestamp value"],
        "if": {
            "properties": { "Timestamp value": { "const": "0" } }
        },
        "then": {
            "properties": { "Timestamp type": { "type": "null" } }
        },
        "else": {
            "required": ["Timestamp type"]
        }
    },
    "emotion": {
        "type": "object",
        "properties": {
            "Fused emotion value": { "type": "number", "minimum": 0 },
            "Text emotion value": { "type": "number", "minimum": 0 },
            "Speech emotion value": { "type": "number", "minimum": 0 },
            "Face emotion value": { "type": "number", "minimum": 0 },
            "Gesture emotion value": { "type": "number", "minimum": 0 },
            "emotion version": {
              "type": "string",
              "pattern": "^[A-Za-z]+-\\d+\\.\\d+$"
            }
        },
        "anyOf": [
            { "required": ["emotion version", "Fused emotion value"] },
            { "required": ["emotion version", "Text emotion value"] },
            { "required": ["emotion version", "Speech emotion value"] },
            { "required": ["emotion version", "Face emotion value"] },
            { "required": ["emotion version", "Gesture emotion value"] }
        ]
    },
     "cogstate": {
        "type": "object",
        "properties": {
            "Fused cogstate value": { "type": "number", "minimum": 0 },
            "Text cogstate value": { "type": "number", "minimum": 0 },
            "Speech cogstate value": { "type": "number", "minimum": 0 },
            "Face cogstate value": { "type": "number", "minimum": 0 },
            "Gesture cogstate value": { "type": "number", "minimum": 0 },
            "cogstate version": {
              "type": "string" ,
              "pattern": "^[A-Za-z]+-\\d+\\.\\d+$"
            }
        },
        "anyOf": [
            { "required": ["cogstate version", "Fused cogstate value"] },
            { "required": ["cogstate version", "Text cogstate value"] },
            { "required": ["cogstate version", "Speech cogstate value"] },
            { "required": ["cogstate version", "Face cogstate value"] },
            { "required": ["cogstate version", "Gesture cogstate value"] }
        ]
},
    "attitude": {
        "type": "object",
        "properties": {
            "Fused attitude value": { "type": "number", "minimum": 0 },
            "Text attitude value": { "type": "number", "minimum": 0 },
            "Speech attitude value": { "type": "number", "minimum": 0 },
            "Face attitude value": { "type": "number", "minimum": 0 },
            "Gesture attitude value": { "type": "number", "minimum": 0 },
            "attitude version": {
              "type": "string",
              "pattern": "^[A-Za-z]+-\\d+\\.\\d+$"
            }
```

```
            },
            "anyOf": [
                { "required": ["attitude version", "Fused attitude value"] },
                { "required": ["attitude version", "Text attitude value"] },
                { "required": ["attitude version", "Speech attitude value"] },
                { "required": ["attitude version", "Face attitude value"] },
                { "required": ["attitude version", "Gesture attitude value"] }
            ]
        }
    },
    "required" : ["cogstate"],
    "required" : ["attitude"],
    "required" : ["emotion"]
}
```

11.4.3 Semantics

1. *Timestamp type* can either be:
1.1. Absolute time (A)
1.2. Relative time, i.e., time from the start of operation (R)
2. *Timestamp value* is as in CAE V1.
2.1. *18 values of Personal Status* that include (see Table 50)
2.1.1. 6 cells for Emotion.
2.1.2. 6 cells for Cognitive State.
2.1.3. 6 cells for Social Attitude.

Table 50 - The table of (Factor, Modality) cells

		Modality					
		Version	Fused value	Text	Speech	Face	Gesture
Factor	Emotion	V.Emotion					
	Cognitive State	V.Cognitive					
	Social Attitude	V.Attitude					

3. The 18 values in the cells are represented as a vector of 18 values, 6 for each Factor:
3.1. The first value is the Version of Emotion/Cognitive State/Social Attitude (VE/VC/VA) represented as two fields:
3.1.1. Field 1: 2 digits of the Version of the MMC standard (e.g., “12”, meaning version 1.2, is expressed as 2 bytes).
3.1.2. Field 2: The sequential number of the Factor dataset. Currently, there is one dataset, given the number 1. New submissions will receive sequential numbers starting from 2, where the sequential number of the dataset is expressed with 1 byte).
3.2. The second value is the current default fused value of the Modality.
3.3. Followed by the 4 values of the Modality.
3.3.1. The value of Text
3.3.2. The value of Speech
3.3.3. The value of Face
3.3.4. The value of Gesture
3.4. The list of possible values of a Modality are (values are in bytes):
3.4.1. Value 0: unable to compute for any reason; error; or no discernible value.
3.4.2. Value 1 up to the largest number of Factor values in the relevant Label Semantics Table.

Therefore, a value of Personal Status is represented by the following table. Timestamp, Emotion, Cognitive State, Social Attitude. Their Descriptors are also present if the information is available.

Table 51 – The information included in the Personal Status

Variable name	Code
Timestamp	Timestamp type
	Timestamp value
Emotion	Emotion version
	Fused Emotion value
	Text Emotion value
	Speech Emotion value
	Face Emotion value
	Gesture Emotion value
Cognitive State	Cognitive State version
	Fused Cognitive State value
	Text Cognitive State value
	Speech Cognitive State value
	Face Cognitive State value
	Gesture Cognitive State value
Social Attitude	Social Attitude version
	Fused Social Attitude value
	Text Social Attitude value
	Speech Social Attitude value
	Face Social Attitude value
	Gesture Social Attitude value

11.4.4 Cognitive State

11.4.4.1 Definition

A Data Type representing an Entity's internal state that reflects the way it understands the Context, such as "Confused", "Dubious", "Convinced". Primary Cognitive State corresponds to General Adjectival and Secondary Cognitive State corresponds to Specific Adjectival in *Table 52*.

11.4.4.2 Syntax

Cognitive State is represented by.

```
{
   "$schema":"http://json-schema.org/draft-07/schema",
   "definitions":{
      "cogstateType":{
         "type":"object",
         "properties":{
            "cogstateDegree":{
              "enum": ["High", "Medium", "Low"]
            },
            "cogstateName":{
               "type":"number"
            },
            "cogstateSetName":{
               "type":"string"
            }
         }
      },
      "type":"object",
      "properties":{
         "primary":{
            "$ref":"#/definitions/cogstateType"
         },
         "secondary":{
            "$ref":"#/definitions/cogstateType"
         }
```

```
    }
}
```

11.4.4.3 Semantics

Name	*Definition*
`cogstateType`	Specifies the Cognitive State that the input carries.
`cogstateDegree`	Specifies the Degree of Cognitive State as one of "Low," "Medium," and "High."
`cogstateName`	Specifies the ID of a Cognitive State listed in *Table 55*.
`cogstateSetName`	Specifies the name of the Cognitive State set which contains the Cognitive State. Cognitive State set of *Table 55* is used as a baseline, but other sets are possible.

Table 52 gives the standardised three-level Basic Cognitive State Label Set.

Table 52 – Basic Cognitive State Label Set

COGNITIVE CATEGORIES	GENERAL ADJECTIVAL	SPECIFIC ADJECTIVAL
AROUSAL	aroused/excited/energetic	cheerful playful lethargic sleepy
ATTENTION	attentive	expectant/anticipating thoughtful distracted/absent-minded vigilant hopeful/optimistic
BELIEF	credulous	sceptical
INTEREST	interested	fascinated curious bored
SURPRISE	surprised	astounded startled
UNDERSTANDING	comprehending	uncomprehending bewildered/puzzled

Table 53 provides the semantics for each label in the GENERAL ADJECTIVAL and SPECIFIC ADJECTIVAL columns above.

Table 53 – Basic Cognitive State Semantics Set

ID	Cognitive State	Meaning
1	aroused/excited/energetic	cognitive state of alertness and energy
2	astounded	high degree of surprised
3	attentive	cognitive state of paying attention
4	bewildered/puzzled	high degree of incomprehension
5	bored	not interested

6	cheerful	energetic combined with and communicating happiness
7	comprehending	cognitive state of successful application of mental models to a situation
8	credulous	cognitive state of conformance to mental models of a situation
9	curious	interest due to drive to know or understand
10	distracted/absent-minded	not attentive to present situation due to competing thoughts
11	expectant/anticipating	attentive to (expecting) future event or events
12	fascinated	high degree of interest
13	interested	cognitive state of attentiveness due to salience or appeal to emotions or drives
14	lethargic	not aroused
15	playful	energetic and communicating willingness to play
16	sceptical	not credulous
17	sleepy	not aroused due to need for sleep
18	surprised	cognitive state due to violation of expectation
19	startled	surprised by a sudden event or perception
20	surprised	cognitive state due to violation of expectation
21	thoughtful	attentive to thoughts
22	uncomprehending	not comprehending

11.4.5 Emotion

11.4.5.1 Definition

A Data Type representing an Entity's internal state that results from its interaction with the Context, such as "Angry", "Sad", "Determined".
Primary Emotion corresponds to General Adjectival and Secondary Emotion corresponds to Specific Adjectival in *Table 54.*

11.4.5.2 Syntax

```
{
   "$schema":"http://json-schema.org/draft-07/schema",
   "definitions":{
      "emotionType":{
         "type":"object",
         "properties":{
            "emotionDegree":{
              "enum": ["High", "Medium", "Low"]
            },
            "emotionName":{
               "type":"number"
            },
            "emotionSetName":{
               "type":"string"
            }
         }
      },
      "type":"object",
      "properties":{
         "primary":{
            "$ref":"#/definitions/emotionType"
         },
         "secondary":{
            "$ref":"#/definitions/emotionType"
         }
      }
   }
```

11.4.5.3 Semantics

Name	*Definition*
`emotionType`	Specifies the Emotion that the input carries.
`emotionDegree`	Specifies the Degree of Emotion as one of "Low," "Medium," and "High."
`emotionName`	Specifies the ID of an Emotion listed in *Table 55*.
`emotionSetName`	Specifies the name of the Emotion set which contains the Emotion. Emotion set of *Table 55* is used as a baseline, but other sets are possible.

Table 54 gives the standardised three-level Basic Emotion Set partly based on Paul Eckman [17].

Table 54 – Basic Emotion Label Set

EMOTION CATEGORIES	GENERAL ADJECTIVAL	SPECIFIC ADJECTIVAL
ANGER	angry	furious irritated frustrated
CALMNESS	calm	peaceful/serene resigned
DISGUST	disgusted	repulsed
FEAR	fearful/scared	terrified anxious/uneasy
HAPPINESS	happy	joyful content delighted amused
HURT	hurt jealous	insulted/offended resentful/disgruntled bitter
PRIDE/SHAME	proud ashamed	guilty/remorseful/sorry embarrassed
RETROSPECTION	nostalgic	homesick
SADNESS	sad	lonely grief-stricken depressed/gloomy disappointed

Table 55 provides the semantics for each label in the GENERAL ADJECTIVAL and SPECIFIC ADJECTIVAL columns above.

Table 55 – Basic Emotion Semantics Set

ID	Emotion	Meaning
1	amused	positive emotion combined with interest (cognitive state)
2	angry	emotion due to perception of physical or emotional damage or threat

3	anxious/uneasy	low or medium degree of fear, often continuing rather than instant
4	ashamed	emotion due to awareness of violating social or moral norms
5	bitter	persistently angry due to disappointment or perception of hurt or injury
6	calm	relatively lacking emotion
7	content	medium or low degree of happiness, continuing rather than instant
8	delighted	high degree of happiness, often combined with surprise
9	depressed/ gloomy	high degree of sadness, continuing rather than instant, combined with lethargy (see AROUSAL)
10	disappointed	sadness due to failure of desired outcome
11	disgusted	emotion due to urge to avoid, often due to unpleasant perception or disapproval
12	embarrassed	shame due to consciousness of violation of social conventions
13	fearful/scared	emotion due to anticipation of physical or emotional pain or other undesired event or events
14	frustrated	angry due to failure of desired outcome
15	furious	high degree of angry
16	grief-stricken	sadness due to loss of an important social contact
17	happy	positive emotion, often continuing rather than instant
18	homesick	sad due to absence from home
19	hurt	emotion due to perception that others have caused social pain or embarrassment
20	insulted/offended	emotion due to perception that one has been improperly treated socially
21	irritated	low or medium degree of angry
22	jealous	emotion due to perception that others are more fortunate or successful
23	joyful	high degree of happiness, often due to a specific event
24	repulsed	high degree of disgusted
25	lonely	sad due to insufficient social contact
26	mortified	high degree of embarrassment
27	nostalgic	emotion associated with pleasant memories, usually of long before
28	peaceful/serene	calm combined with low degree of happiness
29	proud	emotion due to perception of positive social standing
30	resentful/disgruntled	emotion due to perception that one has been improperly treated
31	resigned	calm due to acceptance of failure of desired outcome, often combined with low degree of sadness
32	sad	negative emotion, often continuing rather than instant, often associated with a specific event
33	terrified	high degree of fear

11.4.6 Social Attitude

11.4.6.1 Definition

A Data Type representing an Entity's internal state related to the way it intends to position itself vis-à-vis the Context, e.g., "Respectful", "Confrontational", "Soothing".

Primary Social Attitude corresponds to General Adjectival and Secondary Social Attitude corresponds to Specific Adjectival in *Table 56.*

11.4.6.2 Syntax

```
{
   "$schema":"http://json-schema.org/draft-07/schema",
   "definitions":{
      "attitudeType":{
         "type":"object",
         "properties":{
            "attitudeDegree":{
              "enum": ["High", "Medium", "Low"]
            },
            "attitudeName":{
               "type":"number"
            },
            "attitudeSetName":{
               "type":"string"
            }
         }
      },
      "type":"object",
      "properties":{
         "primary":{
            "$ref":"#/definitions/attitudeType"
         },
         "secondary":{
            "$ref":"#/definitions/attitudeType"
         }
      }
   }
```

11.4.6.3 Semantics

Name	*Definition*
`attitudeType`	Specifies the Social Attitude that the input carries.
`attitudeDegree`	Specifies the Degree of Social Attitude as one of "Low," "Medium," and "High."
`attitudeName`	Specifies the ID of a Social Attitude listed in *Table 57*.
`attitudeSetName`	Specifies the name of the Social Attitude set which contains the Social Attitude. Social Attitude set of *Table 57* is used as a baseline, but other sets are possible.

Table 56 gives the standardised three-level Basic Social Attitude Set.

Table 56 – Basic Social Attitude Label Set

SOCIAL ATTITUDE CATEGORIES	GENERAL ADJECTIVAL	SPECIFIC ADJECTIVAL
ACCEPTANCE	accepting exclusive/cliquish	welcoming/inviting friendly unfriendly/hostile
AGREEMENT, DISAGREEMENT	like-minded argumentative/disputatious	sarcastic
AGGRESSION	aggressive peaceful submissive	combative/belligerent passive-aggressive mocking
APPROVAL, DISAPPROVAL	admiring/approving disapproving indifferent	awed contemptuous

ACTIVITY, PASSIVITY	assertive passive	controlling permissive/lenient
COOPERATION	cooperative/agreeable uncooperative	flexible subversive/undermining uncommunicative stubborn disagreeable
RESPONSIVENESS	responsive/demonstrative emotional/passionate unresponsive/undemonstrative unemotional/detached	enthusiastic unenthusiastic passionate dispassionate
EMPATHY	empathetic/caring kind uncaring/callous	sympathetic merciful merciless/ruthless self-absorbed selfish/self-serving selfless/altruistic generous
EXPECTATION	optimistic pessimistic	positive sanguine negative/defeatist cynical
EXTROVERSION, INTRO-VERSION	outgoing/extroverted uninhibited/unreserved	sociable approachable
DEPENDENCE	dependent independent	helpless
MOTIVATION	motivated apathetic/indifferent	inspired excited/stimulated discouraged/dejected dismissive
OPENNESS, TRUST	open honest/sincere reasonable trusting	candid/frank closed/distant dishonest/deceitful responsible/trustworthy/de-pendable irresponsible distrustful
PRAISING, CRITICISM	laudatory critical	congratulatory flattering belittling
RESENTMENT, FOR-GIVENESS	forgiving unforgiving/vindictive/spiteful/ vengeful	understanding petty
SELF-PROMOTION	boastful modest/humble/unassuming	
SELF-ESTEEM	conceited/vain self-deprecating/self-effacing	smug

SOCIAL DOMINANCE, CONFIDENCE	arrogant confident submissive	overconfident forward/presumptuous brazen
SEXUALITY	seductive lewd/bawdy/indecent prudish/priggish	suggestive/risqué/naughty
SOCIAL RANK	polite/courteous/respectful rude/disrespectful commanding/domineering pompous/pretentious obedient rebellious/defiant	condescending/patroniz-ing/snobbish pedantic unaffected servile/obsequious

Table 57 provides the semantics for each label in the GENERAL ADJECTIVAL and SPECIFIC ADJECTIVAL columns above.

Table 57 – Basic Social Attitude Semantics Set

ID	Social Attitude	Meaning
1	accepting	attitude communicating willingness to accept into relation-ship or group
2	admiring/approving	attitude due to perception that others' actions or results are valuable
3	aggressive	tending to physically or metaphorically attack
4	apathetic/indifferent	showing lack of interest
5	approachable	sociable and not inspiring inhibition
6	argumentative	tending to argue or dispute
7	arrogant	emotion communicating social dominance
8	assertive	taking active role in social situations
9	awed	approval combined with incomprehension or fear
10	belittling	criticising by understating victim’s achievements, personal attributes, etc.
11	boastful	tending to praise or promote self
12	brazen	high degree of forwardness/presumption
13	candid/frank	open in linguistic communication
14	closed/distant	not open
15	commanding/domineering	tending to assert right to command
16	combative/belligerent	high degree of aggression, often physical
17	communicative	evincing willingness to communicate as needed
18	conceited/vain	evincing undesirable degree of self-esteem
19	condescending/patroniz-ing/snobbish	disrespectfully asserting superior social status, experience, knowledge, or membership
20	confident	attitude due to belief in own ability
21	congratulatory	wishing well related to another’s success or good luck
22	contemptuous	high degree of disapproval and perceived superiority
23	controlling	undesirably assertive
24	cool	repressing outward reaction, often to indicate confidence or dominance, especially when confronting aggression, panic, etc.
25	cooperative/agreeable	communicating willingness to cooperate

26	critical	attitude expressing disapproval
27	cynical	habitually negative, reflecting disappointment or disillusionment
28	dependent	evincing inability to function without aid
29	discouraged/dejected	unmotivated because goals or rewards were not achieved
30	disagreeable	not agreeable
31	disapproving	not approving
32	dishonest/deceitful/insincere	not honest
33	dismissive	actively indicating lack of interest or motivation
34	distrustful	not trusting
35	emotional/passionate	high degree of responsiveness to emotions
36	empathetic/caring	interested in or vicariously feeling others' emotions
37	enthusiastic	high degree of positive response, especially to specific occurrence
38	excited/stimulated	attitude indicating cognitive and emotional arousal
39	exclusive/cliquish	not welcoming into a social group
40	flattering	praising with intent to influence, often insincere
41	flexible	willing to adjust to changing circumstances or needs
42	forward/presumptuous	not observing norms related to intimacy or rank
43	forgiving	tending to forgive improper behaviour
44	friendly	welcoming or inviting social contact
45	generous	tending to give to others, materially or otherwise
46	guilty/remorseful/sorry	regret due to consciousness of hurting or damaging others
47	helpless	high degree of dependence
48	honest/sincere	tending to communicate without deception
49	independent	not dependent
50	indifferent	neither approving nor disapproving
51	inhibited/reserved/introverted/withdrawn	unable or unwilling to participate socially
52	inspired	motivated by some person, event, etc.
53	irresponsible	not responsible
54	kind	tending to act as motivated by empathy or sympathy
55	laudatory	praising
56	lewd/bawdy/indecent	evoking sexual associations in ways beyond social norms
57	like-minded	attitude expressing agreement
58	melodramatic	high or excessive degree of responsiveness or demonstrativeness
59	merciful	tending to avoid punishing others, often motivated by empathy or sympathy
60	merciless/ruthless	not merciful
61	mocking	communicating non-physical aggression, often by imitating a disapproved aspect of the victim
62	modest/humble/unassuming	not boastful
63	motivated	communicating goal-directed emotion and cognitive state
64	negative/defeatist	expressing pessimism, often habitually
65	obedient	evincing tendency to obey commands
66	open	tending to communicate without inhibition

67	optimistic	tending to expect positive events or results
68	outgoing/extroverted/uninhibited/unreserved	not inhibited
69	passive	not assertive
70	passive-aggressive	covertly and non-physically aggressive
71	peaceful	not aggressive
72	pedantic	excessively displaying knowledge or academic status
73	permissive	allowing activity that social norms might restrict
74	pessimistic	tending to expect negative events or results
75	petty	unforgiving concerning small matters
76	polite/courteous/respectful	tending to respect social norms
77	pompous/pretentious	excessively displaying social rank, often above actual status
78	positive	expressing optimism, often habitually
79	prudish/priggish	expressing disapproval of even minor social transgressions, especially related to sex
80	reasonable	evincing willingness to resolve issues through reasoning
81	rebellious/defiant	evincing unwillingness to obey
82	responsible/trustworthy/dependable	evincing characteristics or behaviour that encourage trust
83	responsive/demonstrative	tending to outwardly react to emotions and cognitive states, often as prompted by others
84	rude/disrespectful	not polite or respectful
85	sanguine	low degree of optimism, often expressed calmly
86	sarcastic	communicating disagreement by pretending agreement in an obviously insincere manner
87	seductive	communicating interest in sexual or related contact
88	self-absorbed	not empathetic due to excessive interest in self
89	self-deprecating/self-effacing	tending to criticize, or fail to praise or promote, self
90	selfish/self-serving	not generous due to excessive interest in own benefit
91	selfless/altruistic	tending to act for others' benefit, sometimes exclusively
92	servile/obsequious	excessively and demonstrably obedient
93	shy	low degree of social inhibition
94	smug	evincing undesirable degree of self-esteem related to perceived triumph
95	stubborn	unwilling to change one's mind or behaviour
96	sociable	comfortable in social situations
97	submissive	tending to submit to social dominance
98	subversive/undermining	communicating intention to work against a victim's goals
99	suggestive/risqué/naughty	evoking sexual associations within social norms
100	supportive	communicating willingness to support as needed
101	sympathetic	empathetic related to others' hurt or suffering
102	trusting	tending to trust others
103	unaffected	not pompous
104	uncaring/callous	not empathetic or caring
105	uncommunicative	not communicative
106	uncooperative	not cooperative
107	understanding	forgiving due to ability to understand motivations

108	unemotional/dispassion-ate/detached	not emotional, even when emotion is expected
109	unenthusiastic	not enthusiastic
110	unfriendly/hostile	not friendly
111	unresponsive/undemon-strative	not responsive or demonstrative
112	welcoming/inviting	high degree of acceptance with emotional warmth

11.5 Miscellanea

11.5.1 Selector

Selector is a multi-variable composed of:

1. Input type:
   0: Speech is used as input and in the subsequent processing.
   1: Text is used in lieu of Speech.
2. Language Preference: expressed as Language.
3. Target Translated Language: expressed as Language.
4. Input Modality to Personal Status Extraction expressed as TSFB where the four variable T,S,F,B are Boolean:
   T=0: Text Object
   T=1: Text Descriptors
   S=0: Speech Object
   S=1: Speech Descriptors
   F=0: Face Object
   F=1: Face Descriptors
   B=0: Body Object
   B=1: Body Descriptors

11.5.2 Instance Identifier

11.5.2.1 Definition

The label of an element of a set including, e.g., objects, humans, etc. – belonging to some levels in a hierarchical classification (taxonomy).

11.5.2.2 Syntax

```
{
  "$schema":"http://json-schema.org/draft-07/schema",
  "title":"InstanceIdentifier",
  "type":"object",
  "properties":{
    "InstanceLabel":{
      "type":"string"
    },
    "LabelConfidenceLevel":{
      "type":"number",
      "minimum":0,
      "maximum":1
    },
    "Classification":{
      "type":"array",
      "items":{
        "type":"string"
      }
    },
    "ClassificationConfidenceLevel":{
      "type":"number",
      "minimum":0,
```

```
      "maximum":1
    }
  },
  "required":[
    "InstanceLabel",
    "LabelConfidenceLevel",
    "Classification",
    "ClassificationConfidenceLevel"
  ]
}
```

11.5.2.3 Semantics

Name	*Definition*
`InstanceIdentifier`	Provides the identifier of the Instance.
`InstanceLabel`	Describes the Instance identified by InstanceIdentifier.
`LabelConfidenceLevel`	Indicates the confidence level of the association between InstanceLabel and the Instance.
`Classification`	Describes the taxonomy inferred for the Instance.
`ClassificationConfidenceLevel`	Indicates the confidence level of the association between Classification and the Instance.

11.5.3 Language

Language preference is expressed by two characters as specified by [8].

11.5.4 Meaning

11.5.4.1 Definition

A Data Type representing an input text such as syntactic and semantic information. It results from natural language (text analysis) and consists of the following elements:

- POS_tagging
- NE_tagging
- Dependency_tagging
- SRL_tagging

Meaning is a synonym of Text Descriptors.

11.5.4.2 Syntax

```
{
   "$schema":"http://json-schema.org/draft-07/schema",
   "definitions":{
      "meaning":{
         "type":"object",
         "properties":{
            "POS_tagging":{
               "POS_tagging_set":{
                  "type":"string"
               },
               " POS_tagging_result":{
                  "type":"string"
               }
            },
            "NE_tagging":{
               "NE_tagging_set":{
                  "type":"string"
```

```
            },
            " NE_tagging_result":{
               "type":"string"
            }
         },
         "dependency_tagging":{
            "dependency_tagging_set":{
               "type":"string"
            },
            "dependency_tagging_result":{
               "type":"string"
            }
         },
         "SRL_tagging":{
            " SRL_tagging_set":{
               "type":"string"
            },
            " SRL_tagging_result":{
               "type":"string"
            }
         }
      }
   },
   "type":"object",
   "properties":{
      "primary":{
         "$ref":"#/definitions/meaning"
      },
      "secondary":{
         "$ref":"#/definitions/meaning"
      }
   }
  }
}
```

11.5.4.3 Semantics

Name	*Definition*
`Meaning`	Provides an abstract of description of natural language analysis results.
`POS_tagging`	Indicates POS tagging results, including information on the POS tagging set and tagged results of the User question. POS: Part of Speech, such as noun, verb, etc.
`NE_tagging`	Indicates NE tagging results, including information on the NE tagging set and tagged results of the User question. NE: Named Entity such as Person, Organisation, Fruit, etc.
`dependency_tagging`	Indicates dependency tagging results, including information on the dependency tagging set and tagged results of the User question. Dependency indicates the structure of the sentence, such as subject, object, head of the relation, etc.
`SRL_tagging`	Indicates SRL (Semantic Role Labelling) tagging results including information on the SRL tagging set and tagged results of the User question. SRL indicates the semantic structure of the sentence, such as agent, location, patient role, etc.

11.5.5 Portable Avatar

11.5.5.1 Definition

A Data Type conveying information about an Avatar such as Avatar ID, Avatar Model, Body Descriptors, Face Descriptors, Speech Data, and Text, and Context information such as Time, Audio-Visual Scene, Spatial Attitude, Language Preference, and Personal Status.

11.5.5.2 Syntax

```
{
  "$id": "https://schemas.mpai.community/PAF/V1.0/PortableAvatarFormat.json",
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "PortableAvatarFormat",
  "type": "object",
  "properties": {
    "ID": {
      "type": "string"
    },
    "Timestamp": {
      "type": "object",
      "properties": {
        "Type": {
          "type": "string"
        },
        "Value": {
          "type": "string",
          "oneOf": [
            {
              "format": "date-time"
            },
            {
              "const": "0"
            }
          ]
        }
      },
      "required": [
        "Value"
      ],
      "if": {
        "properties": {
          "Value": {
            "const": "0"
          }
        }
      },
      "then": {
        "properties": {
          "Type": {
            "type": "null"
          }
        }
      },
      "else": {
        "required": [
          "Type"
        ]
      }
    },
    "Placement": {
      "type": "object",
      "properties": {
        "Spatial Attitude": {
          "type": "object",
          "properties": {
            "Position": {
              "type": "array",
              "contains": {
                "type": "number"
              },
              "minContains": 3,
              "maxContains": 3
            },
            "Orientation": {
              "type": "array",
              "contains": {
                "type": "number"
              },
```

```
              "minContains": 3,
              "maxContains": 3
            }
          }
        },
        "Visual Environment": {
          "$ref": "https://schemas.mpai.community/PAF/V1.0/data/VisualEnvironment.json"
        }
      }
    },
    "Visual": {
      "type": "object",
      "properties": {
        "Model": {
          "$ref": "https://schemas.mpai.community/PAF/V1.0/data/AvatarModel.json"
        },
        "BodyDescriptors": {
          "$ref": "https://schemas.mpai.community/PAF/V1.0/data/BodyDescriptors.json"
        },
        "FaceDescriptors": {
          "$ref": "https://schemas.mpai.community/PAF/V1.0/data/FaceDescriptors.json"
        },
      }
    }
  },
  "Audio": {
    "type": "object",
    "properties": {
      "LanguagePreference": {
        "type": "string",
        "minLength": 2,
        "maxLength": 2
      },
      "Speech": {
        "type": "object",
        "properties": {
          "Encoding": {
            "enum": [
              "MP3",
              "AAC"
            ]
          },
          "Utterance": {
            "type": "array",
            "contains": {
              "type": "integer"
            }
          }
        }
      }
    }
  },
  "Text": {
    "type": "string"
  },
  "PersonalStatus": {
    "$ref": "https://schemas.mpai.community/MMC/V2.0/data/PersonalStatus.json"
  }
}
```

11.5.5.3 Semantics

Note: All elements in Table 58 are optional.

Table 58 – Variables composing the Portable Avatar Format

Variable name	Comments
ID	String
Time	
Type	0=Relative

	1=Absolute
Value	Seconds from - 0000/00/00T00:00 (relative time) - 1970/01/01T00:00 (absolute time)
Visual	
AVSceneDescriptors	Describes the Audio-Visual Scene where the Avatar Model will be placed with the Spatial Attitude.
SpatialAttitude	Position, Orientation, and time derivatives up to the second order.
Avatar Model	Specified by MPAI-PAF
BodyDescriptors	Specified by MPAI-PAF
FaceDescriptors	Specified by MPAI-PAF
Language	
LanguagePreference	Specified by MPAI-MMC V2
Speech	
SpeechType	String identifies compression
Speech	Byte stream conveying speech
Text	Specified by MPAI-MMC V2
PersonalStatus	Specified by MPAI-MMC V2

Annex 1 - MPAI Basics (Informative)

In recent years, Artificial Intelligence (AI) and related technologies have been introduced in a broad range of applications affecting the life of millions of people. These technologies are expected to do even more in the future. Digital media standards have positively influenced industry and billions of people, and AI-based data coding standards are expected to have a similar positive impact. However, certain AI technologies may carry inherent risks, e.g., by introducing bias toward some classes of users or application domains. Thus the need for standardisation becomes more important and urgent than ever.

These considerations have prompted the establishment of the international, unaffiliated, not-for-profit Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) organisation. The group's mission is to develop *AI-enabled data coding standards* to enable the development of AI-based products, applications, and services.

Technical Specification: Governance of the MPAI Ecosystem (MPAI-GME) [1] provides the technical foundations of the MPAI Ecosystem composed of:

1. *MPAI* developing and maintaining:
    a. Technical Specification.
    b. Reference Software Specification.
    c. Conformance Testing Specification.
    d. Performance Assessment Specification.
    e. Technical Report
2. *Implementers* developing implementations of MPAI Technical Specifications.
3. *Performance Assessors*, appointed by MPAI but independent of it, verifying various aspects of an implementation, such as Reliability, Robustness, Fairness and Replicability.
4. *The MPAI Store*, collecting and testing implementations for Conformance and making available to End Users those that pass.
5. *End Users* downloading implementations from the MPAI Store.

Figure 49 depicts the MPAI ecosystem operation for conforming MPAI implementations.

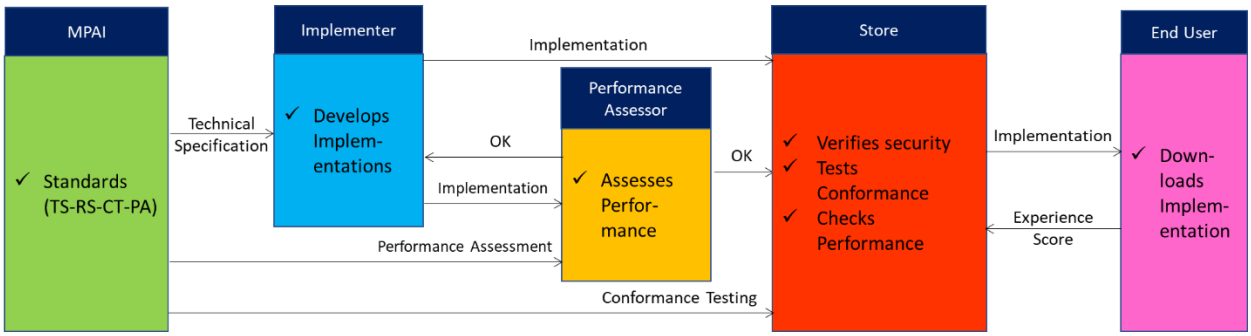

Figure 49 – The MPAI ecosystem operation

MPAI Technical Specifications are developed in compliance with a rigorous process [14] in service of the following policies:

1. While closely accommodating a given AI use case, so far as possible, remain agnostic to the technology – AI or DP – used in an implementation.
2. Facilitate the exploitation of a Technical Specification once adopted by MPAI.
3. Attempt to attract various industries, end users, and regulators.

4. Address three levels of standardisation, any of which an implementer can freely decide to adopt: the data exchanged by AIMs ("Data Types"), AIMs, and AIWs.
5. Specify the data exchanged by AIMs with clear, humanly understandable semantics, so far as possible.

Technical Specification: AI Framework (MPAI-AIF) V2, depicted in *Figure 50,* enables dynamic configuration, initialisation, and control of AIWs in a standard environment called AI Framework (AIF) [2].

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

Thus, users can exercise AIWs that are both proprietary or standardised by MPAI – i.e., with standard functions and interfaces, and with an explicit computing workflow. Developers can compete in providing AIMs with standard functions and interfaces that may have improved performance compared to other implementations. AIMs can execute data processing or Artificial Intelligence algorithms and can be implemented in hardware, software, or hybrid hardware/software.

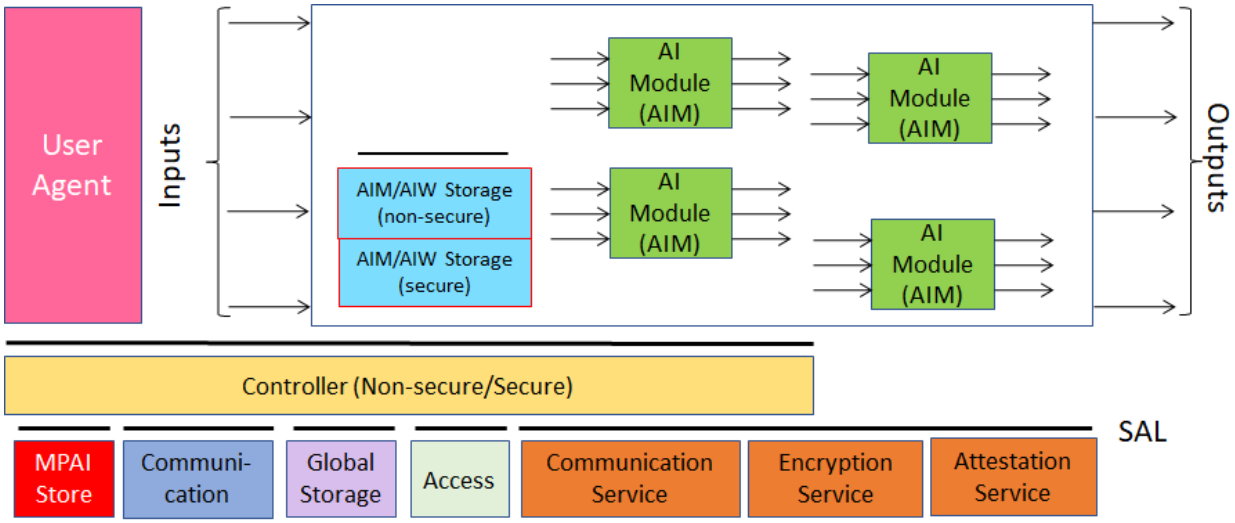

Figure 50 - The AI Framework (MPAI-AIF) V2 Reference Model

An AIW is defined by its Function and input/output Data and by its AIM topology. Likewise, an AIM is defined by its Function and input/output Data. MPAI standards do not restrict the technology used to implement the AIM, which may be based on AI or data processing, and implemented in software, hardware, or hybrid software and hardware technologies.

An AIW and its AIMs may be rated at one of three interoperability levels:
Level 1 – Proprietary 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.

Note the following points:
1. AIMs can implement basic or aggregated functionalities. In the former case, an AIM is called Basic, and in the latter, Composite. A Composite AIM may include other Composite AIMs.

This Technical Specification includes several examples of Composite AIMs that include other Composite AIMs, which in turn include Basic AIMs.

2. The distinction between Basic and Component AIMs is implementation specific. An implementation might not expose the internal AIMs of a Composite AIM, or it might implement a Basic AIM with interconnected AIMs.
3. AIMs may include functionalities related to the decoding of the payload embedded in the watermarking that an upstream AIM may have inserted [19].
4. AIMs are specified by the following:

4.1. The functions they perform.
4.2. The Reference Model.
4.3. The Input/Output Data.
4.4. The component AIMs (called SubAIMs).

The MPAI-HMC specification will consistently use this structure throughout the Specification.

Annex 2 - Notices and Disclaimers Concerning MPAI Standards (Informative)

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

In the following, "Standard" means the collection of four MPAI-approved and published 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. 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. 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 to determine the status of any given published MPAI Standard.

Description 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. 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 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 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.

Annex 3 - General MPAI Terminology

The capitalised Terms used in this standard that are not already included in Table 1 are defined in Table 59.

NOTE: A hyphenated entry for e.g., "- Testing" should be read as adding that word to the closest non-hyphenated entry above it – in this case, "Conformance," giving "Conformance Testing" as the complete entry name.

Table 59 - 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 Model (AIM)	A data 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 outputs according to the AIW Function.
Application Standard	An MPAI Standard designed to enable a particular application domain.
Assessment Laboratory	A laboratory accredited to Assess the Grade of Performance of Implementations.
Channel	A connection between an output port of an AIM and an input port of an AIM. The term "connection" is also used as synonymous.
Communication	The infrastructure that implements message passing between AIMs.
Component	One of the 7 AIF elements: Access, Communication, Controller, Internal Storage, Global Storage, Store, and User Agent
Composite AIM	An AIM aggregating more than one AIM.
Component	One of the 7 AIF elements: Access, Communication, Controller, Internal Storage, Global Storage, Store, and User Agent
Conformance	The attribute of an Implementation of being a correct technical Implementation of a Technical Specification.
- Testing	The normative document specifying the Means to Test the Conformance of an Implementation.
- Testing Dataset	A dataset used to Test the Conformance of an implementation to a Technical Specification.
- Testing Means	Procedures, tools, data sets and/or data set characteristics to Test the Conformance of an Implementation.

- Testing Procedure	The sequence of steps to be performed to Test the Conformance of an implementation.
- Testing Tools	Devices and/or software used 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.
- Format	The standard digital representation of Data.
- Type	An instance of Data with a specific Data Format.
- Semantics	The meaning of Data.
Descriptor	Coded representation of a text, audio, speech, or visual feature.
Digital Representation	Data corresponding to and representing a physical entity.
Ecosystem	The ensemble of actors making it possible for a User to execute an application composed of an AIF, one or more AIWs, each with one or more AIMs potentially sourced from independent implementers.
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.
AIM/AIW Storage	A Component to store data of the individual 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) conforming with a Use Case of an MPAI Application Standard.
Implementer	A legal entity implementing MPAI Technical Specifications.
ImplementerID (IID)	A unique name assigned by the ImplementerID Registration Authority to an Implementer.
ImplementerID Registration Authority (IIDRA)	The entity appointed by MPAI to assign ImplementerID's to Implementers.
Instance ID	Instance of a class of Objects and the Group of Objects the Instance belongs to.
Interoperability	The ability to functionally replace an AIM with another AIW having the same Interoperability Level
- Level	The attribute of an AIW and its AIMs to be executable in an AIF Implementation and to:

	1. Be proprietary (Level 1) 2. Pass the Conformance Testing (Level 2) of an Application Standard 3. Pass the Performance Testing (Level 3) of an Application Standard.
Knowledge Base	Structured and/or unstructured information made accessible to AIMs via MPAI-specified interfaces
Message	A sequence of Records transported by Communication through Channels.
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.
- Assessment	The normative document specifying the Means to Assess the Grade of Performance of an Implementation.
- Assessment Means	Procedures, tools, data sets and/or data set characteristics to Assess the Performance of an Implementation.
- Assessor	An entity Assessing the Performance of an Implementation.
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	A data structure 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.
Standard	A set of Technical Specification, Reference Software, Conformance Testing, Performance Assessment, and Technical Report 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.
Time Base	The protocol specifying how 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 4 - Patent Declarations

Technical Specification: Human and Machine Communication (MPAI-HMC) has been developed according to the process outlined in the MPAI Statutes [13] and the MPAI Patent Policy [14] using elements already developed in other MPAI Technical Specifications with the addition of a few more new elements.

The following table will include references to the entities declaring to agree to licence their standard essential patents reading on *Technical Specification: Human and Machine Communication (MPAI-HMC)* according to the MPAI-HMC Framework Licence [15]:

Entity	Name	email address