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

MPAI Technical Specification

Multimodal Conversation
MPAI-MMC

V2 .1

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 3 - Notices and Disclaimers.

Technical Specification
Multimodal Conversation (MPAI-MMC) V2.1

1 Introduction (Informative)

From the moment a human built the first machine, there was a need to "communicate" with it. As more complex machines were built, the need for more sophisticated communication methods arose. Today, as personal devices become more pervasive, and the use of information and other online services become ubiquitous, human-machine communication often becomes more direct and even "personal". In the past, humans communicated with more primitive machines by touch, later by characters and then with speech and even visual means.

The ability of Artificial Intelligence to learn from interactions with humans gives machines the ability to improve their "conversational" capabilities by better understanding the meaning of what a human types or says and by providing more pertinent responses. If properly trained, machines can also learn to understand additional or hidden meanings of a sentence by analysing a human's text, speech, or gestures. Machines can also be made to develop and rely on "internal statuses" comparable to those driving the attitudes of conversing humans. Thus, they can provide responses – in text, speech, and gestures – that are more human-like and richer in content.

Technical Specification: Multimodal Conversation (MPAI-MMC) V2 has been developed by MPAI – Moving Picture, Audio, and Data Coding by Artificial Intelligence, the international, unaffiliated, non-profit organisation developing standards for Artificial Intelligence (AI)-based data coding with clear Intellectual Property Rights licensing frameworks in compliance with the rigorous MPAI Process [16] in pursuit of the following policies:

1. Be friendly to the AI context but, to the extent possible, agnostic to the technology – AI or Data Processing – used in an implementation.
2. Be attractive to different industries, end users, and regulators.
3. Address three levels of standardisation any of which an implementer can freely decide to adopt:
    a. Data types, i.e., the data exchanged by systems.
    b. Components (called AI Modules - AIM).
    c. Connections of components (called AI Workflows - AIW).
4. Specify the data exchanged by components with a semantic that is clear to the extent possible.

Technical Specification: AI Framework (MPAI-AIF) V2 [2] enables dynamic configuration, initialisation, and control of AIWs in a standard environment called AI Framework (AIF). Figure 1 depicts the AI Framework.

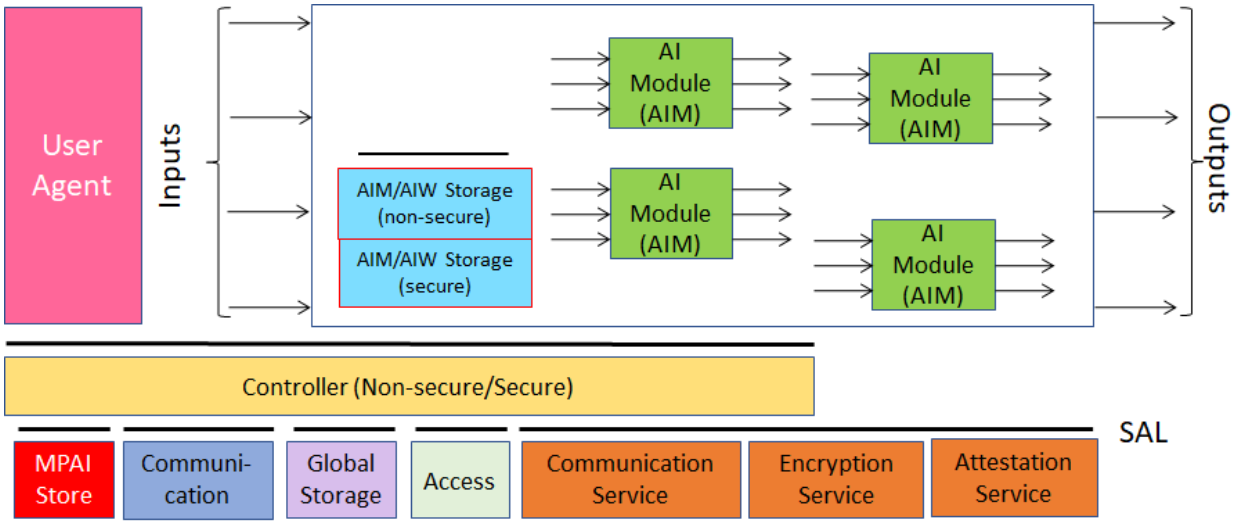

Figure 1 - The AI Framework (MPAI-AIF) V2 Reference Model

AIWs and AIMs have standard interfaces. AIMs can execute data processing or Artificial Intelligence algorithms and can be implemented in hardware, software, or hybrid hardware/software. AI Module can be Composite if they include connected AI Modules.

The MPAI-AIF-specified AIF environment enables the secure execution of AIWs constituted by AIMs. Thus, users can have machines implementing AIMs whose internal operation they understand to some degree, rather than machines that are just "black boxes" resulting from unknown training with unknown data. AIM developers can provide components with standard interfaces that can have improved performance compared to other implementations.

An AIW and its AIMs may have 3 interoperability levels any of which implementers can freely adopt:

Level 1 – Implementer-specific and satisfying the MPAI-AIF Standard.
Level 2 – Specified by an MPAI Application Standard.
Level 3 – Specified by an MPAI Application Standard and certified by a Performance Assessor.

As manager of the MPAI Ecosystem specified by Governance of MPAI Ecosystem (MPAI-GME) [1], MPAI ensures that a user can:

1. Operate a reference implementation of the Technical Specification, by providing a Reference Software Specification with annexed software.
2. Test the conformance of an implementation with the Technical Specification, by providing Conformance Testing Specification.
3. Assess the performance of an implementation of a Technical Specification, by providing the Performance Assessment Specification.
4. Get conforming implementations possibly with a performance assessment report from a trusted source through the MPAI Store.

The MPAI-MMC V2 Technical Specification will be accompanied by the Reference Software, Conformance Testing, and Performance Assessment Specifications. Conformance Testing specifies methods enabling users to ascertain whether a data type generated by an AIM, an AIM, or an AIW conform with this Technical Specification.

The MPAI-MMC V2 Technical Specification provides the technologies supporting the implementation of a subset or the totality of the possibilities envisaged by this Introduction:

1. It is organised in Use Cases collected in Chapter 5, such as Conversation with Personal Status, Multimodal Question Answering, and Unidirectional Speech Translation, corresponding to AI Workflows.
2. Each Use Case provides:
    a. The functions.
    b. The Input/Output Data of the AIW implementing it.
    c. The Reference Model specifying the AIM topology.
    d. The AIMs specified in terms of functions performed and Input/Output Data.
3. A single chapter (Chapter 7) collects all data formats referenced in the specification.
4. Annexes provide the JSON metadata of the AIWs, Composite AIM, and AIMs.

In this Introduction and in the following Chapters, Terms beginning with a capital letter are defined in *Table 1* if they are specific to this Technical Specification and in *Table 58* if they are common to all MPAI Technical Specifications. The chapters and the Annexes are Normative unless they are labelled as Informative.

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

2 Scope of Standard

Multimodal Conversation (MPAI-MMC) specifies:

1. Data Formats for analysis of text, speech, and other non-verbal components as used in human-machine and machine-machine conversation applications.
2. Use Cases implemented in the AI Framework using Data Formats from MPAI-MMC and other MPAI standards and providing recognised applications in the Multimodal Conversation domain.

This Technical Specification includes the following Use Cases:

1. "*Conversation with Personal Status*" (CPS), enabling conversation and question answering with a machine able to extract the inner state of the entity it is conversing with and showing

itself as a speaking digital human able to express a Personal Status. By adding or removing minor components to this general Use Case, five Use Cases are spawned:

2. "*Conversation About a Scene*" (CAS) where a human converses with a machine pointing at the objects scattered in a room and displaying Personal Status in their speech, face, and gestures while the machine responds displaying its Personal Status in speech, face, and gesture.
3. *"Virtual Meeting Secretary"* (VSV) where an avatar not representing a human in a virtual avatar-based video conference extracts Personal Status from Text, Speech, Face, and Gestures, displays a summary of what other avatars say, and receives and act on comments.
4. *"Human-Connected Autonomous Vehicle Interaction"* (HCI) where humans converse with a machine displaying Personal Status after having been properly identified by the machine with their speech and face in outdoor and indoor conditions while the machine responds displaying its Personal Status in speech, face, and gesture.
5. "*Conversation with Emotion*" (CWE), supporting audio-visual conversation with a machine impersonated by a synthetic voice and an animated face.
6. "*Multimodal Question Answering*" (MQA), supporting request for information about a displayed object.
7. Three Uses Cases supporting text and speech translation applications. In each Use Case, users can specify whether speech or text is used as input and, if it is speech, whether their speech features are preserved in the interpreted speech:

7.1. "*Unidirectional Speech Translation*" (UST).
7.2. "*Bidirectional Speech Translation*" (BST).
7.3. "*One-to-Many Speech Translation*" (MST).

8. The *"Personal Status Extraction"* Composite AIM that estimates the Personal Status conveyed by Text, Speech, Face, and Gesture – of an Entity, i.e., a real or digital human.

Note that:

1. Each Use Case normatively defines:

1.1. The Functions of the AIW implementing it and of the AIMs.
1.2. The Connections between and among the AIMs
1.3. The Semantics and the Formats of the input and output data of the AIW and the AIMs.

2. Each Composite AIM normatively defines:

2.1. The Functions of the Composite AIM implementing it and of the AIMs.
2.2. The Connections between and among the AIMs
2.3. The Semantics and the Formats of the input and output data of the AIW and the AIMs.

The word *normatively* implies that an Implementation claiming Conformance to:

1. An *AIW*, shall:
   a. Perform the AIW function specified in the appropriate Section of Chapter 5.
   b. All AIMs, their topology and connections should conform with the AIW Architecture specified in the appropriate Section of Chapter 5.
   c. The AIW and AIM input and output data should have the formats specified in the appropriate Sections of Chapter 7.
2. An *AIM*, shall:
   a. Perform the functions specified by the appropriate Section of Chapter 5 or 6.
   b. Receive and produce the data specified in the appropriate Section of Chapter 7.
3. A data *Format*, the data shall have the format specified in Chapter 7.

Users of this Technical Specification should note that:

1. This Technical Specification defines Interoperability Levels but does not mandate any.
2. Implementers decide the Interoperability Level their Implementation satisfies.

3. Implementers can use the Reference Software of this Technical Specification to develop their Implementations.
4. The Conformance Testing specification can be used to test the conformity of an Implementation to this Standard.
5. Performance Assessors can assess the level of Performance of an Implementation based on the Performance Assessment specification of this Standard.
6. Implementers and Users should consider Annex 2 - Notices and Disclaimers.

The current Version of MPAI-MMC has been developed by the MPAI Multimodal Conversation Development Committee (MM-DC). MPAI expects to produce future MPAI-MMC Versions extending the scope of the Use Cases and/or add new Use Cases within the Multimodal Conversation scope.

3 Terms and Definitions

Terms beginning with a capital letter have the meaning defined in *Table 1*. 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 *Table 1* 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 – Table of terms and definitions

Term	Definition
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 Audio and Visual Object in a 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.
- Object	Coded representation of Audio information with its metadata. An Audio Object can be a combination of Audio Objects.
- Scene	The Audio Objects of an Environment with Object location metadata.
Audio-Visual Object	Coded representation of Audio-Visual information with its metadata. An Audio-Visual Object can be a combination of Audio-Visual Objects.
Audio-Visual Scene	(AV Scene) The Audio-Visual Objects of an Environment with Object location metadata.
Avatar	An animated 3D object representing a real or fictitious person in a Virtual Space.
- Model	An inanimate avatar exposing interfaces enabling animation.
Cognitive State	The coded representation of the internal state reflecting the way a human or avatar understands the Environment, such as "Confused", "Dubious", "Convinced".

Colour (of speech)	The timber of an identifiable voice independent of a current Personal Status and language.
Connected Autonomous Vehicle	A vehicle able to autonomously reach an assigned geographical position by: 1. Understanding human utterances. 2. Planning a route. 3. Sensing and interpreting the Environment. 4. Exchanging information with other CAV. 5. Acting on the CAV's motion actuation subsystem.
Context	Additional information about a communication emitted by an Entity, such as language, culture etc..
Data	Information in digital form.
- Format	The standard digital representation of Data.
- Type	An instance of Data with a specific Data Format.
Descriptor	Coded representation of text, audio, speech, or visual feature.
Digital Representation	Data corresponding to and representing a real 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 real or Digital Human
Environment	A Virtual Space containing a Scene.
Face	The portion of a 2D or 3D digital representation corresponding to the face of a human.
Factor	One of Emotion, Cognitive State and Attitude.
Gesture	A movement of the body or part of it, such as the head, arm, hand, and finger, often a complement to a vocal utterance.
Grade	The intensity of a Factor.
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 avatar or an object.
Instance	An element of a set of entities – Objects, users etc. – belonging to some levels in a hierarchical classification (taxonomy).
Intention	The result of analysis of the goal of an input question.
Manifestation	The manner of showing the Personal Status, or a subset of it, in any one of Speech, Face, and Gesture.
Meaning	Information extracted from Text such as syntactic and semantic information, Personal Status, and other information, such as an Object Identifier.
Modality	One of Text, Speech, Face, or Gesture.
Object Descriptors	Attribute of the coded representation of an object in a Scene, including its Spatial Attitude.
Orientation	The set of the 3 roll, pitch, yaw angles indicating the rotation around the principal axis (x) of an Object, its y axis having an angle of 90° counterclockwise (right-to-left) with the x axis and its z axis pointing up toward the viewer.
Personal Status	The ensemble of information internal to a person, including Emotion,

	Cognitive State, and Attitude.
Portable Avatar	A Data Type representing an Avatar and its Context.
Pitch	The fundamental frequency of Speech. Pitch is the attribute that makes it possible to judge sounds as "higher" and "lower."
Point of View	The Spatial Attitude of a human or avatar looking at an Environment.
Position	The 3 coordinates (x,y,z) of a representative point of an object in the Real and Virtual Space.
Refined Text	The Text resulting from the analysis of the Text produced by Automatic Speech Recognition made by Natural Language Understanding.
Scene	A structured composition of Objects.
Speech	Digital representation of analogue speech sampled at a frequency between 8 kHz and 96 kHz with a number of bits/sample of 8, 16 and 24, and non-linear and linear quantisation.
- Features	Aspects of a speech segment that enable its description and reproduction, e.g., degree of vocal tension, Pitch, etc., and that can be automatically recognised and extracted for speech synthesis or other related purposes.
- Rate	The number of Speech Units per second.
- Unit	Phoneme, syllable, or word as a segment of Speech.
Summary	An abridged outline of the content of the utterance(s) of one or more Users possibly including their Personal Statuses.
Text	A sequence of characters drawn from a finite alphabet.
Visual Object	Coded representation of Visual information with its metadata. A Video Object can be a combination of Video Objects.
Vocal Gesture	Utterance, such as cough, laugh, hesitation, etc. Lexical elements are excluded.

4 References

4.1 Normative References

This standard normatively references the following documents, both from MPAI and other standards organisations. MPAI standards are publicly available at https://mpai.community/standards/resources/.

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: Portable Avatar Format (MPAI-PAF) V1; https://mpai.community/standards/mpai-paf/.
4. Technical Specification: Context-based Audio Enhancement (MPAI-CAE) V2; https://mpai.community/standards/mpai-cae/.
5. Technical Specification: Connected Autonomous Vehicle (MPAI-CAV) – Architecture V1; https://mpai.community/standards/mpai-cav/.
6. Technical Specification: Visual Object and Scene Description (MPAI-OSD) V1; https://mpai.community/standards/mpai-osd/.
7. Khronos; Graphics Language Transmission Format (glTF); October 2021; https://registry.khronos.org/glTF/specs/2.0/glTF-2.0.html
8. ISO 639; Codes for the Representation of Names of Languages – Part 1: Alpha-2 Code.
9. ISO/IEC 10646; Information technology – Universal Coded Character Set.

10. 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.
11. ISO/IEC 14496-10; Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding.
12. ISO/IEC 14496-12; Information technology – Coding of audio-visual objects – Part 12: ISO base media file format.
13. ISO/IEC 23008-2; Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 2: High Efficiency Video Coding.
14. ISO/IEC 23094-1; Information technology – General video coding – Part 1: Essential Video Coding.
15. MPAI; The MPAI Statutes; https://mpai.community/statutes/.
16. MPAI; The MPAI Patent Policy; https://mpai.community/about/the-mpai-patent-policy/.
17. MPAI; Framework Licence of the Multimodal Conversation Technical Specification (MPAI-MMC) V1; https://mpai.community/standards/mpai-mmc/framework-licence/mpai-mmc-v1-framework-licence/.
18. MPAI; Framework Licence of the Multimodal Conversation Technical Specification (MPAI-MMC) V2; https://mpai.community/standards/mpai-mmc/call-for-technologies/mpai-mmc-v2-call-for-technologies/.

4.2 Informative References

The references provided here are for information purpose.

19. Ekman, Paul (1999), "Basic Emotions", in Dalgleish, T; Power, M (eds.), Handbook of Cognition and Emotion (PDF), Sussex, UK: John Wiley & Sons.
20. Emotion Markup Language (EmotionML) 1.0; https://www.w3.org/TR/2010/WD-emotionml-20100729/diffmarked.html.
21. Hobbs J.R., Gordon A.S. (2011) The Deep Lexical Semantics of Emotions. In: Ahmad K. (eds) Affective Computing and Sentiment Analysis. Text, Speech, and Language Technology, vol 45. Springer, Dordrecht, https://people.ict.usc.edu/~gordon/publications/EMOT08.PDF and https://www.researchgate.net/publication/227251103_The_Deep_Lexical_Semantics_of_Emotions.

5 Use Cases

5.1 General

Interoperable implementations of the MPAI-MMC V2 Use Cases require standardisation of a set of Data Types specified by MPAI-MMC V2 and by other MPAI Technical Specifications. Chapter 7 specifies the Formats of all the Data Types in Table 2.

Table 2 - Data Types of Multimodal Conversation (MPAI-MMC) V2

Section	Data Type	Technical Specification
7.1	*Media*	
7.1.1	Audio File	MPAI-MMC
7.1.2	Text	MPAI-MMC
7.1.3	Video	MPAI-MMC
7.1.4	Video File	MPAI-MMC
7.2	*Descriptors*	
7.2.1	Audio Scene Descriptors	MPAI-CAE

7.2.2	Face Descriptors	MPAI-PAF
7.2.3	Gesture Descriptors	MPAI-PAF
7.2.4	Speech Descriptors	MPAI-MMC
7.2.5	Speech Features	MPAI-MMC
7.2.6	Text Descriptors	MPAI-MMC
7.2.7	Visual Scene Descriptors	MPAI-MMC
7.3	*Personal Status*	
7.3.1	Factors and Modalities	MPAI-MMC
7.3.2	Personal Status Data	MPAI-MMC
7.3.3	Cognitive State	MPAI-MMC
7.3.4	Emotion	MPAI-MMC
7.3.5	Social Attitude	MPAI-MMC
7.4	*Objects and Scenes*	
7.4.1	Spatial Attitude and Point of View	MPAI-CAE
7.4.2	Audio Object and Scene	MPAI-OSD
7.4.3	Visual Object and Scene	MPAI-OSD
7.5	*Miscellanea*	
7.5.1	Instance Identifier	MPAI-MMC
7.5.2	Intention	MPAI-MMC
7.5.3	Language Identifier	MPAI-MMC
7.5.4	Meaning	MPAI-MMC
7.5.5	Spatial Attitude	MPAI-MMC
7.5.6	Query Format of Video of Faces KB	MPAI-MMC

Each Use Case is implemented as an AI Workflow (AIW) composed of AI Modules (AIMs) and includes the following elements:

1	Functions of the AIW	The functions performed by the AIW implementing the MPAI-MMC Use Case.
2	Reference Model of the AIW	The Topology of AIMs in the AIW.
3	Input and Output Data of the AIW	Input and Output Data of the AIW.
4	Functions of the AIMs	Functions performed by all AIMs of the AIW.
5	Input and Output Data of the AIMs	Input and Output Data of all AIMs of the AIW.
6	Specification of AIMs and JSON Metadata	Links to summary specification on the web of each AIM used in the standard and the corresponding JSON Metadata as specified by MPAI-AIF

5.2 Conversation with Personal Status (MMC-CPS)

5.2.1 Scope of Conversation with Personal Status

When humans have a conversation with other humans, they use speech and, in constrained cases, text. Their interlocutors perceive speech and/or text supplemented by visual information related to the speaker's face and gesture of a conversing human. Text, speech, face, and gesture may convey information about the internal state of the speaker that MPAI calls Personal Status. Therefore, handling of Personal Status information in a human-machine conversation and, in the future, even

machine-machine conversation, is a key feature of a machine trying to understand what the speakers' utterances mean because Personal Status recognition can improve understanding of the speaker's utterance and help a machine produce better replies.
Conversation with Personal Status (MMC-CPS) is a general Use Case of an entity – a real or digital human – conversing and question answering with a machine. The machine captures and understands Speech, extracts Personal Status from the Text, Speech, Face, and Gesture Factors, fuses the Factors into an estimated Personal Status of the entity to achieve a better understanding of the context in which the entity utters Speech.

5.2.2 Reference Model of Conversation with Personal Status

Figure 2 gives the Conversation with Personal Status Reference Model including the input/output data, the AIMs, and the data exchanged between and among the AIMs.

The operation of the Conversation with Personal Status Use Case develops as follows:

1. Input Selector is used to inform the machine whether the human employs Text or Speech in conversation with the machine.
2. Visual Scene Description extracts the Scene Geometry, the Visual Objects and the Face and Body Descriptors of humans in the Scene.
3. Audio Scene Description extracts the Scene Geometry, and the Speech Objects in the Scene.
4. Visual Object Identification assigns an Identifier to each Visual Object indicated by a human.
5. Audio-Visual Alignment uses the Audio Scene Description and Visual Scene Description to assign unique Identifiers to Audio, Visual, and Audio-Visual Objects.
6. Automatic Speech Recognition recognises Speech utterances.
7. Natural Language Understanding refines Text and extracts Meaning.
8. Personal Status Extraction extracts a human's Personal Status.
9. Entity Dialogue Processing produces the machine's response and its Personal Status.
10. Personal Status Display produces a speaking Avatar expressing Personal Status.

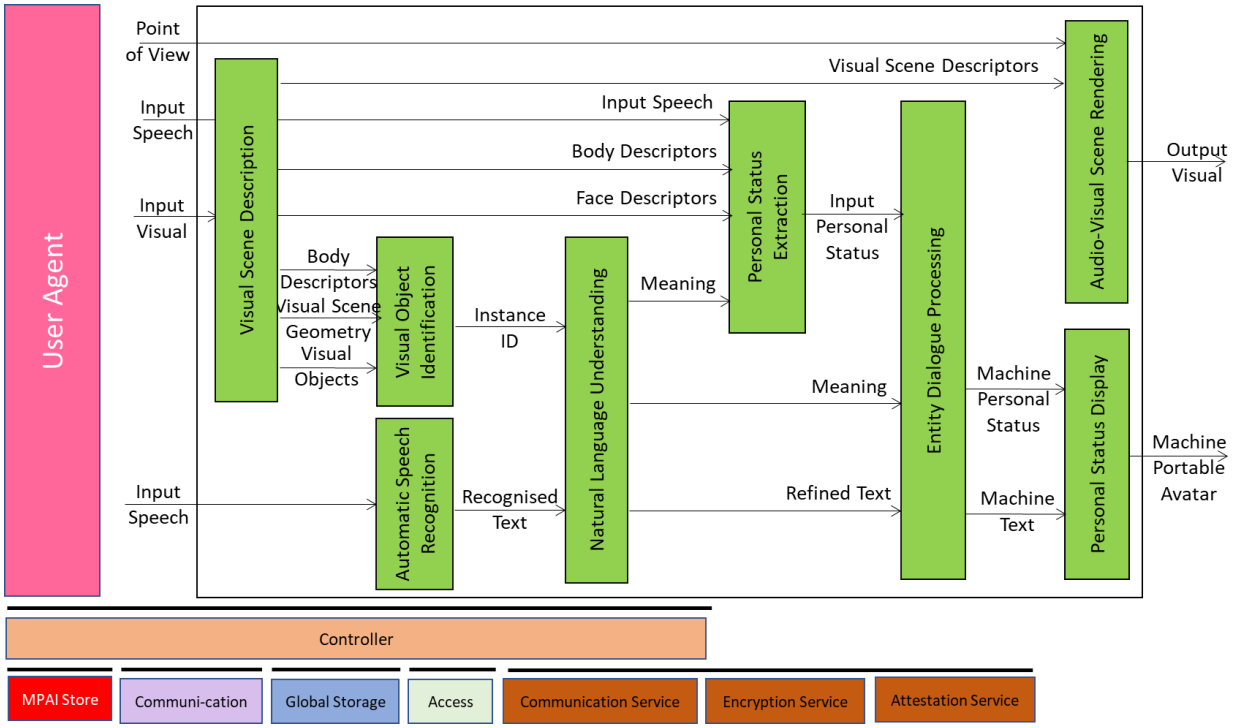

Figure 2 – Reference Model of Conversation with Personal Status

5.2.3 I/O Data of Conversation with Personal Status

Table 3 gives the input and output data of the Conversation with Personal Status Use Case:

Table 3 – I/O Data of Conversation with Personal Status

Input	Descriptions
Input Text	Text typed by the human as additional information stream or as a replacement of the Speech.
Input Speech	Speech of the human having a conversation with the machine.
Input Visual	Visual information of the Face and Body of the human having a conversation with the machine.
Input Selector	Data determining the use of Speech vs Text.
Output	**Descriptions**
Machine Portable Avatar	Text of the Speech produced by the machine.

5.2.4 Functions of AI Modules of Conversation with Personal Status

Table 4 provides the functions of the Conversation with Personal Status Use Case.

Table 4 - Functions of AI Modules of Conversation with Personal Status

AIM	Function
Visual Scene Description	1. Receives Input Visual. 2. Provides Visual Objects and Visual Scene Geometry.
Audio Scene Description	1. Receives Input Audio. 2. Provides Speech Objects and Audio Scene Geometry.
Visual Object Identification	1. Receives Visual Scene Geometry, Body Descriptors, and Visual Objects. 2. Provides Visual Object Instance IDs.
Automatic Speech Recognition	1. Receives Input Speech. 2. Extracts Recognised Text.
Natural Language Understanding	1. Receives Recognised Text. 2. Refines Text and extracts Meaning.
Personal Status Extraction	1. Receives Meaning, Refined Text, Body Descriptors, and Face Descriptors. 2. Extracts Personal Status.
Entity Dialogue Processing	1. Receives Refined Text and Personal Status. 2. Produces machine's Text and Personal Status.
Personal Status Displays	1. Receives Machine Text and Personal Status. 2. Synthesises Machine Portable Avatar.

5.2.5 I/O Data of AI Modules of Conversation with Personal Status

Table 5 provides the I/O Data of the AI Modules of the Conversation with Personal Status Use Case.

Table 5 - I/O Data of AI Modules of Conversation with Personal Status

AIM	Receives	Produces
Visual Scene Description	Input Visual	1. Face Descriptors

		2. Body Descriptors 3. Visual Scene Geometry 4. Visual Objects
Audio Scene Description	Input Audio	1. Speech 2. Audio Scene Geometry
Visual Object Identification	1. Body Descriptors 2. Visual Scene Geometry 3. Visual Objects	Visual Object ID
Automatic Speech Recognition	Input Speech	Recognised Text
Natural Language Understanding	1. Visual Object ID 2. Input Text 3. Recognised Text 4. Input Selector	1. Meaning 2. Refined Text
Personal Status Extraction	1. Body Descriptors 2. Face Descriptors 3. Meaning 4. Speech	Input Personal Status
Entity Dialogue Processing	1. Input Text 2. Refined Text 3. Input Personal Status 4. Input Selector	1. Machine Personal Status 2. Machine Text
Personal Status Displays	1. Machine Text 2. Machine Personal Status	Machine Portable Avatar

5.2.6 Specification of AIMs and JSON Metadata of Conversation with Personal Status

Table 6 – AIMs and JSON Metadata

MMC-CPS			Conversation With Personal Status	X
-	OSD-VSD		Visual Scene Description	X
-	CAE-ASD		Audio Scene Description	X
	-	CAE-AAT	Audio Analysis Transform	X
	-	CAE-ASL	Audio Source Localisation	X
	-	CAE-ASE	Audio Separation and Enhancement	X
	-	CAE-AST	Audio Synthesis Transform	X
	-	CAE-AMX	Audio Descriptor Multiplexing	X
-	OSD-VSD		Visual Scene Description	X
-	OSD-VOI		Visual Object Identification	X
	-	OSD-VDI	Visual Direction Identification	X
	-	OSD-VOE	Visual Object Extraction	X
	-	OSD-VII	Visual Instance Identification	X
-	OSD-AVA		Audio-Visual Alignment	X
-	MMC-ASR		Automatic Speech Recognition	X
-	MMC-NLU		Natural Language Understanding	X
-	MMC-PSE		Personal Status Extraction	X
	-	MMC-ITD	Input Text Description	X
	-	MMC-ISD	Input Speech Description	X
	-	PAF-IFD	Input Face Description	X
	-	PAF-IBD	Input Body Description	X

	-	MMC-PTI	PS-Text Interpretation	X
	-	MMC-PSI	PS-Speech Interpretation	X
	-	PAF-PFI	PS-Face Interpretation	X
	-	PAF-PGI	PS-Gesture Interpretation	X
	-	MMC-PMX	Personal Status Multiplexing	X
-	MMC-EDP		Entity Dialogue Processing	X
-	PAF-PSD		Personal Status Display	X
	-	MMC-TTS	Text-to-Speech	X
	-	PAF-IFD	Input Face Description	X
	-	PAF-IBD	Input Body Description	X
	-	PAF-PMX	Portable Avatar Multiplexing	X

5.3 Conversation About a Scene (MMC-CAS)

5.3.1 Scope of Conversation About a Scene

This Use Case addresses the case of a human holding a conversation with a Machine:

1. The human converses with the Machine indicating the object in the Environment s/he wishes to talk to or ask questions about it using Speech, Face, and Gesture.
2. The Machine

2.1. Sees and hears an Environment containing a speaking human and some scattered objects.

2.2. Recognises the human's Speech and obtains the human's Personal Status by capturing Speech, Face, and Gesture.

2.3. Understands which object the human is referring to and generates an avatar that:

2.3.1. Utters Speech conveying a synthetic Personal Status that is relevant to the human's Personal Status as shown by his/her Speech, Face, and Gesture, and

2.3.2. Displays a face conveying a Personal Status that is relevant to the human's Personal Status and to the response the Machine intends to make.

2.4. Renders the Scene that it perceives from a human-selected Point of View. The objects in the scene are labelled with the Machine's understanding of their semantics so that the human can understand how the Machine sees the Environment.

5.3.2 Reference Model of Conversation About a Scene

Figure 3 gives the Conversation About a Scene Reference Model including the input/output data, the AIMs, and the data exchanged between and among the AIMs.

The Machine operates according to the following workflow:

1. Visual Scene Description produces Body Descriptors, Visual Scene Geometry and Visual Objects from Input Visual.
2. Automatic Speech Recognition produces Recognised Text from Input Speech.
3. Visual Object Identification produces Visual Object Instance ID from Visual Objects, Body Descriptors, and Visual Scene Geometry.
4. Natural Language Understanding produces Meaning and Refined Text from Recognised Text and Visual Object ID.
5. Personal Status Extraction produces Input Personal Status from Meaning, Input Speech, Face Descriptors, and Body Descriptors.
6. Entity Dialogue Processing produces Machine Text and Machine Personal Status from Input Personal Status, Meaning, and Refined Text.
7. Personal Status Display produces Machine Portable Avatar from Machine Text, and Machine Personal Status.

8. Audio-Visual Scene Rendering rendered the Scene as seen from the user-selected Point of View using the Visual Scene Descriptors. The rendering is constantly updated as the machine improves its understanding of the scene and its objects.

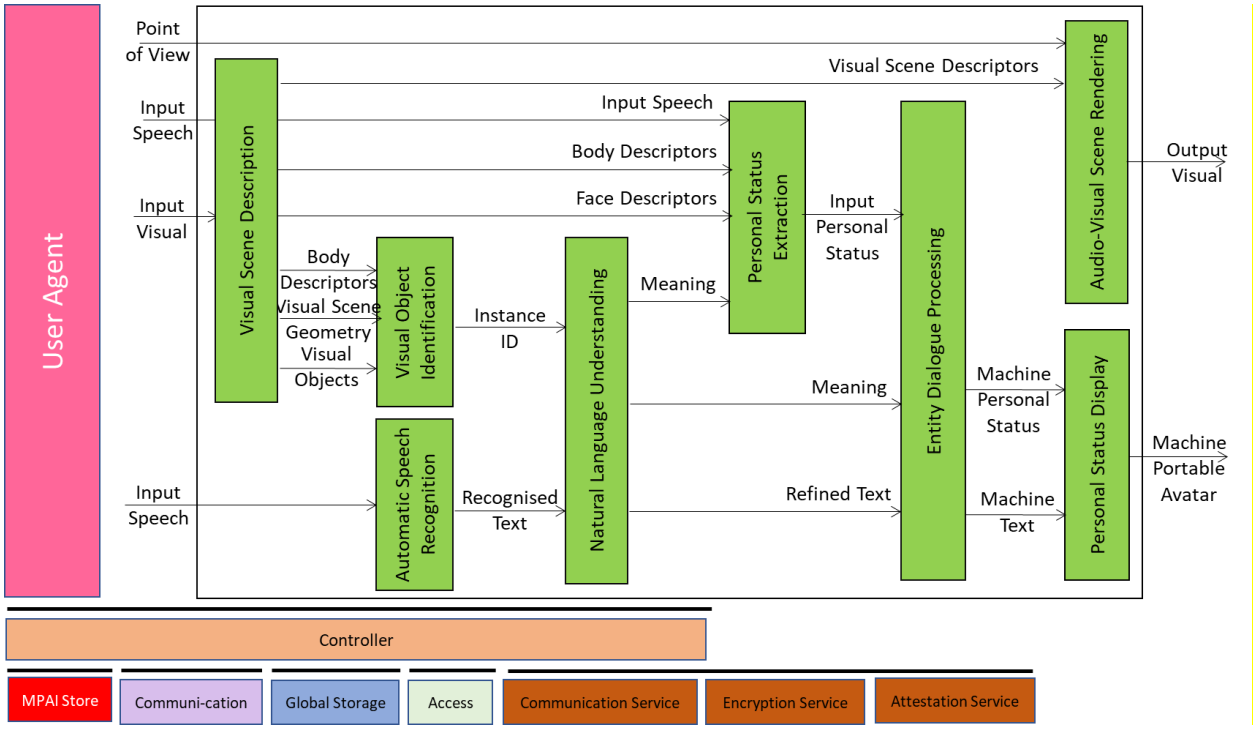

Figure 3 – Reference Model of Conversation About a Scene

5.3.3 I/O Data of Conversation About a Scene

Table 7 gives the input/output data of Conversation About a Scene.

Table 7 – I/O data of Conversation About a Scene

Input data	From	Description
Input Visual	Camera	Points to human and scene.
Input Speech	Microphone	Speech of human.
Point of View	Human	The point of view of the scene displayed by Scene Presentation.
Output data	**To**	**Descriptions**
Output Visual	Human	Rendering of the Scene containing labelled objects as perceived by Machine and seen from the Point of View.
Machine Portable Avatar	Human	Machine's avatar.

5.3.4 Functions of AI Modules of Conversation About a Scene

Table 24 provides the functions of the Conversation About a Scene Use Case.

Table 8 - Functions of AI Modules of Conversation About a Scene

AIM	Functions
Visual Scene Description	1. Receives Input Visual 2. Provides Visual Objects and Visual Scene Geometry.

Visual Object Identification	1. Receives Body Descriptors and non-human Visual Objects 2. Provides the Instance ID of the Visual Object indicated by the human.
Automatic Speech Recognition	1. Receives Input Speech 2. Provides Recognised Text.
Natural Language Understanding	1. Receives Instance ID and Recognised Text 2. Refines Text and extracts Meaning.
Personal Status Extraction	1. Receives Input Speech, Body Descriptors, Face Descriptors, and Meaning. 2. Provides Personal Status.
Entity Dialogue Processing	1. Receives Refined Text and Personal Status. 2. Produces Machine's Text and Personal Status.
Audio-Visual Scene Rendering	1. Receives the Descriptors of the Visual Scene perceived by Machine. 2. Renders the Visual Scene from the Point of View selected by human.
Personal Status Display	1. Receives Machine's Personal Status and Text. 2. Provides Machine Portable Avatar.

5.3.5 I/O Data of AI Modules of Conversation About a Scene

Table 9 gives the list of AIMs with their I/O Data.

Table 9 – AI Modules of Conversation About a Scene

AIM	Receives	Produces
Visual Scene Description	Input Visual	1. Visual Scene Descriptors 2. Body Descriptors 3. Face Descriptors 4. Visual Scene Geometry 5. Visual Objects
Visual Object Identification	1. Body Object 2. Visual Objects 3. Visual Scene Geometry	Visual Object Instance ID
Automatic Speech Recognition	Input Speech	Recognised Text
Natural Language Understanding	1. Recognised Text 2. Visual Object ID	1. Meaning 2. Refined Text
Personal Status Extraction	1. Body Object 2. Face Object 3. Input Speech 4. Meaning	Personal Status
Entity Dialogue Processing	1. Personal Status 2. Meaning 3. Refined Text	Machine Personal Status
Audio-Visual Scene Rendering	1. Visual Scene Descriptors 2. Point of View	Output Visual
Personal Status Display	1. Machine Text 2. Machine Personal Status	Machine Portable Avatar

5.3.6 Specification of Conversation About a Scene AIMs and JSON Metadata

Table 10 – AIMs and JSON Metadata

AIW and AIMs	Name and AIW/AIM Specification	JSON
MMC-CAS	Conversation About a Scene	X
- OSD-VSD	Visual Scene Description	X
- OSD-VOI	Visual Object Identification	X
- OSD-VDI	Visual Direction Identification	X
- OSD-VOE	Visual Object Extraction	X
- OSD-VII	Visual Instance Identification	X
- MMC-ASR	Automatic Speech Recognition	X
- MMC-NLU	Natural Language Understanding	X
- MMC-PSE	Personal Status Extraction	X
- MMC-ITD	Input Text Description	X
- MMC-ISD	Input Speech Description	X
- PAF-IFD	Input Face Description	X
- PAF-IBD	Input Body Description	X
- MMC-PTI	PS-Text Interpretation	X
- MMC-PSI	PS-Speech Interpretation	X
- PAF-PFI	PS-Face Interpretation	X
- PAF-PGI	PS-Gesture Interpretation	X
- MMC-PMX	Personal Status Multiplexing	X
- PAF-AVR	Audio-Visual Scene Rendering	X
- PAF-PSD	Personal Status Display	X
- OSD-AVS	Audio-Visual Scene Description	X
- MMC-TTS	Text-to-Speech	X
- PAF-IFD	Input Face Description	X
- PAF-IBD	Input Body Description	X
- PAF-PMX	Portable Avatar Multiplexing	X

5.4 Virtual Meeting Secretary (MMC-VMS)

5.4.1 Scope of Virtual Meeting Secretary

In Avatar Based Videoconference [3], i.e., a videoconference where avatars participate realistically impersonating the human participants, the Virtual Secretary is tasked with:

1. Listening to the Speech of each avatar.
2. Monitoring their Personal Status.
3. Drafting a Summary using the avatars' Personal Status and Text obtained from Automatic Speech Recognition or directly via Text input in the meeting's common language handled in two different ways:
   3.1. Transferred to an external application so that participants can edit the Summary.
   3.2. Displayed to avatars:
      3.2.1. Avatars make Speech comments or Text comments (e.g., offline via chat).

3.2.2. The Virtual Secretary edits the Summary interpreting Text, and the avatars' Personal Statuses.

Chapter 5 of Annex 1 - MPAI Basics provides additional information on the Avatar-Based Videoconference Use Case.

5.4.2 Reference Model of Virtual Meeting Secretary

Figure 4 specifies the architecture of the Virtual Secretary AIW. It is assumed that Meaning represents both meaning of Input Text and meaning of Refined Text.

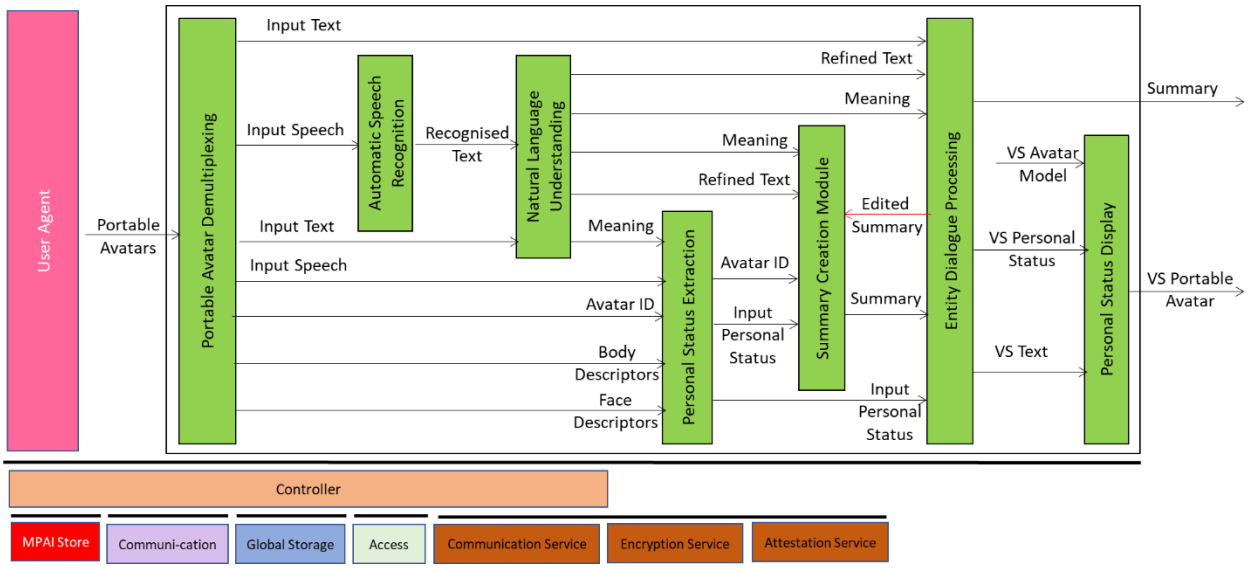

Figure 4 – Reference Model of the Virtual Meeting Secretary Use Case

The Virtual Secretary processes one avatar at a time according to the following workflow:

1. Portable Avatar Demultiplexing produces Input Text, Input Speech, Avatar ID, Body Descriptors, and Face Descriptors.
2. Automatic Speech Recognition extracts Text from Avatar Speech.
3. Natural Language Understanding:
   3.1. Receives Recognised Text.
   3.2. Produces Refined Text (of Recognised Text) and Meaning.
4. Personal Status Extraction:
   4.1. Receives Meaning, Speech, and Body and Face Descriptors.
   4.2. Produces the Personal Status of the avatar it is interacting with.
5. Summary Creation Module:
   5.1. Receives Refined Text, Personal Status, and Meaning
   5.2. Produces Summary using Personal Status and Text in the meeting's common language.
   5.3. Receives Edited Summary from Entity Dialogue Processing.
6. Entity Dialogue Processing:
   6.1. Sends Summary to external application.
   6.2. Sends Edited Summary produced from Refined Text (from Speech), Avatar's Text (from chat), Meaning, and Summary back to Summary Creation Module.
   6.3. Produces VS Text and VS Personal Status.
7. Personal Status Display produces VS Portable Avatar containing VS Avatar Model, VS Text, VS Speech, and VS Avatar Descriptors.

5.4.3 I/O Data of Virtual Meeting Secretary

Table 11 gives the input/output data of Virtual Meeting Secretary.

Table 11 – I/O data of Virtual Meeting Secretary

Input data	From	Description
Portable Avatar	Server	Portable Avatars as re-multiplexed by Server
Output data	**To**	**Descriptions**
VS Portable Avatar	Server	VS Portable Avatar to Server
Summary	Server	Summary of avatars' interventions

5.4.4 Functions of AI Modules of Virtual Meeting Secretary

Table 12 gives the functions of Virtual Meeting Secretary AIMs.

Table 12 – Functions of Virtual Meeting Secretary AI Modules

AIM	Functions
Portable Avatar Demultiplexing	1. Receives Portable Avatar. 2. Provides the Data required by Virtual Secretary's AIMs.
Automatic Speech Recognition	1. Receives Speech. 2. Provides Recognised Text.
Natural Language Understanding	1. Refines Recognised Text. 2. Extracts Meaning.
Personal Status Extraction	1. Receives Meaning, Input Speech, Body Descriptors, Face Descriptors. 2. Extracts Personal Status.
Summary Creation Module	1. Receives Meaning, Refined Text, Avatar ID, Input Personal Status of Avatar ID, and Edited Summary (from Entity Dialogue Processing.) 2. Produces and refines Summary using Edited Summary.
Entity Dialogue Processing	1. Receives Input Text, Refined Text, Meaning, Summary, Input Personal Status. 2. Produces Text, Virtual Secretary Personal Status, and Edited Summary.
Personal Status Display	1. Receives Virtual Secretary's Avatar Model, Personal Status, and Text. 2. Shows Virtual Secretary as Virtual Secretary Portable Avatar.

5.4.5 I/O Data of AI Modules of Virtual Meeting Secretary

Table 13 gives the AI Modules of the Virtual Meeting Secretary depicted in Figure 4.

Table 13 – AI Modules of Virtual Meeting Secretary

AIM	Receives	Produces
Portable Avatar Demultiplexing	Portable Avatar	1. Input Text 2. Input Speech 3. AvatarID 4. Body Descriptors 5. Face Descriptors
Automatic Speech Recognition	Speech	Recognised Text
Natural Language Understanding	Recognised Text	1. Refined Text

		2. Meaning
Personal Status Extraction	1. Meaning 2. Speech 3. Face Descriptors 4. Body Descriptors	Personal Status
Summary Creation Module	1. Meaning 2. Refined Text 3. Edited Summary	Summary
Entity Dialogue Processing	1. Refined Text 2. Personal Status 3. Meaning 4. Summary	1. VS Personal Status 2. VS Text 3. Edited Summary
Personal Status Display	1. VS Text 2. VS Personal Status	PersonalVS Avatar

5.4.6 Specification of Virtual Meeting Secretary AIMs and JSON Metadata

Table 14 – AIMs and JSON Metadata

AIW and AIMs	Name and AIW/AIM Specification	JSON
MMC-VMS	Virtual Meeting Secretary	X
- PAF-PDX	Portable Avatar Demultiplexing	X
- MMC-ASR	Automatic Speech Recognition	X
- MMC-NLU	Natural Language Understanding	X
- MMC-PSE	Personal Status Extraction	X
- MMC-ITD	Input Text Description	X
- MMC-ISD	Input Speech Description	X
- PAF-IFD	Input Face Description	X
- PAF-IBD	Input Body Description	X
- MMC-PTI	PS-Text Interpretation	X
- MMC-PSI	PS-Speech Interpretation	X
- PAF-PFI	PS-Face Interpretation	X
- PAF-PGI	PS-Gesture Interpretation	X
- MMC-PMX	Personal Status Multiplexing	X
- MMC-SCM	Summary Creation Module	X
- MMC-EDP	Entity Dialogue Processing	X
- PAF-PSD	Personal Status Display	X
- MMC-TTS	Text-to-Speech	X
- PAF-IFD	Input Face Description	X
- PAF-IBD	Input Body Description	X
- PAF-PMX	Portable Avatar Multiplexing	X

5.5 Human-Connected Autonomous Vehicle (CAV) Interaction (MMC-HCI)

5.5.1 Functions of Human-CAV Interaction Subsystem

The MPAI Connected Autonomous Vehicle (CAV) – Architecture specifies the Reference Model of a Vehicle – called Connected Autonomous Vehicle (CAV) – able to reach a destination by understanding the environment using its own sensors, exchanging information with other CAVs and actuating motion. The Reference Model subdivides a CAV in four Subsystems. Annex 1 - MPAI Basics Chapter 6 introduces MPAI-CAV and [5] provides the full specification.

The Human-CAV interaction (HCI) Subsystem has the function to recognise the human owner or renter, respond to humans' commands and queries, converse with humans during the travel, exchange information with the Autonomous Motion Subsystem in response to humans' requests, and communicate with HCIs on board other CAVs.

5.5.2 Reference Model of Human-CAV Interaction Subsystem

Figure 5 represents the Human-CAV Interaction (HCI) Reference Model.

Note that it is assumed that Natural Language Understanding produces a Refined Text that is either the refined Recognised Text or the Input Text, depending on which one is active. Meaning is always computed based on the available text - Refined or Input. Personal Status Extraction is unaware of the decisions made by Natural Language Understanding.

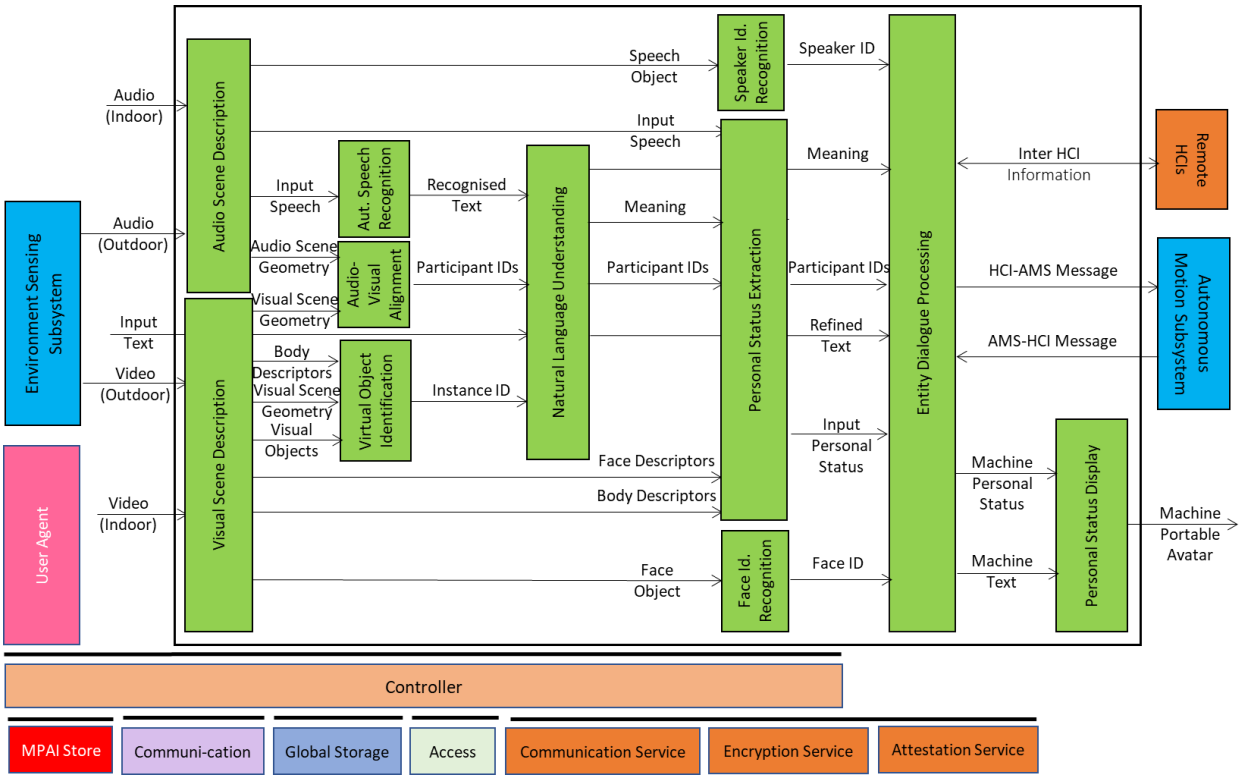

Figure 5 – Human-CAV Interaction Reference Model

A group of humans approaches the CAV outside the CAV: or sitting seats inside the CAV:

1. Audio Scene Description AIM creates the Audio Scene Description in the form of Audio (Speech) Objects corresponding to each speaking human in the Environment (close to the CAV) and Audio Scene Geometry.
2. Visual Scene Description creates the Visual Scene Descriptors in the form of Descriptors of the Faces and the Bodies corresponding to each human in the Environment (close to the CAV) and Visual Scene Geometry.

3. Automatic Speech Recognition recognises the speech of each human and produces Recognised Text.
4. Audio-Visual Alignment produces the Audio-Visual Scene Geometry.
5. Visual Object Identification produces Object ID from Visual Objects, Body Descriptors, and Visual Scene Geometry.
6. Natural Language Understanding extracts Meaning and produces Refined Text from the Recognised Text of each Input Speech and Visual Object.
7. The Speaker Identity Recognition and Face Identity Recognition AIMs authenticate the humans that the HCI is interacting with using Speech and Face Descriptors.
8. The Personal Status Extraction AIM extracts the Personal Status of the humans.
9. The Personal Status Display produces the ready-to-render Machine Portable Avatar [**Error! Reference source not found.**] conveying Machine Speech and Machine Personal Status.
   9.1. Issues commands to the Autonomous Motion Subsystem.
   9.2. Receives and processes responses from the Autonomous Motion Subsystem.
   9.3. Communicates with Remote HCIs.

The HCI interacts with the humans in the cabin in several ways:
1. By responding to commands/queries from one or more humans at the same time, e.g.:
   1.1.Commands to go to a waypoint, park at a place, etc.
   1.2.Commands with an effect in the cabin, e.g., turn off air conditioning, turn on the radio, call a person, open window or door, search for information etc.
2. Note: this document does not address the format in which the interactions the interaction of HCI with AMS (e.g., commands and responses regarding selection of Route by human) and with remote HCIs (see Figure 5).
3. By conversing with and responding to questions from one or more humans at the same time about travel-related issues (in-depth domain-specific conversation), e.g.:
   3.1.Humans request information, e.g., time to destination, route conditions, weather at destination, etc.
   3.2.CAV offers alternatives to humans, e.g., long but safe way, short but likely to have interruptions.
   3.3.Humans ask questions about objects in the cabin.
4. By following the conversation on travel matters held by humans in the cabin if 1) the passengers allow the HCI to do so, and 2) the processing is carried out inside the CAV.

Note that the version of the Audio Scene Description provides all the Speech Objects in the Audio Scene, removing all other audio sources. The Speaker Identity Recognition and Automatic Speech Recognition AIMs support multiple Speech Objects as input. Each Speech Object has an identifier to enable the Speaker Identity Recognition and Automatic Speech Recognition AIMs to provide labelled Speaker IDs and Recognised Texts. If the Face Identity Recognition AIM provides Face IDs corresponding to the Speaker IDs, the Entity Dialogue Processing AIM can correctly associate the Speaker IDs (and the corresponding Recognised Texts) with the Face IDs.

5.5.3 I/O Data of Human-CAV Interaction

Table 15 gives the input/output data of Human-CAV Interaction.

Table 15 - I/O data of Human-CAV Interaction

Input data	From	Description
Input Audio (Outdoor))	Environment Sensing Sub-system	User authentication User command

		User conversation
Input Audio (Indoor)	Cabin Passengers	User's social life Commands/interaction with HCI
Input Visual (Outdoor)	Environment Sensing Sub-system	Commands/interaction with HCI
Input Visual (Indoor)	Cabin Passengers	User's social life Commands/interaction with HCI
AMS-HCI Message	Autonomous Motion Sub-system	Includes response to HCI-AMS Message
Inter HCI Information	Remote HCI	HCI-to-HCI information
Output data	**To**	**Comments**
Inter HCI Information	Remote HCI	HCI-to-HCI information
HCI-AMS Message	Autonomous Motion Sub-system	HCI-to-AMS information
Machine Portable Avatar	Cabin Passengers	HCI's avatar.

5.5.4 Functions of AI Modules of Human-CAV Interaction

Table 16 gives the functions of all Human-CAV Interaction AIMs.

Table 16 – Functions of Human-CAV Interaction's AI Modules

AIM	Function
Audio Scene Description	1. Receives Input Audio captured by the appropriate (indoor or outdoor) Microphone Array. 2. Produces the Audio Scene Descriptors.
Visual Scene Description	1. Receives Input Visual captured by the appropriate (indoor or outdoor) visual sensors. 2. Produces the Visual Scene Descriptors.
Automatic Speech Recognition	1. Receives Input Speech from one of the human. 2. Converts speech into Recognised Text.
Audio-Visual Alignment	1. Receives Audio and Visual Scene Geometries and Audio and Visual Objects. 2. Re-identifies the Audio and Visual Objects having the same Spatial Attitudes.
Visual Object Identification	1. Receives Body Descriptors, Visual Scene Geometry, and Visual Objects. 2. Provides the ID of the class of objects of which the Visual Object is an Instance
Natural Language Understanding	1. Receives Recognised Text, Input Text, Visual Object Instance ID. 2. Produces Refined Text and Meaning.
Speaker Identity Recognition	1. Receives Speech Object. 2. Provides Speaker ID.
Personal Status Extraction	1. Receives Input Speech, Meaning, Body Descriptors, Face Descriptors. 2. Provides Input Personal Status of human.
Face Identity Recognition	1. Receives Face Object. 2. Provides Face ID.

Entity Dialogue Processing	1. Receives Speaker ID, Meaning, Refined Text, Input Personal Status, Face ID. 2. Provides Machine (HCI) Text and Personal Status.
Personal Status Display	1. Receives Machine Personal Status and Text. 2. Produces Machine Portable Avatar.

5.5.5 I/O Data of AI Modules of Human-CAV Interaction

Table 17 gives the AI Modules of the Human-CAV Interaction depicted in Figure 3.

Table 17 – AI Modules of Human-CAV interaction

AIM	Receives	Produces
Audio Scene Description	Input Audio (outdoor) Input Audio (indoor)	Speech Objects
Visual Scene Description	Input Video (outdoor) Input Video (indoor)	Face Objects Visual Objects Body Descriptors Face Descriptors
Automatic Speech Recognition	Speech Object	Recognised Text
Audio-Visual Alignment	Audio Scene Geometry Visual Scene Geometry	Participant ID
Visual Object Identification	Visual Object Visual Scene Geometry Body Descriptors	Visual Object Instance ID
Natural Language Understanding	Recognised Text Participant ID Visual Object Instance ID	Meaning Refined Text Participant ID
Speaker Identity Recognition	Speech Descriptors	Speaker ID
Personal Status Extraction	Input Speech Meaning Participant ID Face Descriptors Body Descriptors	Personal Status Participant ID
Face Identity Recognition	Face Object	Face ID
Entity Dialogue Processing	Participant ID Speaker ID Meaning Refined Text Personal Status Face ID AMS-HCI Response	AMS-HCI Commands Output Text Output Personal Status
Personal Status Display	Machine Text Output Personal Status	Machine Portable Avatar

5.5.6 Specification of Human-CAV Interaction AIMs and JSON Metadata

Table 18 - AIMs and JSON Metadata

AIMs	Name	JSON
MMC-HCI	Human-CAV Interaction	X
- CAE-ASD	Audio Scene Description	X
- CAE-AAT	Audio Analysis Transform	X
- CAE-ASL	Audio Source Localisation	X
- CAE-ASE	Audio Separation and Enhancement	X
- CAE-AST	Audio Synthesis Transform	X
- CAE-AMX	Audio Descriptor Multiplexing	X
- OSD-VSD	Visual Scene Description	X
- MMC-ASR	Automatic Speech Recognition	X
- OSD-AVA	Audio-Visual Alignment	X
- OSD-VOI	Visual Object Identification	X
- OSD-VDI	Visual Direction Identification	X
- OSD-VOE	Visual Object Extraction	X
- OSD-VII	Visual Instance Identification	X
- MMC-NLU	Natural Language Understanding	X
- MMC-SIR	Speaker Identity Recognition	X
- MMC-PSE	Personal Status Extraction	X
- MMC-ITD	Input Text Description	X
- MMC-ISD	Input Speech Description	X
- PAF-IFD	Input Face Description	X
- PAF-IBD	Input Body Description	X
- MMC-PTI	PS-Text Interpretation	X
- MMC-PSI	PS-Speech Interpretation	X
- PAF-PFI	PS-Face Interpretation	X
- PAF-PGI	PS-Gesture Interpretation	X
- MMC-PMX	Personal Status Multiplexing	X
- MMC-EDP	Entity Dialogue Processing	X
- PAF-FIR	Face Identity Recognition	X
- PAF-PSD	Personal Status Display	X
- MMC-TTS	Text-to-Speech	X
- PAF-IFD	Input Face Description	X
- PAF-IBD	Input Body Description	X
- PAF-PMX	Portable Avatar Multiplexing	X

5.6 Conversation with Emotion (MMC-CWE)

5.6.1 Scope of Conversation with Emotion

In the Conversation with Emotion (MMC-CWE) Use Case, a machine responds to a human's textual and/or vocal utterance in a manner consistent with the human's utterance and emotional state, as detected from the human's text, speech, or face. The machine responds using text, synthetic

speech, and a face whose lip movements are synchronised with the synthetic speech and the synthetic machine emotion.

5.6.2 Reference Model of Conversation with Emotion

Figure 6 gives the Reference Model of Conversation with Emotion including the input/output data, the AIMs, the AIM topology, and the data exchanged between and among the AIMs.

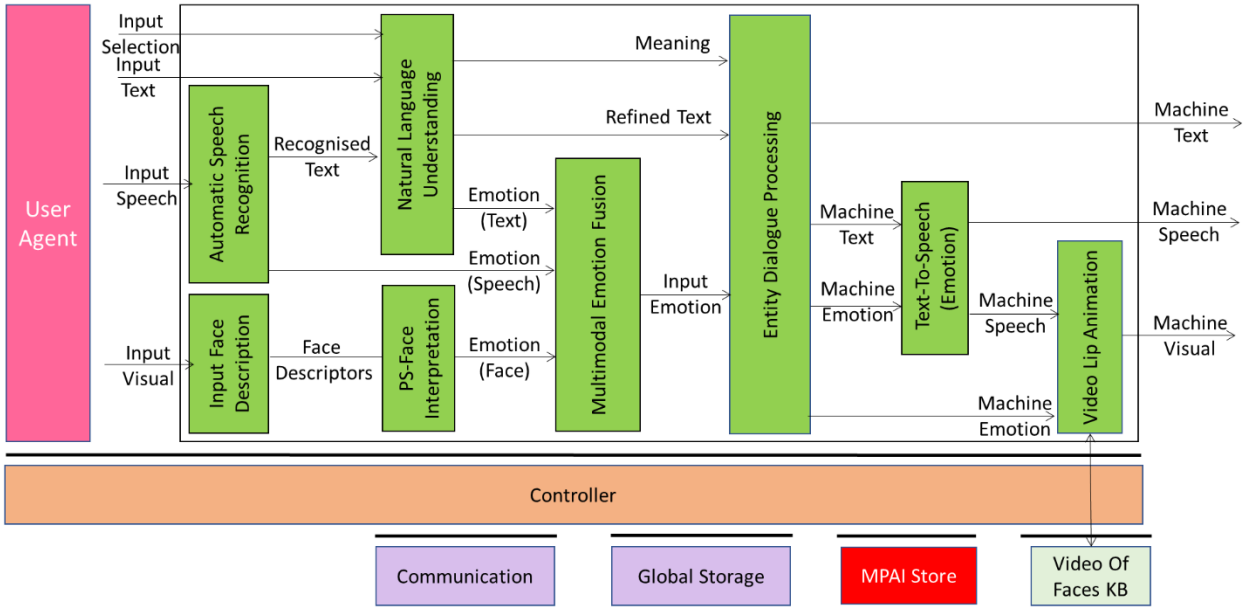

Figure 6 – Reference Model of Conversation With Emotion

The operation of Conversation with Emotion develops as follows:

1. Automatic Speech Recognition (Emotion) recognises Speech and produces Recognised Text and Emotion (Speech).
2. The Natural Language Understanding AIM refines Recognised Text and produces Meaning.
3. Emotion is recognised by the machine:

3.1. The Natural Language Understanding, Automatic Speech Recognition, and the combination of Input Face Description and PS-Face Interpretation AIMs independently extract a set of Emotion-related cues from Input Text, Input Speech, and Input Visual.

3.2. The Multimodal Emotion Fusion AIM fuses all Emotions into the Fused Emotion.

4. The Entity Dialogue Processing AIM produces a reply based on the Fused Emotion and Meaning.
5. The Text-To-Speech (Emotion) AIM produces Output Speech from Text with Emotion.
6. The Lips Animation AIM animates the lips of a Face drawn from the Video of Faces KB consistently with the Output Speech and the Output Emotion.

5.6.3 I/O Data of Conversation with Emotion

The input and output data of the Conversation with Emotion Use Case are:

Table 19 – I/O Data of Conversation with Emotion

Input	Descriptions
Input Selector	Data determining the use of Speech vs Text.
Text Object	Text typed by the human as additional information stream or as a replacement of the speech depending on the value of Input Selector.

Speech Object	Speech of the human having a conversation with the machine.
Face Object	Visual information of the Face of the human having a conversation with the machine.
Output	**Descriptions**
Text Object	Text of the Speech produced by the Machine.
Speech Object	Synthetic Speech produced by the Machine.
Face Object	Video of a Face whose lip movements are synchronised with the Output Speech and the synthetic machine emotion.

5.6.4 Functions of AI Modules of Conversation with Emotion

Table 20 provides the functions of the Conversation with Emotion AIMs.

Table 20 - Functions of AI Modules of Conversation with Emotion

AIM	Function
Automatic Speech Recognition	1. Receives Speech Object. 2. Produces Recognised Text.
Input Speech Description	1. Receives Speech Object. 2. Produces Speech Descriptors
Input Face Description	1. Receives Face Object. 2. Extracts Face Descriptors.
Natural Language Understanding	1. Receives Input Selector, Text Object, Recognised Text. 2. Produces Meaning (i.e., Text Descriptors), Refined Text.
PS-Speech Interpretation	1. Receives Speech Descriptors. 2. Provides the Emotion of the Face.
PS-Face Interpretation	1. Receives Face Descriptors. 2. Provides the Emotion of the Face.
PS-Text Interpretation	1. Receives Text Descriptors. 2. Provides the Emotion of the Text.
Multimodal Emotion Fusion	1. Receives Emotion (Text), Emotion (Speech), Emotion (Face). 2. Provides human's Input Emotion by fusing Emotion (Text), Emotion (Speech), and Emotion (Video).
Entity Dialogue Processing	1. Receives Refined Text, Meaning, Input Emotion. 2. Analyses Meaning and Input Text or Refined Text, depending on the value of Input Selector. 3. Produces Machine Emotion and Machine Text.
Text-To-Speech	1. Receives Machine Text and Machine Emotion. 2. Produces Output Speech.
Video Lip Animation	1. Receives Machine Speech and Machine Emotion. 2. Animates the lips of a video obtained by querying the Video Faces KB, using the Output Emotion. 3. Produces Face Object with synchronised Speech Object (Machine Object).

5.6.5 I/O Data of AI Modules of Conversation with Emotion

The AI Modules of Conversation with Emotion perform the Functions specified in Table 21.

Table 21 - AI Modules of Conversation with Emotion

AIM	Receives	Produces

Automatic Speech Recognition	Speech Object	Recognised Text
Input Speech Description	Speech Object	Speech Descriptors
Input Face Description	Input Visual	Face Descriptors
Natural Language Understanding	Recognised Text	Refined Text Meaning
PS-Speech Interpretation	Speech Descriptors	Emotion (Speech)
PS-Face Interpretation	Face Descriptors	Emotion (Face)
PS-Text Interpretation	Text Descriptors	Emotion (Text)
Multimodal Emotion Fusion	1. Emotion (Text) 2. Emotion (Speech) 3. Emotion (Face)	Input Emotion
Entity Dialogue Processing	1. Meaning 2. Based on Input Selector 2.1. Refined Text 2.2. Input Text. 3. Input Emotion.	1. Machine Text 2. Machine Emotion
Text-To-Speech (Emotion)	1. Machine Text 2. Machine Emotion	Output Speech.
Lips Animation	1. Machine Emotion 2. Machine Speech	Output Visual.

5.6.6 Specification of Conversation with Emotion AIMs and JSON Metadata

Table 22 – AIMs and JSON Metadata

MMC-CWE	Conversation With Emotion	X
- MMC-ASR	Automatic Speech Recognition	X
- MMC-ISD	Input Speech Description	X
- PAF-IFD	Input Face Description	X
- MMC-NLU	Natural Language Understanding	X
- MMC-PSI	PS-Speech Interpretation	X
- PAF-PFI	PS-Face Interpretation	X
- MMC-PTI	PS-Text Interpretation	X
- MMC-MEF	Multimodal Emotion Fusion	X
- MMC-EDP	Entity Dialogue Processing	X
- MMC-TTS	Text-to-Speech	X
- MMC-VLA	Video Lip Animation	X

5.7 Multimodal Question Answering (MMC-MQA)

5.7.1 Scope of Multimodal Question Answering

In a Question Answering (QA) System, a machine provides answers to a user's question presented in natural language. Multimodal Question Answering improves current QA systems that are only able to deal with text or speech inputs by offering the requesting human the ability to present both speech or text and images. For example, users might ask "Where can I buy this tool?" while showing the picture of the tool, even without showing their faces. In the Multimodal Question

Answering (MMC-MQA) Use Case, a machine responds to a question expressed by a user in text or speech while showing an object. The machine's response may use text and synthetic speech.

5.7.2 Reference Model of Multimodal Question Answering

Figure 7 gives the Multimodal Question Answering Reference Model including the input/output data, the AIMs, and the data exchanged between and among the AIMs.

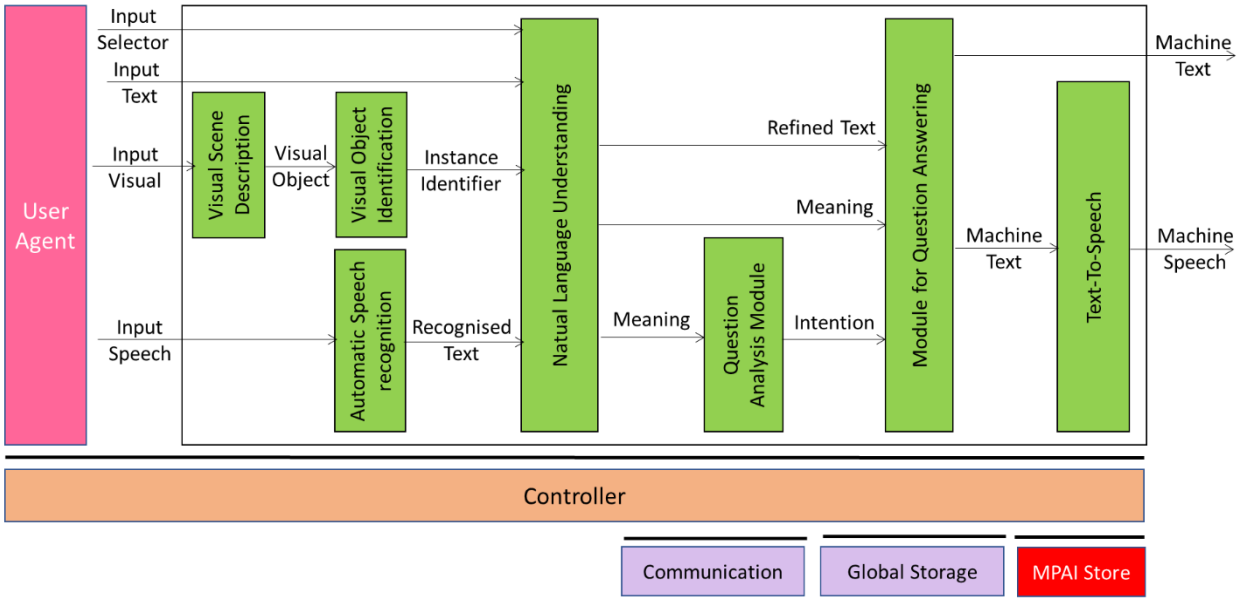

Figure 7 – Reference Model of Multimodal Question Answering

The operation of Multimodal Question Answering develops in the following way:

1. Input Selector is used to inform the machine whether the human employs Text or Speech to query the machine.
2. Depending on the value of Input Selector, Natural Language Understanding:

2.1. Extracts the Meaning of the question from Recognised Text and refines Recognised Text.

2.2. Extracts the Meaning of the question from Input Text.

3. Visual Scene Description extracts the Visual Object.
4. Visual Object Identification identifies the Visual Object.
5. Question Analysis Module determines the Intention of the question.
6. Module for Question Answering uses Intention and Meaning to produce the answer as Machine Text.
7. Text-To-Speech produces the Output Speech from Machine Text.

5.7.3 I/O Data of Multimodal Question Answering

The input and output data of the Multimodal Question Answering Use Case are:

Table 23 – I/O Data of Multimodal Question Answering

Input	**Descriptions**
Input Text	Text typed by the human as a replacement for Input Speech.
Input Selector	Data determining the use of Speech or Text.
Input Video	Video of the human showing an object held in hand.
Input Speech	Speech of the human asking a question the Machine.
Output	**Descriptions**

Machine Text	The Text generated by Machine in response to human input.
Machine Speech	The Speech generated by Machine in response to human input.

5.7.4 Functions of AI Modules of Multimodal Question Answering

Table 24 provides the functions of the Multimodal Question Answering Use Case.

Table 24 - Functions of AI Modules of Multimodal Question Answering

AIM	Function
Visual Scene Description	Extracts the Visual Object in the Visual Scene.
Visual Object Identification	Identifies the Visual Object.
Automatic Speech Recognition	Recognises Speech.
Natural Language Understanding	Extracts Meaning and refines Text from Recognised Text.
Question Analysis Module	Extracts Intention from Text.
Answer to Question Module	Produces the Machine response to the query.
Text-To-Speech	Synthesises Speech from Text.

5.7.5 I/O Data of AI Modules of Multimodal Question Answering

The AI Modules of Multimodal Question Answering are given in *Table 25*.

Table 25 – AI Modules of Multimodal Question Answering

AIM	Receives	Produces
Visual Scene Description	Input Video	Visual Object
Object Identification	Visual Object	Visual Object Identifier
Automatic Speech Recognition	Input Speech	Recognised Text
Natural Language Understanding	Input Text or Recognised Text based on Input Selector	Refined Text Meaning
Question Analysis Module	Meaning	Intention
Module for Question Answering	1. Input or Recognised Text (based on Input Selector) 2. Intention 3. Meaning	Machine Text
Text-To-Speech	Machine Text	Machine Speech

5.7.6 JSON Metadata of Multimodal Question Answering

Table 26 - Acronyms and URs of JSON Metadata

MMC-MQA	Multimodal Question Answering	X
- OSD-VSD	Visual Scene Description	X
- OSD-VOI	Visual Object Identification	X
- OSD-VDI	Visual Direction Identification	X
- OSD-VOE	Visual Object Extraction	X
- OSD-VII	Visual Instance Identification	X

- MMC-ASR Automatic Speech Recognition X
- MMC-NLU Natural Language Understanding X
- MMC-QAM Question Analysis Module X
- MMC-AQM Answer to Question Module X
- MMC-TTS Text-to-Speech X

5.8 Unidirectional Speech Translation (MMC-UST)

5.8.1 Scope of Unidirectional Speech Translation

The goal of the Unidirectional Speech Translation (MMC-UST) Use Case is to translate speech segments expressed in a source language into a target language or to produce the textual version of the translated speech. If the desired output is speech, the user can specify whether their speech features (voice colour, emotional charge, etc.) should be preserved in the translated speech.

The flow of control is from Input Speech or Input Text to Translated Text, and then to Output Speech and Output Text. Depending on the value of Input Selector:

1. Input Text in Language A is translated into Translated Text in Language B and pronounced as Speech in Language B.
2. The Speech features (voice colour, emotional charge, etc.) in Language A are preserved in Language B.

5.8.2 Reference Model of Unidirectional Speech Translation

Figure 8 describes the input/output data, the AIMs and the data exchanged between AIMs.

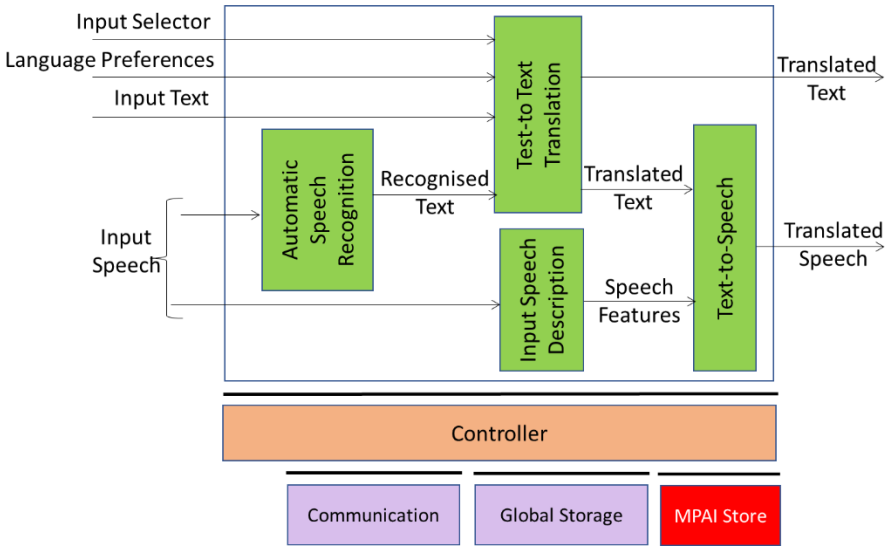

Figure 8 – Reference Model of Unidirectional Speech Translation (UST)

5.8.3 I/O Data of Unidirectional Speech Translation

The input and output data of the Unidirectional Speech Translation Use Case are:

Table 27 – I/O Data of Unidirectional Speech Translation

Input	Descriptions
Input Selector	Determines whether: 1. The input will be in Text or Speech

	2. The Input Speech features are preserved in the Output Speech.
Language Preferences	User-specified input Language (A) and output Language (B).
Input Speech	Speech produced in Language A by a human desiring translation into language B.
Input Text	Alternative textual source information to be translated into and pronounced in language B depending on the value of Input Selector.
Output	**Descriptions**
Translated Speech	Input Speech translated into language B preserving the Input Speech features in the Output Speech, depending on the value of Input Selector.
Translated Text	Text of Input Speech or Input Text translated into language B, depending on the value of Input Selector.

5.8.4 Functions of AI Modules of Unidirectional Speech Translation

Table 28 gives the functions of Unidirectional Speech Translation AIMs.

Table 28 – Functions of Unidirectional Speech Translation AI Modules

AIM	Functions
Automatic Speech Recognition	Recognises Speech
Text-to-Text Translation	Translates Recognised Text
Input Speech Description	Extracts Speech Features
Text-To-Speech (Features)	Synthesises Translated Text adding Speech Features

5.8.5 I/O Data of AI Modules of Unidirectional Speech Translation

The AI Modules of Unidirectional Speech Translation are given in *Table 29*.

Table 29 – AI Modules of Unidirectional Speech Translation

AIM	Receives	Produces
Automatic Speech Recognition	Input Speech Segment	Recognised Text
Text-to-Text Translation	1. Input Text 2. Recognised Text (Based on Input Selector)	Translated Text
Input Speech Description	Input Speech	Speaker-specific Speech Features
Text-To-Speech (Features)	1. Translated Text 2. Speech Features (depending on Input Selector)	Produces Output Speech.

5.8.6 Specification of Unidirectional Speech Translation AIMs and JSON Metadata

Table 30 – AIMs and JSON Metadata

MMC-BST	Bidirectional Speech Translation	X
- MMC-ASR	Audio Scene Description	X
- MMC-TTT	Text-to-Speech	X

- MMC-ISD Input Speech Description X
- MMC-TTS Text-to-Speech X

5.9 Bidirectional Speech Translation (MMC-BST)

5.9.1 Scope of Bidirectional Speech Translation

The goal of the Bidirectional Speech Translation (MMC-BST) Use Case is to support a conversation between two people, each speaking a different language. The machine translates each input speech segment into the selected language as speech or text. If the desired output is speech, users can specify whether their speech features (voice colour, emotional charge, etc.) should be preserved in the translated speech.
The flow of control (from Input Speech to Translated Text to Output Speech) is identical to that of the Unidirectional case. The difference is that, rather than one such flow, two flows are provided in two different channels – the first from language A to language B, and the second from language B to language A.

Depending on the value of Input Selector:
1. Input Text in Language A is translated into Translated Text in Language B and pronounced as Speech in Language B.
2. The Speech features (voice colour, emotional charge, etc.) in Language A are preserved in Language B.

The same applies for the Language-B-to-Language-A channel.

5.9.2 Reference Model of Bidirectional Speech Translation

Figure 9 depicts the AIMs and the data exchanged between AIMs.

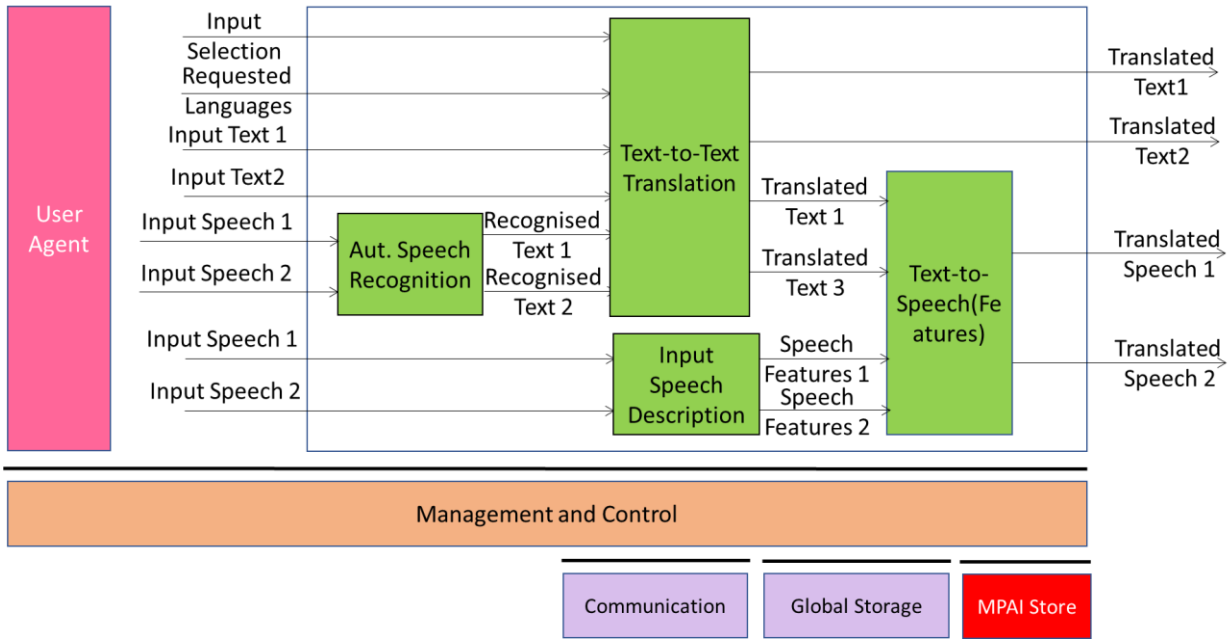

Figure 9 – Reference Model of Bidirectional Speech Translation (BST)

5.9.3 I/O Data of Bidirectional Speech Translation

The input and output data of the Bidirectional Speech Translation Use Case are:

Table 31 – I/O Data of Bidirectional Speech Translation

Input	Descriptions
Input Selector	Determines whether the input will be Text or Speech.
Requested languages	User-specified input language and output languages
Input Speech1	Speech by human1 desiring spoken translation in the specified language.
Input Text1	Alternative Input Text to be translated to the specified language.
Input Speech2	Speech by human2 desiring spoken translation in the specified language.
Input Text2	Alternative Input Text to be translated to the specified language.
Output	**Descriptions**
Output Speech1	Translated Speech of Speaker 1.
Output Text1	Text of the translated Speech of Speaker 1.
Output Speech2	Translated Speech of Speaker 2.
Output Text2	Text of the translated Speech of Speaker 2.

5.9.4 Functions of AI Modules of Bidirectional Speech Translation

Table 32 gives the functions of Bidirectional Speech Translation AIMs.

Table 32 – Functions of Bidirectional Speech Translation AI Modules

AIM	Functions
Automatic Speech Recognition	Recognises Speech
Text-to-Text Translation	Translates Recognised Text
Input Speech Description	Extracts Speech Features
Text-To-Speech (Features)	Synthesises Translated Text adding Speech Features

5.9.5 I/O Data of AI Modules of Bidirectional Speech Translation

Table 33 gives the I/O Data of the AI Modules.

Table 33 – AI Modules of Bidirectional Speech Translation

AIM	Receives	Produces
Automatic Speech Recognition	1. Input Speech 1 Segment 2. Input Speech 2 Segment	1. Recognised Text 1 2. Recognised Text 2.
Text-to-Text Translation	1. Input Text 1 or Recognised Text 1 2. Input Text 2 or Recognised Text 2 3. based on the value of Input Selector	1. Translated Text 1 2. Translated Text 2.
Input Speech Description	1. Input Speech 1 2. Input Speech 2	1. Speech Features 1 2. Speech Features 2.
Text-To-Speech (Features)	1. Translated Text 1 and 2. Translated Text 2 and Speech Features 3. Speech Features 1 and 2 based on Input Selector	1. Translated Speech 1 3. Translated Speech 2

5.9.6 Specification of Bidirectional Speech Translation AIMs and JSON Metadata

Table 34 – AIMs and JSON Metadata

MMC-BST	Bidirectional Speech Translation	X
- MMC-ASR	Audio Scene Description	X
- MMC-TTT	Text-to-Speech	X
- MMC-ISD	Input Speech Description	X
- MMC-TTS	Text-to-Speech	X

5.10 One-to-Many Speech Translation (MMC-MST)

5.10.1 Scope of One-to-Many Speech Translation

The goal of the One-to-Many Speech Translation (MMC-MST) Use Case is to enable one person speaking his or her language to broadcast to two or more audience members, each listening and responding in a different language, presented as speech or text. If the desired output is speech, users can specify whether their speech features (voice colour, emotional charge, etc.) should be preserved in the translated speech.

The flow of control (from Recognised Text to Translated Text to Output Speech) is identical to that of the Unidirectional case. However, rather than one such flow, multiple paired flows are provided – the first pair from language A to language B and B to A; the second from A to C and C to A; and so on.

Depending on the value of Input Selector (text or speech):

1. Input Text in Language A is translated into Translated Text in and pronounced as Speech of all Requested Languages.
2. The Speech features (voice colour, emotional charge, etc.) in Language A are preserved in all Requested Languages.

5.10.2 Reference Model of One-to-Many Speech Translation

Figure 10 depicts the AIMs and the data exchanged between AIMs.

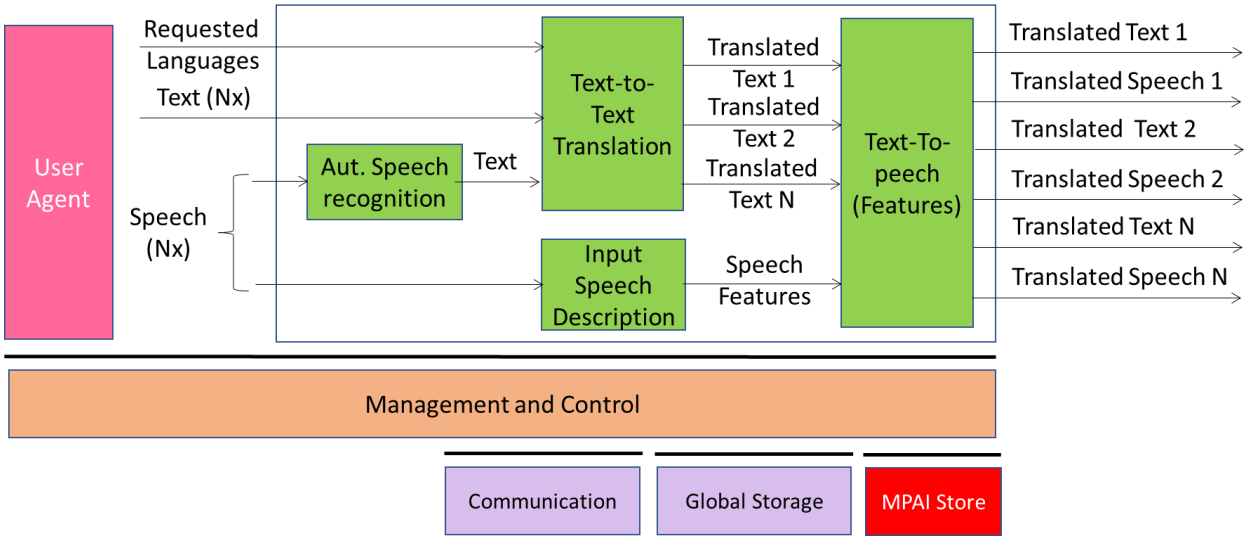

Figure 10 – Reference Model of One-to-Many Speech Translation (MST)

5.10.3 I/O Data of One-to-Many Speech Translation

The input and output data of the One-to-Many Speech Translation Use Case are:

Table 35 – I/O Data of One-to-Many Speech Translation

Input	Descriptions
Input Selector	Determines whether the input will be in Text or Speech.
Desired Languages	User-specified input language and translated languages
Input Speech	Speech produced by human desiring translation and interpretation in a specified set of languages.
Input Text	Alternative textual source information.
Output	**Descriptions**
Translated Speech	Speech translated into the Requested Languages.
Translated Text	Text translated into the Requested Languages.

5.10.4 Functions of AI Modules of One-to-Many Speech Translation

Table 36 gives the functions of One-to-Many Speech Translation AIMs.

Table 36 – Functions of One-to-Many Speech Translation AI Modules

AIM	Functions
Automatic Speech Recognition	Recognises Speech
Text-to-Text Translation	Translates Recognised Text
Input Speech Description	Extracts Speech Features
Text-To-Speech (Features)	Synthesises Translated Text adding Speech Features

5.10.5 I/O Data of AI Modules of One-to-Many Speech Translation

Table 37 gives the I/O Data of the AI Modules.

Table 37 – AI Modules of One-to-Many Speech Translation

AIM	Receives	Produces
Automatic Speech Recognition	Input Speech Segment	Recognised Text
Text-to-Text Translation	Text input	Translated Texts in the Requested Languages.
Input Speech Description	Input Speech	Speaker-specific Speech Features.
Text-To-Speech (Features)	1. Translated Texts 2. Speech Features (based on Input Selector)	Speech Segments in the Desired Languages.

5.10.6 Specification of One-to-Many Speech Translation AIMs and JSON Metadata

Table 38 – AIMs and JSON Metadata

MMC-BST	One-to-Many Speech Translation	X
- MMC-ASR	Audio Scene Description	X
- MMC-TTT	Text-to-Speech	X
- MMC-ISD	Input Speech Description	X
- MMC-TTS	Text-to-Speech	X

6 Composite AI Modules

Composite AIMs are AI Modules composed of multiple AI Modules. Several Use Cases of MPAI-MMC and other MPAI Technical Specifications use Composite AIMs. This Chapter specifies the Personal Status Extraction (PSE) Composite and Text and Speech Translation (MMC-TST) AIMs using a format aligned with the one adopted for Uses Cases.

6.1 Personal Status Extraction (MMC-PSE)

Personal Status Extraction (PSE) is a composite AIM that extracts Cognitive State, Emotion, and Social Attitude called Factors conveyed by each of Text, Speech, Face, and Gesture, called Modalities, and provides an estimate of the Personal Status, intended as a combination of Factors. The Personal Status Composite AIM is used in MPAI-MMC and other Use Cases as a replacement for the combination of AIMs depicted in Figure 11. Note that the Personal Status Data Type need not convey information on all Factors and all Modalities.

6.1.1 Scope of Personal Status Extraction

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

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

An implementation may combine two or more of the AIMs implementing the steps.

6.1.2 Reference Model of Personal Status Extraction

Figure 11 depicts the Personal Status extraction process:

1. *Descriptors are extracted* from Text, Speech, Face Object, and Body Object. An AI Module upstream can provide Descriptors, depending on the value of the Input Selectors, indicating to PSE whether a Modality or its Descriptors are used.
2. *Descriptors are interpreted* and the specific indicators of the Personal Status in the Text, Speech, Face, and Gesture Modalities are derived.
3. *Personal Status is obtained* by combining the estimates of different Modalities of the Personal Status.

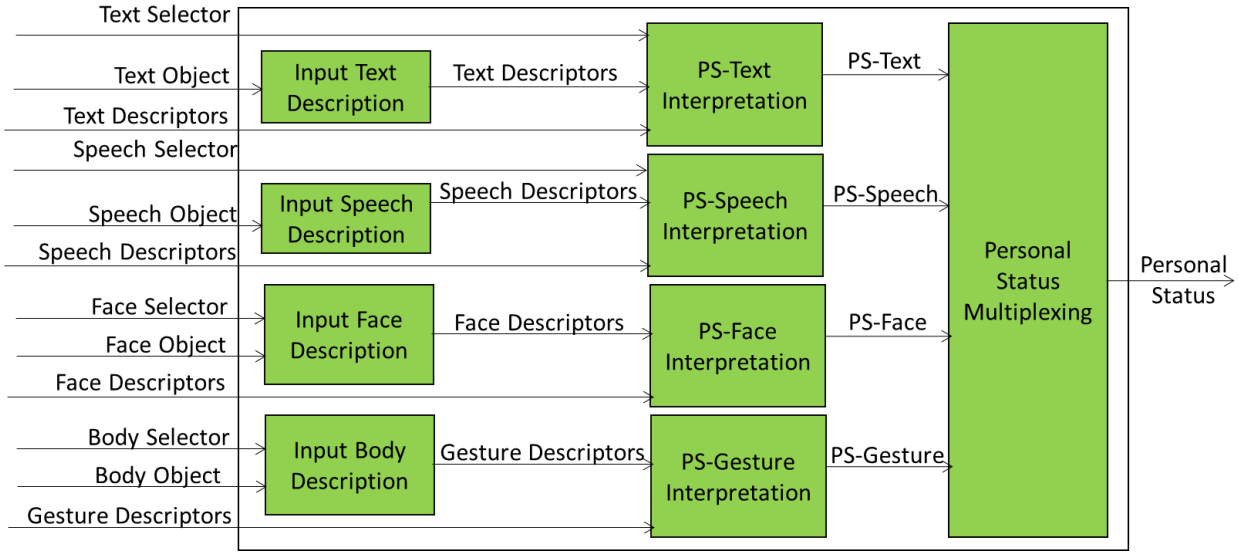

Figure 11 – Reference Model of Personal Status Extraction

Note that:

1. A Modality can be input into the Personal Status Extraction Composite AIM as a Modality or as Descriptors. Both Modality Descriptors have the same syntax and semantics. Text Descriptors are equivalent to Meaning. Gesture Description extracts Gesture Descriptors from Body Object. In the future other Descriptors may be extracted from Body Object.
2. An Implementation can combine, e.g., the Gesture Description and PS-Gesture Interpretation AIMs into one AIM, and directly provide PS-Gesture from a Body Object without exposing PS-Gesture Descriptors.

6.1.3 I/O Data of Personal Status Extraction

Table 39 gives the input/output data of Personal Status Extraction.

Table 39 – I/O data of Personal Status Extraction

Input data	From	Description
Input Text Selector	An external signal	Text/Descriptors Selector
Text	Keyboard or AIM	Text or Recognised Text
Text Descriptors	An upstream AIM	Descriptors of Text
Input Speech Selector	An external signal	Speech/Descriptors Selector
Speech	Microphone	Speech of human.
Speech Descriptors	An upstream AIM	Descriptors of Speech
Input Face Selector	An external signal	Face/Descriptors Selector
Face Object	Visual Scene Description	The face of the human
Face Descriptors	An upstream AIM	Descriptors of Face
Input Gesture Selector	An external signal	Body/Descriptors Selector
Body Object	Visual Scene Description	The body of the human
Gesture Descriptors	An upstream AIM	Descriptors of Gesture
Output data	**To**	**Descriptions**
Personal Status	A downstream AIM	For further processing

6.1.4 Functions of AI Modules of Personal Status Extraction

Table 40 gives functions of the AIMs.

Table 40 - AI Modules of Personal Status Extraction

AIM	Modules
Input Text Description	Extracts the Descriptors of Text.
Input Speech Description	Extracts the Descriptors of Speech.
Input Face Description	Extracts the Descriptors of Face.
Input Body Description	Extracts the Descriptors of Body.
PS-Text Interpretation	Interprets the Personal Status Descriptors of Text.
PS-Speech Interpretation	Interprets the Personal Status Descriptors of Speech.
PS-Face Interpretation	Interprets the Personal Status Descriptors of Face.
PS-Gesture Interpretation	Interprets the Personal Status Descriptors of Body.
Personal Status Multiplexing	Produces the Personal Status.

6.1.5 I/O Data of AI Modules of Personal Status Extraction

Table 41 gives the list of the AIMs with their functions.

Table 41 - AI Modules of Personal Status Extraction

AIM	Receives	Produces
Input Text Description	Text	Text Descriptors
Input Speech Description	Speech	Speech Descriptors
Input Face Description	Face Object	Face Descriptors
Input Body Description	Body Object	Gesture Descriptors
PS-Text Interpretation	PS-Text Descriptors	PS-Text
PS-Speech Interpretation	PS-Speech Descriptors	PS-Speech
PS-Face Interpretation	PS-Face Descriptors	PS-Face
PS-Gesture Interpretation	PS-Gesture Descriptors	PS-Gesture
Personal Status Multiplexing	PS-Text PS-Speech PS-Face PS-Gesture	Personal Status

6.1.6 AIM and JSON Metadata Specification of Personal Status Extraction

Table 42 – AIM and JSON Metadata

AIM	Name	JSON
MMC-PSE	Personal Status Extraction	X
MMC-ITD	Input Text Description	X
MMC-ISD	Input Speech Description	X
- PAF-IFD	Input Face Description	X
- PAF-IBD	Input Body Description	X
- MMC-PTI	PS-Text Interpretation	X
- MMC-PSI	PS-Speech Interpretation	X
- PAF-PFI	PS-Face Interpretation	X
- PAF-PGI	PS-Gesture Interpretation	X
- MMC-PMX	Personal Status Multiplexing	X

6.2 Text and Speech Translation (MMC-TST)

6.2.1 Functions of Speech and Text Translation

Text and Speech Translation (MMC-TST):

1. Receives:

1.1. Input Selection determining whether the input is provided as text or speech. If the desired output is speech, the user can specify whether their speech features (voice colour, emotional charge, etc.) should be preserved in the translated speech.

1.2. Target language.

1.3. Input Text.

1.4. Input Speech.

2. Produces translated text or speech segments in the target language.

6.2.2 Reference Model of Text-and-Speech Translation

Figure 12 depicts the Reference Model of the Text-and-Speech Translation Composite AIM.

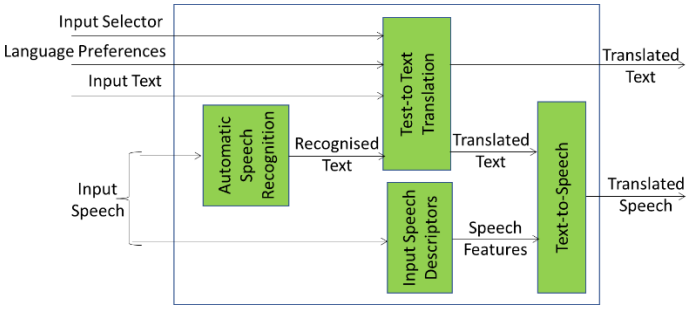

Figure 12 – The Text-and-Speech Translation Composite AIM

6.2.3 I/O Data of Text-and-Speech Translation

Table 43 provides the list of the I/O Data of the Text-and-Speech Translation Composite AIM.

Table 43 – I/O Data of Text-and-Speech Translation

Input	Semantics
Input Selection	Determines whether: 3. The input will be in Text or Speech 4. The Input Speech features are preserved in the Output Speech.
Language Preferences	User-specified input Language (A) and output Language (B).
Input Speech	Speech produced in Language A by a human desiring translation into language B.
Input Text	Alternative textual source information to be translated into and pronounced in language B depending on the value of Input Selection.
Output	**Description**
Translated Speech	Input Speech in language A translated into language B preserving the Input Speech features in the Output Speech, depending on the value of Input Selection.
Translated Text	Text of Input Speech or Input Text translated into language B, depending on the value of Input Selection.

6.2.4 I/O Data of AI Modules of Text-and-Speech Translation

Table 28 gives the functions of Text-and-Speech Translation AIMs.

Table 44 – Functions of Text-and-Speech Translation AI Modules

AIM	Functions
Automatic Speech Recognition	Recognises Input Speech.
Text-to-Text Translation	Translates Recognised Text into Translated Text.
Input Speech Descriptors	Extracts Speech Descriptors (a.k.a. Features) from Input Speech.
Text-to-Speech (Features)	Synthesises Translated Text adding Speech Features

6.2.5 I/O Data of AI Modules of Text-and-Speech Translation

The AI Modules of Text-and-Speech Translation are given in *Table 29.*

Table 45 – AI Modules of Text-and-Speech Translation

AIM	Receives	Produces
Automatic Speech Recognition	Input Speech Segment	Recognised Text
Text-to-Text Translation	1. Input Text 2. Recognised Text (Based on Input Selection)	Translated Text
Input Speech Descriptors	Input Speech	Speaker-specific Speech Features.
Text-to-Speech (Features)	1. Translated Text 2. Speech Features (Based on Input Selection)	Produces Output Speech.

6.2.6 Specification of Speech-and-Text Translation AIMs and JSON Metadata

Table 46 – AIMs and JSON Metadata

AIM	Name	JSON
MMC-TST	Text and Speech Translation	X
- MMC-ASR	Automatic Speech Recognition	X
- MMC-TTT	Text-to-Text Translation	X
- MMC-ISD	Input Speech Description	X
- MMC-TTS	Text-to-Speech	X

7 Data Types

This Technical Specification specifies the Data Types listed in *Table 47*. The reader is alerted that some data Formats are shared with the Context-based Audio Enhancement (MPAI-CAE) Standard [3]. At the current date, the specification of such data Formats is repeated verbatim in both Standards.

The first column gives the name of the data Format, the second the subsection where the data Format is specified and the third the Use Case(s) making use of it.

Table 47 – Data formats

Name of Data Format	Subsection	Use Case
Media		
Audio File	7.1.1	ABV
		BST
		CAS
		CWE
		HCI
		MST
		UST
		VSV
Text	7.1.2	BST
		CWE
		MQA
		MST
		UST
Video	7.1.3	CWE
Video File	7.1.4	ARP
Descriptors		
Audio Scene Descriptors	7.2.1	ABV
		HCI
Face Descriptors	7.2.2	ABV
		CWE
		HCI
		VSV
Gesture Descriptors	7.2.3	ABV
		CWE
		HCI
		VSV
Speech Descriptors	7.2.4	ABV
		CWE
		HCI
		VSV
Speech Features	7.2.5	UST
Text Descriptors	7.2.6	ABV
		CWE
		HCI
		VSV
Visual Scene Descriptors	7.2.7	ABV
		CAS
		HCI
Personal Status		
Personal Status	7.3.2	ABV
		CAS
		HCI
Cognitive State	7.3.2.2	CAS
		HCI
		VSV
Emotion	7.3.4	ABV

		CWE
		HCI
		VSV
Social Attitude	7.3.5	CAS
		HCI
Object and Scenes		
Spatial Attitude and Point of View	7.4.1	CAS
		HCI
Audio Object and Scene	7.4.2	HCI
Visual Object and Scene	7.4.3	CAS
		HCI
Miscellanea		
Instance Identifier	7.5.1	ABV
		CPS
		HCI
		MQA
Intention	7.5.2	MQA
Language Identifier	7.5.3	BST
		MST
		UST
Meaning	7.5.4	CAS
		CWE
		HCI
Query Format for Video of Faces KB	7.5.5	CWE

MPAI plans on creating a future specification that will contain all data Formats that are shared by more than one MPAI Technical Specification.

7.1 Media

7.1.1 Audio File

Audio data is packaged in a .wav file [10].

7.1.2 Text

The Format of Input Text, Output Text and Recognised Text is provided by ISO/IEC 10646 [9].

7.1.3 Video

Video satisfies the following specifications:

1. Pixel shape: square
2. Bit depth: 8 or 10 bits/pixel
3. Aspect ratio: 4/3 or 16/9
4. 640 < # of horizontal pixels < 1920
5. 480 < # of vertical pixels < 1080
6. Frame frequency 50-120 Hz
7. Scanning: progressive
8. Colorimetry: ITU-R BT709 or BT2020
9. Colour format: RGB or YUV
10. Compression:
    10.1. Uncompressed.

10.2. If compressed, compression according to one of the following standards: MPEG-4 AVC [10], MPEG-H HEVC [13], MPEG-5 EVC [14].

7.1.4 Video File

The Format of a Video MP4 File Format [12].

7.2 Descriptors

7.2.1 Audio Scene Descriptors

Audio Scene Descriptors are specified in MPAI-CAE V2 [3].

7.2.2 Face Descriptors

Face Descriptors used by Personal Status Extraction are specified in MPAI-PAF V1 [3].

7.2.3 Gesture Descriptors

Gesture Descriptors used by Personal Status Extraction are specified in MPAI-PASF V1 [3].

7.2.4 Speech Descriptors

Speech Descriptors act as Speech Features defined in Personal Status Extraction.

7.2.5 Speech Features

Speech Features are digitally represented as follows.

7.2.5.1 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"
      }
   }
}

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

7.2.5.2 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 as in *Table 48*.
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 (decibel).
speed	Indicates the Speech Rate as a real number indicating specified linguistic 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.

Name	*Definition*
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 7.3.

Table 48 – Basic Tones

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

7.2.6 Text Descriptors

Meaning acts as Text Descriptors defined in Personal Status Extraction.

7.2.7 Visual Scene Descriptors

Visual Scene Descriptors are specified in MPAI-OSD [5].

7.3 Personal Status

7.3.1 Factors and Modalities

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").

All these Factors can be expressed via several Personal Status *Modalities*: Text, Speech, Face, and Gestures. (Other Modalities, such as body posture, may be handled in future MPAI Versions.) 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 Emotion, Cognitive State, and Social Attitude Factors 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 and 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 indicated the subversion.

7.3.2 Personal Status Data

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 49)
2.1.1. 6 cells for Emotion.
2.1.2. 6 cells for Cognitive State.
2.1.3. 6 cells for Social Attitude.

Table 49 - The table of (Factor, Modality) cells

		Modality					
		Version	Fused value	Text	Speech	Face	Gesture
F	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, or error, or no discernible value.
3.4.2. Value 1 up to the largest number of Factor values in the relevant Label Semantics Table.

7.3.2.1 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"]
}
```

7.3.2.2 Semantics

An instance of Personal Status is represented by the following table. Timestamp, Emotion, Cognitive State, Social Attitude, and their Descriptors are present if the corresponding information is available.

Table 50 – The variables composing 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

7.3.3 Cognitive State

Cognitive State is represented by the following Syntax and Semantics. Primary Cognitive State corresponds to General Adjectival and Secondary Cognitive State corresponds to Specific Adjectival in *Table 51*.

The Syntax and Semantics of Cognitive State are given by the following clauses.

7.3.3.1 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"
         }
      }
   }
}
```

7.3.3.2 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 54*.
`cogstateSetName`	Specifies the name of the Cognitive State set which contains the Cognitive State. Cognitive State set of *Table 54* is used as a baseline, but other sets are possible.

Table 51 gives the standardised three-level Basic Cognitive State Label Set.

Table 51 – 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 52 provides the semantics for each label in the GENERAL ADJECTIVAL and SPECIFIC ADJECTIVAL columns above.

Table 52 – 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

7.3.4 Emotion

The Syntax and Semantics of Emotion are given by the following clauses. Emotions are expressed vocally through combinations of prosody (pitch, rhythm, and volume variations); separable speech effects (such as degrees of voice tension, breathiness, etc.); and vocal gestures (laughs, sobs, etc.). Emotion is represented by the following Syntax and Semantics. Primary Emotion corresponds to General Adjectival and Secondary Emotion corresponds to Specific Adjectival in *Table 53*.

7.3.4.1 Syntax

Emotion is represented by:

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

7.3.4.2 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 54*.
`emotionSetName`	Specifies the name of the Emotion set which contains the Emotion. Emotion set of *Table 54* is used as a baseline, but other sets are possible.

Table 53 gives the standardised three-level Basic Emotion Set partly based on Paul Eckman [19].

Table 53 – 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 54 provides the semantics for each label in the GENERAL ADJECTIVAL and SPECIFIC ADJECTIVAL columns above.

Table 54 – 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

7.3.5 Social Attitude

Social Attitude is represented by the following Syntax and Semantics. Primary Social Attitude corresponds to General Adjectival and Secondary Social Attitude corresponds to Specific Adjectival in *Table 55.*

7.3.5.1 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"
        }
     }
  }
}
```

7.3.5.2 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 56*.
`attitudeSetName`	Specifies the name of the Social Attitude set which contains the Social Attitude. Social Attitude set of *Table 56* is used as a baseline, but other sets are possible.

Table 55 gives the standardised three-level Basic Social Attitude Set.

Table 55 – 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	flexible

	uncooperative	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	suggestive/risqué/naughty

	lewd/bawdy/indecent prudish/priggish	
SOCIAL RANK	polite/courteous/respectful rude/disrespectful commanding/domineering pompous/pretentious obedient rebellious/defiant	condescending/patroniz-ing/snobbish pedantic unaffected servile/obsequious

Table 56 provides the semantics for each label in the GENERAL ADJECTIVAL and SPECIFIC ADJECTIVAL columns above.

Table 56 – 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 disillu-sionment
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/trustwor-thy/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-effac-ing	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 per-ceived 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

7.4 Objects and Scenes

7.4.1 Spatial Attitude and Point of View

Specified by Object and Scene Description [6].

7.4.2 Audio Objects and Scene

Specified by Object and Scene Description [6].

7.4.3 Visual Objects and Scene

Specified by Object and Scene Description [6].

7.5 Miscellanea

7.5.1 Instance Identifier

Instance is an element of a set of entities – Visual Objects, users etc. – belonging to some levels in a hierarchical classification (taxonomy).
The syntax and semantics of Instance Identifier are specified below.

7.5.1.1 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"
  ]
}
```

7.5.1.2 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.

7.5.2 Intention

Intention,is the result of Question Analysis AIM. It consists of the following elements:

- qtopic
- qfocus
- qLAT
- qSAT

7.5.2.1 Syntax

```
{
  "$schema": "http://json-schema.org/draft-07/schema",
  "definitions": {
    "Intention": {
      "type": "object",
      "properties": {
        "qtopic": {
          "type": "string"
        },
        "qfocus": {
          "type": "string"
        },
        "qLAT": {
          "type": "string"
        },
        "qSAT": {
          "type": "string"
        },
        "qdo-main": {
          "type": "string"
        }
      }
    }
  },
  "type": "object",
  "properties": {
    "primary": {
      "$ref": "#/definitions/intention"
    },
    "second-ary": {
      "$ref": "#/definitions/intention"
    }
  }
}
```

7.5.2.2 Semantics

Name	*Definition*
Intention	Provides abstracts of Intention of User Question using properties: qtopic, qfocus, qLAT, qSAT and qdomain
qtopic	Indicates the topic of the question. Question topic is the object or event that the question is about. Ex. of Qtopic is King Lear in "Who is the author of King Lear?".
qfocus	Indicates the focus of the question, which is the part of the question that, if replaced by the answer, makes the question a stand-alone statement. Ex. What, where, who, what policy. Which river, etc. Example. Question: Who is the president of USA? (The word "Who" is the focus of the question and it will be replaced by "Biden" in the Answer.) Answer: Biden is the president of USA.
qLAT	Indicates the lexical answer type of the question.
qSAT	Indicates the semantic answer type of the question. QSAT corresponds to Named Entity type of the language analysis results.
qdomain	Indicates the domain of the question such as "science", "weather", "history". Ex. Who is the third king of Yi dynasty in Korea? (qdomain: history)

The following example shows the question analysis result of the user's question, "Who is the author of King Lear?" The question analysis result in the example shows that the domain of the question is "Literature," the topic of the question is "King Lear", and the focus of the question is "Who."

```
{
   "intention":[
      {
         "qdomain":"Literature",
         "qtopic":"King Lear ",
         "qfocus":"who ",
         "qLAT":"author ",
         "qSAT":"person "
      }
   ]
}
```

The following example shows the result of the analysed question of "How do you make Kimchi?" The question analysis result in the example shows that the domain of the question is "Cooking", the topic of the question is "Kimchi", the focus of the question is "how".

```
{
   "intention":[
      {
         "qdomain":"Cooking",
         "qtopic":"Kimchi",
         "qfocus":"How ",
         "qLAT":"cooking method ",
         "qSAT":"method "
      }
   ]
}
```

7.5.3 Language Preference

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

7.5.4 Meaning

This subclause specifies data formats to describe meaning which is the result of the analysis of text expressing natural language. The “meaning” consists of the following elements:

- POS_tagging
- NE_tagging
- Dependency_tagging
- SRL_tagging

7.5.4.1 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"
         }
      }
   }
}
```

7.5.4.2 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.

7.5.5 Query Format of Video of Faces KB

Data Specification: All faces in the Video of Faces KB shall be aligned.
Input: Emotion.
Output: Video File of a human face.

Annex 1 - MPAI Basics (Informative)

1 General

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 and are expected to do so much more in the future. As digital media standards have positively influenced industry and billions of people, so AI-based data coding standards are expected to have a similar positive impact. In addition, some AI technologies may carry inherent risks, e.g., in terms of bias toward some classes of users or application domains making the need for standardisation more important and urgent than ever.

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

2 Governance of the MPAI Ecosystem

The technical foundations of the MPAI Ecosystem are currently provided by the Governance of the MPAI Ecosystem [1] developed and maintained by MPAI:

1. Technical Specification.
2. Reference Software Specification.
3. Conformance Testing Specification.
4. Performance Assessment Specification.
5. Technical Report

An MPAI Standard is a collection of a variable number of the 5 document types.

Figure 13 depicts the MPAI ecosystem operation for conforming MPAI implementations.

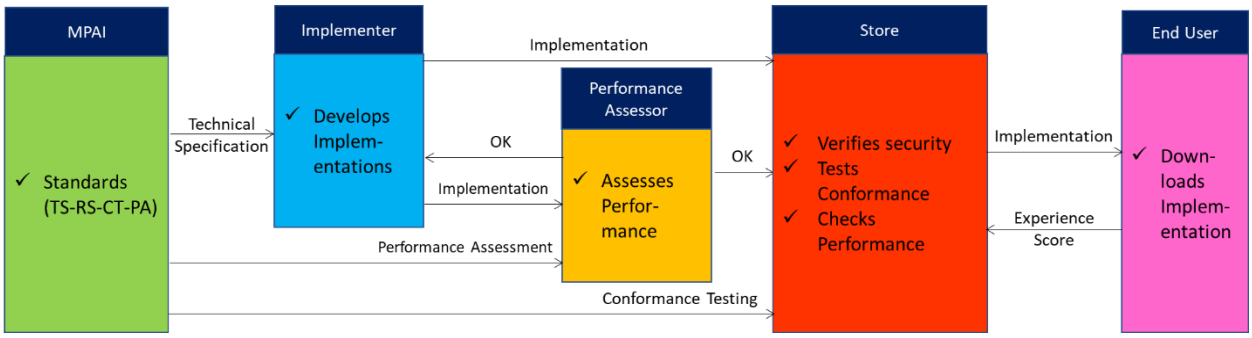

Figure 13 – The MPAI ecosystem operation

Technical Specification: Governance of the MPAI Ecosystem [1] identifies the following roles in the MPAI Ecosystem:

Table 57 - Roles in the MPAI Ecosystem

MPAI	Publishes Standards. Establishes the not-for-profit MPAI Store. Appoints Performance Assessors.
Implementers	Submit Implementations to Performance Assessors.

	Submit Implementations to the MPAI Store.
Performance Assessors	Inform Implementation submitters and the MPAI Store if Implementation Performance is acceptable.
MPAI Store	Assign unique ImplementerIDs (IID) to Implementers in its capacity as ImplementerID Registration Authority (IIDRA)[1]. Verifies security and Tests Conformance of Implementations.
Users	Download Implementations and report their experience to the MPAI Store.

3 AI Framework

MPAI develops standards in compliance with a rigorous process [16] pursuing the following policies:

5. Be friendly to the AI context but, to the extent possible, agnostic to the technology – AI or Data Processing – used in an implementation.
6. Be attractive to different industries, end users, and regulators.
7. Address three levels of standardisation any of which an implementer can freely decide to adopt:
    a. Data types, i.e., the data exchanged by systems.
    b. Components called AI Modules (AIM).
    c. Connected components called AI Workflows (AIW).
8. Specify the data exchanged by components with a clear semantic to the extent possible.

Technical Specification: AI Framework (MPAI-AIF) V2 enables dynamic configuration, initialisation, and control of AIWs in a standard environment called AI Framework (AIF). *Figure 14* depicts the AI Framework.

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

Thus, users can exercise AIWs that are both proprietary or standardised by MPAI – i.e., with standard functions and interfaces, 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.

[1] At the time of publication of this Technical Report, the MPAI Store was assigned as the IIDRA.

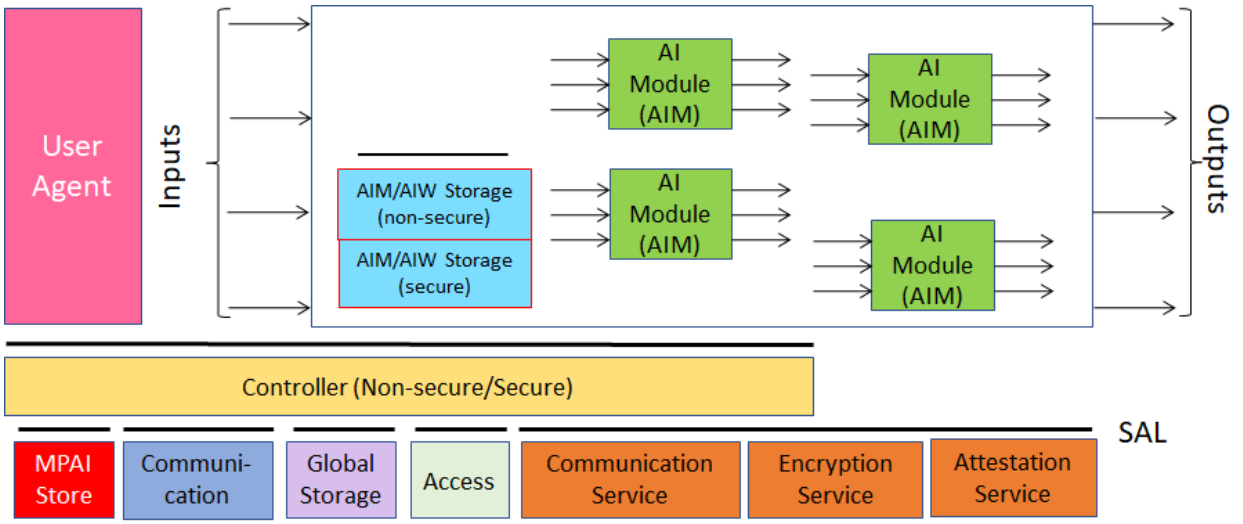

Figure 14 - The AI Framework (MPAI-AIF) V2 Reference Model

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 are silent on 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.

AIW and its AIMs may have 3 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.

4 Audio-Visual Scene Description

The ability to describe (i.e., digitally represent) an audio-visual scene is a key requirement of several MPAI Technical Specifications and Use Cases. MPAI has developed Technical Specification: Context-based Audio Enhancement (MPAI-CAE) [4] that includes Audio Scene Descriptors and uses a subset of Graphics Language Transmission Format (glTF) [7] to describe a visual scene.

Audio Scene Description is a Composite AI Module (AIM) specified by Technical Specification: Context-based Audio Enhancement (MPAI-CAE) [4]. The position of an Audio Object is defined by Azimuth, Elevation, Distance.

The Composite AIM and its composing AIMs are depicted in Figure 20.

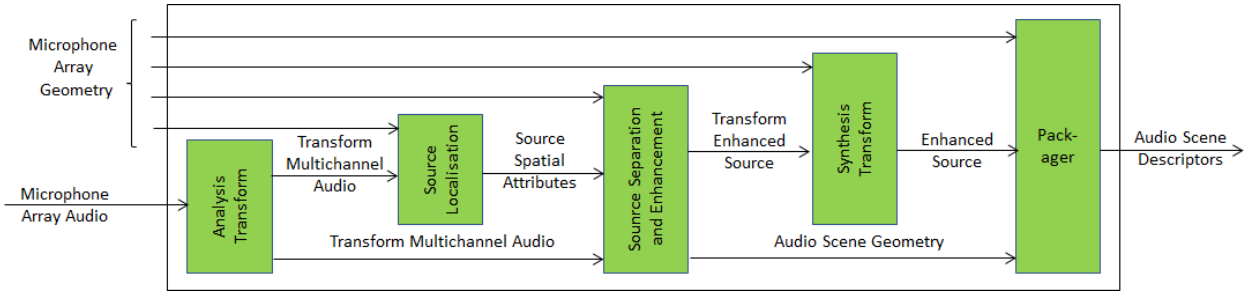

Figure 15 - The Audio Scene Description Composite AIM

4.1 Visual Scene Descriptors

MPAI uses a subset of Graphics Language Transmission Format (glTF) [7] to describe a visual scene.

5 Avatar-Based Videoconference

Technical Report: Avatar-Based Videoconference (MPAI-ARA) specifies AIWs and AIMs of a Use Case where geographically distributed humans hold a videoconference represented by their avatars. Figure 16 depicts the components of the system supporting the conference of a group of humans participating through avatars having their visual appearance and uttering the participants' real voice.

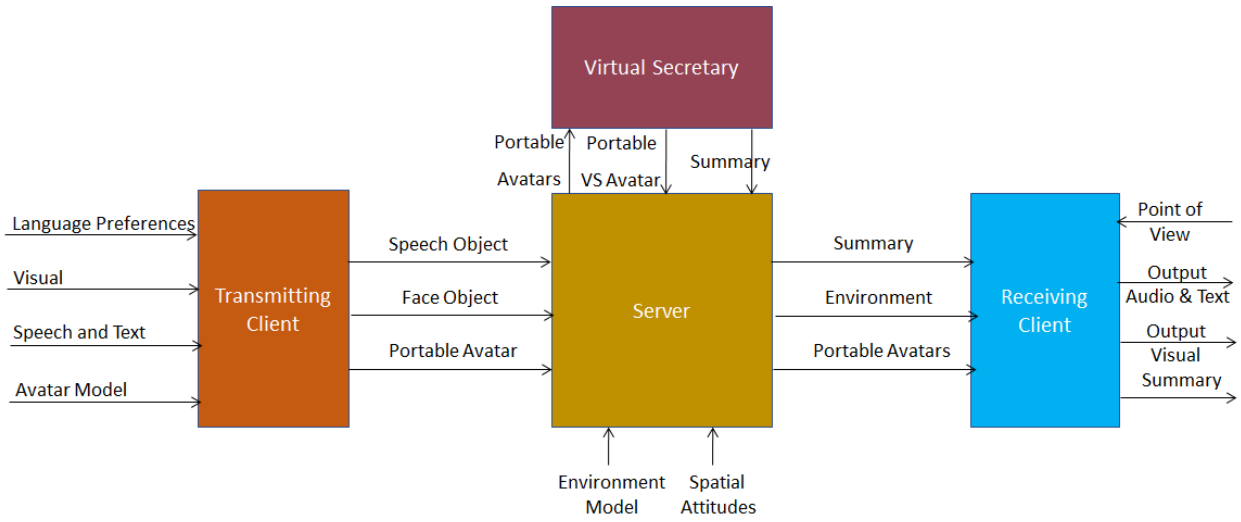

Figure 16 – Avatar-Based Videoconference end-to-end diagram

Figure 17 contains the Reference Models of the four AW Workflows constituting the Avatar-Based Videoconference: Client (Transmission side), Server, Virtual Secretary, and Client (Receiving side).

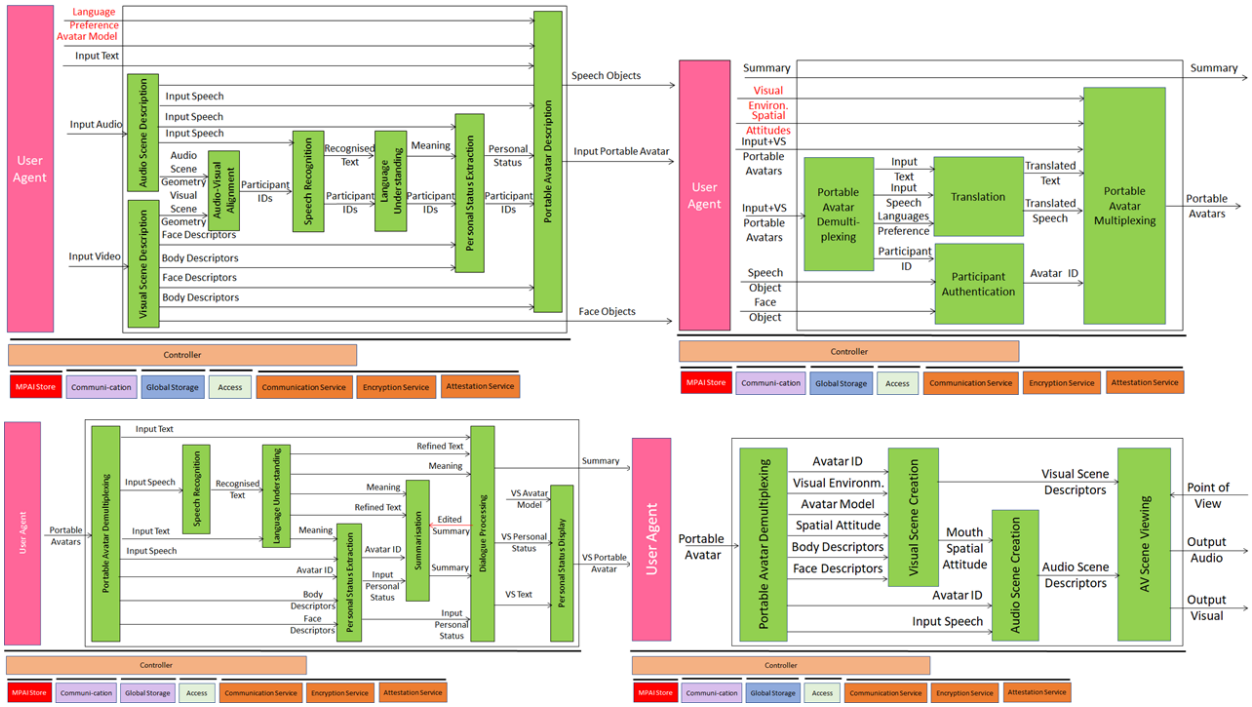

Figure 17 - The AIWs of Avatar-Based Videoconference

6 Connected Autonomous Vehicle

MPAI defines a Connected Autonomous Vehicle (CAV), as a physical system that:

1. Converses with humans by understanding their utterances, e.g., a request to be taken to a destination.
2. Acquires information with a variety of sensors on the physical environment where it is located or traverses like the one depicted in *Figure 18.*
3. Plans a Route enabling the CAV to reach the requested destination.
4. Autonomously reaches the destination by:

4.1. Moving in the physical environment.
4.2. Building Digital Representations of the Environment.
4.3. Exchanging elements of such Representations with other CAVs and CAV-aware entities.
4.4. Making decisions about how to execute the Route.
4.5. Acting on the CAV motion actuation to implement the decisions.

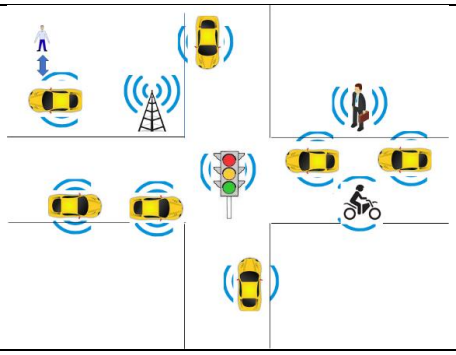

Figure 18 - An environment of CAV operation

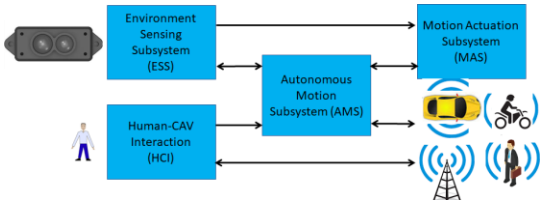

Figure 19 – The MPAI-CAV subsystems

MPAI believes in the capability of standards to accelerate the creation of a global competitive CAV market and has published Technical Specification: Connected Autonomous Vehicle (MPAI-CAV) – Architecture that includes (see *Figure 19*):

1. A CAV Reference Model broken down into four Subsystems.

2. The Functions of each Subsystem.
3. The Data exchanged between Subsystems.
4. A breakdown of each Subsystem in Components of which the following is specified:

4.1. The Functions of the Components.
4.2. The Data exchanged between Components.
4.3. The Topology of Components and their Connections.

5. Subsequently, Functional Requirements of the Data exchanged.
6. Eventually, standard technologies for the Data exchanged.

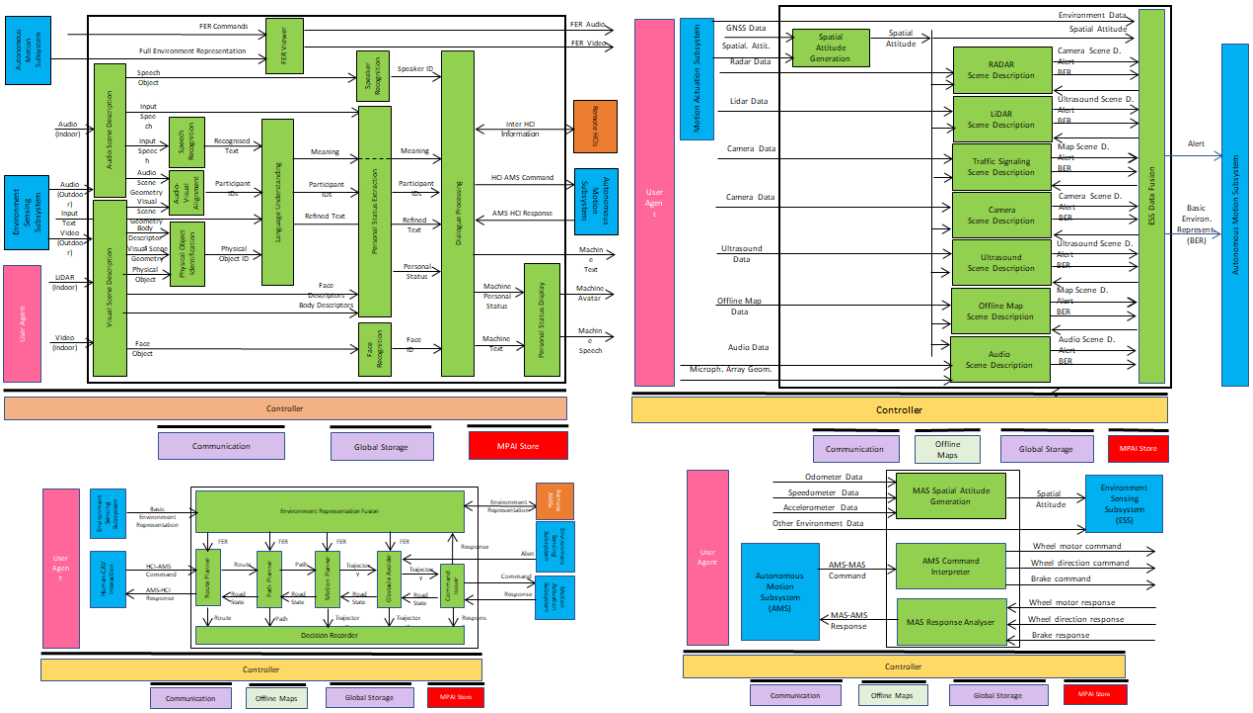

Figure 20 - The MPAI-CAV Subsystems with their Components

Subsystems are implemented as AI Workflows and Components as AI Modules according to Technical Specification: AI Framework (MPAI-AIF) [2].

Annex 2 - MPAI-wide terms and definitions

The Terms used in this standard whose first letter is capital and are not already included in *Table 1* are defined in Table 58. To concentrate in one place all the Terms that are composed of a common name followed by other words (e.g., the word Data followed by one of the words Format, Type, or Semantics), the definition given to a Terms preceded by a dash "-" applies to a Term composed by that Term without the dash preceded by the Term that precedes it in the column without a dash.

Table 58 - 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.
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 Means	Procedures, tools, data sets and/or data set characteristics to Test the Conformance of an Implementation.
Connection	A channel connecting an output port of an AIM and an input port of an AIM.
Controller	A Component that manages and controls the AIMs in the AIF, so that they execute in the correct order and at the time when they are needed
Data	Information in digital form.
- 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.
Testing Laboratory	A laboratory accredited to Assess the Grade of Performance of Implementations.
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 3 - 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.

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

Coverage and Applicability of MPAI Standards

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

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

IN NO EVENT SHALL MPAI BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: THE NEED TO PROCURE SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF

THE PUBLICATION, USE OF, OR RELIANCE UPON ANY STANDARD, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

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

Users of MPAI Standards should consider all applicable laws and regulations when using an MPAI Standard. The validity of Conformance Testing is strictly technical and refers to the correct implementation of the MPAI Standard. Moreover, positive Performance Assessment of an implementation applies exclusively in the context of the MPAI Governance 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 4 - Patent declarations (Informative)

The MPAI Multimodal Conversation (MPAI-MMC) Technical Specification has been developed according to the process outlined in the MPAI Statutes [15] and the MPAI Patent Policy [16].

The following entities have agreed to licence their standard essential patents reading on the MPAI Multimodal Conversation (MPAI-MMC) Technical Specification according to the MPAI-MMC Framework Licence [17]:

Table 59 - Companies having submitted a patent declaration (MPAI-MMC V1)

Entity	Name	Email address
ETRI	Songwon Lee	lsw84@etri.re.k
KLleon	Jisu Kang	jisu.kang@klleon.io
Speech Morphing, Inc.	Fathy Yassa	fathy@speechmorphing.com

Patents declarations of Table 59 concern Version 1. Declarations for Version 2 will be published when patent declarations in response to requests for declarations will be received.

Annex 5 - Personal Status (Informative)

The study of "personal status" – of emotion, cognitive states, attitudes, and other status factors that a person can express at a given time – is not new: many aspects have long been studied. Now, however, technological, and scientific advances promise accelerating understanding. MPAI's aim is to establish standards in various current and future use cases involving Personal Status – for instance, to enable computational systems to recognize users' emotions and react to them most helpfully. Thus, the need arises to at least roughly characterize and survey Emotions, Cognitive States, and Attitudes.

To begin meeting this need, this document proposes *definitions*, *listings*, and *semantic characterizations* of these three factors. These proposals are indeed rough and subject to disagreement or revision on many levels. Accordingly, they can in fact be revised for particular use cases and as the relevant studies move ahead. Revision procedures are specified in the Conclusion below.

This Annex offers definitions and examples of each status factor, with brief discussion. Listings of labels and accompanying semantics per factor are given in Section 4.2.

Emotions are states of physiological arousal accompanied by changes in facial expressions, gestures, posture, or subjective feelings. Examples include joy, sadness, disgust, fear, and anger. Innate elements of emotions – there may be learned components as well – are controlled by the subcortical regions of the brain, including the amygdala, ventral striatum, and hypothalamus.

Sensations like pain, pleasure, taste, vision, hearing, and so on are likewise largely innate, but we'll try to distinguish them from Emotions as such. Unlike Emotions, sensations will not be defined or listed here.

Cognitive states are the results of information processing: a cognitive system accepts input patterns – in humans, initially perceptual patterns, whether new or stored – and produces output patterns, which may include actions that can affect the world outside the system. To perform this processing, the system must recognize the input patterns, perhaps influenced by priming ("expectations"), and then associate them with other patterns, often in a sequence of steps or flow, until the output pattern is reached. The recognition, associations, and sequencing giving rise to *Cognitive States* may sometimes be innate; but in humans, they're predominantly learned.

This high-level definition of cognition and Cognitive States could describe not only human or other biological information processing, but artificial processing as well – such as that carried out by self-driving vehicles, which must recognize other vehicles, signs and signals, etc., based on patterns conveyed by sensors, and, through processing, derive appropriate action patterns. Clearly, then, the definition is meant to exclude emotion, since the vehicles have none, and in fact probably lack sensations ("qualia") of any sort, much less consciousness. In humans, however, the separation between emotion and cognition is much harder to make cleanly, since much information processing is at least partly driven by drives which are associated with emotions. Even so, it's helpful to maintain the separation for analytical purposes; so this Annex will treat Cognitive States as those information processing states which even a system lacking emotions might be able to enter – the processing states that Star Trek's "purely logical" Mr. Spock might be found in.
However, while observing the distinction between Emotions and Cognitive States as an analytical aid, we certainly recognize (1) borderline cases (like Curiosity, which does involve a *drive* to obtain new information, but might still be modelled by a system which pursued that goal in

numerical terms without emotion, as Mr. Spock might do) and (2) hybrid or overlapping states in which both cognitive processing and emotion play parts (like Positive or Negative Surprise, in which a human is both surprised – as even Mr. Spock might be – but also emotionally pleased or displeased by the unexpected event or discovery).

Since we're defining and listing Emotions and Cognitive States for the limited purposes of near-term human-machine interaction, we'll avoid a wide range of human emotional and cognitive concerns. Again, we're bypassing discussion of *sensation* or *consciousness*. Likewise, we'll avoid concern with the *emotional factors in human decision-making* (related to issues of bias and free will); with *abnormal psychology* (related to psychosis, obsessive-compulsive disorder, amnesia, etc.); or with many more psychological areas.

So, for example, while we will currently be *interested* (clearly a Cognitive State, though also viewable as borderline, hybrid, or both) in the following states, among others:

1. *Interest: determination that certain percepts are relevant to goals*
2. *Curiosity: bias toward seeking or attending to new percepts or information*
3. *Confusion: disorderly information processing*
4. *Certainty: conclusion that percepts or processing results are reliable (e.g., as basis for action)*
5. *Attention: bias to process some percepts and not others; bias to direct processing through a certain sequence and not others*

… we will for now avoid discussion of states like these:

1. *Amnesia: loss of long-term memory*
2. *Psychosis: a cognitive disorder in which mental percepts are sometimes confused with objectively real ones*
3. *Priming: cognitive bias to recognize or process percepts in a certain way*
4. *Consciousness: reportable awareness, augmented by self-concept, self-history, awareness of being aware, etc.*
5. *Subconscious processing: information processing without awareness or consciousness*

A person's attitudes are ways of *relating to* exterior elements – most often, to other humans, but also to situations, facts, etc. They're ways of *feeling or thinking about* those elements, and/or ways of *behaving toward them*, prompted by those Emotions and Cognitive States.

For MPAI's purposes, *Attitudes* are of interest for analysis of *relations* within use cases between people, and/or between people and computational systems. How can a machine communicate a *helpful* Attitude – the hybrid combination of Emotion and Cognitive State that constitutes a desire to be useful? How can a machine recognize a *resentful* Attitude – perhaps arising from a user's anger (Emotion) at her belief (Cognitive State) that she has been treated unfairly in a transaction?

The prompting or engendering of Attitudes by relevant Emotions and Cognitive States can be depicted in various ways, as in the Figures 1 and 2 below; but, whatever the graphic description, for the purposes of MPAI's standardization efforts, the focus will remain on the *relational* aspect of Attitudes, and especially on social relations.

Given that Emotions and Cognitive States themselves are difficult to describe precisely, we can't expect definitive listings or semantic characterizations of the Attitudes that arise from them.

Even so, we hope that those in Section 4.2 can prove useful in facilitating coordination among modules.

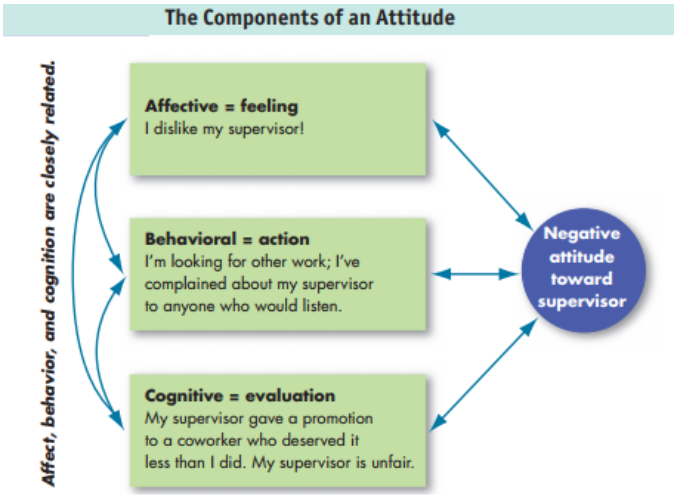

Figure 21 - Components of Attitude

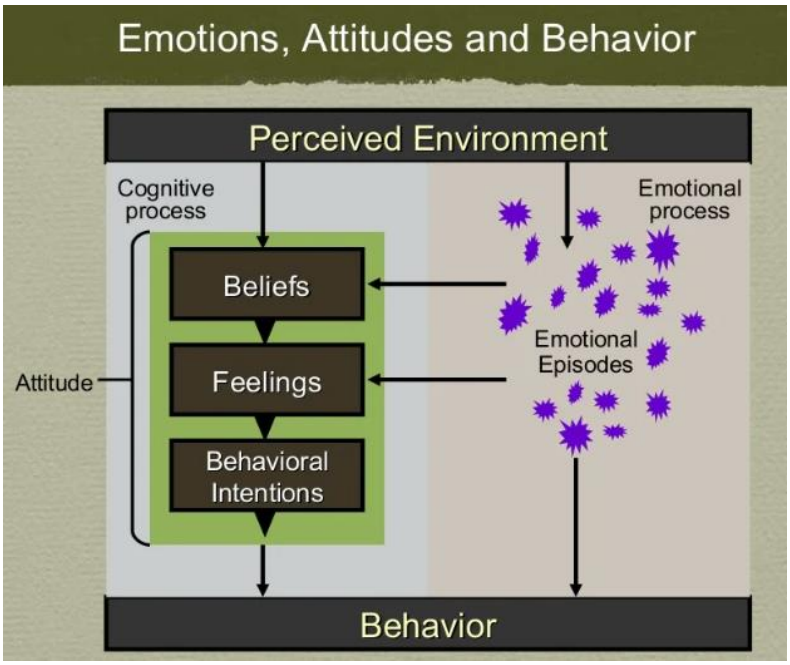

Figure 22 - Process of the Behaviour from the Emotion and Attitudes

Annex 6 - Communication Among AIM Implementors (Informative)

To the extent possible, AIM input and output data are specified so that the inner implementation of an AIM *need not* be known or considered by AIMs cooperating in an AIW or in a Composite AIM. In other words, so far as possible, cooperating AIMs are designed to interact as black boxes. However, AIMs based upon the neural network technology currently prevalent in AI systems will generally require closer cooperation – in effect, greater transparency. An AIM receiving neural input in the form of features (vectors) will require some assistance in processing them. The downstream AIM will need either:
(1) The neural network model used to train the upstream AIM, or
(2) A precise specification of the syntax and semantics of the features,
so that the downstream AIM can handle the features received from the upstream AIM.

A core design principle of MPAI is modularity: AI Modules or AIMs and their interfaces must be defined such that each AIM can be built by an independent implementor, without damage to the function of a use case as a whole.

However, MPAI also recognises that AIMs and their implementors may sometimes profit from communication and interchange of data and/or components. Such exchanges can be especially appropriate for AIMs featuring neural network components or comparable elements for machine learning – an increasingly common and important situation in the design of cooperative artificial intelligence modules.

The Unidirectional Speech Translation workflow provides a good example. It is designed to enable addition to the Translated Speech (that is, to the target language or output speech) of Speech Features extracted from the input, or source language, speech. This addition can enable the spoken translation to express the original emotion, or to employ the original speaker's voice quality to give the impression that he or she is pronouncing the translation. For these purposes, a Speech Feature Extraction AIM can extract relevant speech features from the input speech and pass them to the Text-To-Speech (Features) AIM. However, while the two AIMs can indeed be independently implemented, the downstream (receiving) AIM, in this case Text-To-Speech (Features), will need to process the received speech features appropriately. If Speech Feature Extraction employs neural network technology and passes the resulting features as vectors, then Text-To-Speech (Features) will need cooperation from Speech Feature Extraction. The downstream AIM will need either (1) the neural network model used to train the upstream AIM, or (2) a precise specification of the syntax and semantics of the features, so that the downstream AIM can handle the features received from the upstream AIM.

There are comparable considerations for the Conversation with Emotion (CWE) use case. And, more generally, they will obtain for any AIMs that exchange neural information. In explicitly providing for such communication among artificial machine learning models and components, MPAI is not only recognising practical requirements for cooperation among such modules, but also acknowledging an analogy with communication among biological neural subsystems.