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

|  |
| --- |
|  |
| **N337** | 2021/08/25 |
| **Source** | Leonardo Chiariglione |
| **Title** | Initial Draft of MMC-HCI Use Case and Functional Requirements |
| **Target** | MPAI Community |

[Abstract 1](#_Toc80356612)

[1 Introduction 2](#_Toc80356613)

[2 The MPAI approach to AI standardisation 3](#_Toc80356614)

[3 Terms and definitions 5](#_Toc80356615)

[4 Description of Human-CAV interaction 6](#_Toc80356616)

[4.1 Reference architecture 6](#_Toc80356617)

[4.2 Input and output data 7](#_Toc80356618)

[4.3 AI Modules 7](#_Toc80356619)

[5 Technologies and Functional Requirements 8](#_Toc80356620)

[5.1 I/O Data summary 9](#_Toc80356621)

[5.2 Data format requirements 9](#_Toc80356622)

[5.2.1 Text 9](#_Toc80356623)

[5.2.2 Audio 9](#_Toc80356624)

[5.2.3 Video 10](#_Toc80356625)

[5.2.4 Object identifier 10](#_Toc80356626)

[5.2.5 Emotion 10](#_Toc80356627)

[5.2.6 Meaning 10](#_Toc80356628)

[5.2.7 Intention 10](#_Toc80356629)

[5.2.8 Reply (speech) 10](#_Toc80356630)

[5.2.9 Reply (face) 10](#_Toc80356631)

[5.2.10 Human commands 10](#_Toc80356632)

[5.2.11 AMS response 11](#_Toc80356633)

[5.2.12 Face identity 11](#_Toc80356634)

[5.2.13 Speaker identity 11](#_Toc80356635)

[5.2.14 Full World Representation 11](#_Toc80356636)

[5.2.15 FWR interaction 11](#_Toc80356637)

[Annex 1 - Terminology 12](#_Toc80356638)

[6 References 13](#_Toc80356639)

# Abstract

*Connected Autonomous Vehicles* (CAV) is one of the 10 current MPAI standard projects that also include Multimodal Conversation (MPAI-MMC), a standard comprising the Conver­sation with Emotion and Multimodal Question Answering Use Cases and 3 Speech Translation Use Cases. For the purpose of standardisation, a CAV has been subdivided in 5 subsystems, one of then called Human-CAV Interaction (HCI) dealing with the technology-rich scenario of the ways a human and a CAV can interact. As the analysis of the HCI standardisation needs has shown that some of the technologies required are already standardised in MPAI-MMC, this document collects all HCI-related material and proposes that HCI become its 6th Use Case.

# The MPAI Connected Autonomous Vehicle

A Connected Autonomous Vehicle (CAV) is a mechanical system capable of executing a command to move its body guided by an analysis of data produced by a range of sensors exploring the environment and information transmitted by other sources in range, e.g., CAVs and roadside units (RSU).

MPAI has subdivided a CAV in 5 main subsystems:

*Human-CAV interaction (HCI)*

Recognises the human CAV rights holder

Responds to humans’ commands and queries

Provides extended environment representation (called Full World Repres­entation) for humans to enjoy

Senses human activities during the travel

Activates other subsystems as required by humans or as deemed necessary by the identified conditions.

*Environment Sensing Subsystem (EDS)*

Acquires information from the physical environment via a variety of sensors

Develops the best environment representation (called Basic World Representation).

*Autonomous Motion Subsystem (AMS)*

Computes the Route to destination

Uses dif­ferent sources of information – CAV sensors, other CAVs and transmitting units – to produce a Full World Representation

Gives command that drive the CAV to the intended destination.

*CAV to Everything Subsystem (V2X)*

Teceives information from external sources, including other CAVs, other vehicles and Roadside Units (RSU).

Sends information to other CAVs.

*Motion Actuation Subsystem (MAS)*

Provides environment information

Actuates motion commands in the environment.

The interaction of the 5 subsystems in depicted in *Figure 1*.



*Figure 1 – The CAV subsystems*

The following high-level workflow illustrates the CAV operation envisaged by this docum­ent.

A *human* with appropriate credentials requests the CAV, via Human-CAV Interaction, to take the human to a given place.

*Human-CAV Interaction* authenticates the human, interprets the request and passes a command to the *Autonomous Motion Subsystem*.

*Autonomous Motion Subsystem*:

Requests *Environment Sensing Subsystem* to provide the current Pose.

Computes the Route.

Issues the start command.

*Environment Sensing Subsystem* computes and sends Basic World Representations to the *Autonomous Motion Subsystem.*

*CAV to Everything*

Becomes aware of other CAVs and external sources (CAVs, RSU etc.).

Shares the CAV’s Basic World Representation with CAVs in range.

*Autonomous Motion Subsystem*:

Receives and processes data broadcasted by external sources (CAVs, RSU etc.).

Computes the Full World Representation.

Shares the CAV’s Full World Representation with CAVs in range.

Computes a Path.

Issues commands to the *Motion Actuation Subsystem* to move the CAV accordingly.

While the CAV moves, *Humans*

Hold conversation with *Human-CAV Interaction* and possibly other hum­ans on board.

Issue commands.

Request views of the environment (Full World Representation) etc.

# The MPAI approach to AI standardisation

Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) is an [international standards developing organisation](http://mpai.community/) with the mission to develop *AI-enabled data coding standards*. Research has shown that data coding with AI-based technologies is generally *more efficient* than with existing technologies. Compression and feature-based description are notable examples of coding.

MPAI Standards enable the implementation of AI applications and services based on the standard execution environment called AI Framework (AIF) being developed in the MPAI-AIF standard. The AIF Reference Model is depicted in *Figure 2*.



*Figure 2 – The Components of the AI Framework (AIF)*

MPAI Application Standards normatively specify the technologies required to support one or more Use Cases. An example of an AIF-based implementation of one MPAI-MMC “Unidirectional Speech Translation” Use Case “interpret a sentence from a specified language to another language preserving the characteristics of the original speech” is given in *Figure 3*.



*Figure 3 – An AIF-based Use Case*

MPAI calls the rectangle containing the green boxes as “AI Workflow” (AIW). MPAI normatively specifies the following characteristics of an AIW:

1. The format of the input data, e.g., “source speech”, “source text” and “input and output lan­guages”.
2. The function, e.g., “interpreting a sentence from a language to another language preserving the characteristics of the original sentence”.
3. The format of the output data, e.g., “speech” and “text”.

An AIW is composed of data processing elements – called AI Modules (AIM). *Figure 4* and *Figure 5* depict two examples of AIMs. The former includes the necessary knowledge (e.g., in a neural network) while the latter accesses that knowledge from an external knowledge base.

|  |  |
| --- | --- |
|  |  |
| *Figure 4 – An AIM with embedded knowledge* | *Figure 5 – An AIM with access to an external knowledge base* |

MPAI normatively specifies the following aspects of an AIM:

1. The format and semantics of the input data, e.g., “Text and Speech Features”.
2. The function, e.g., “produce a synthetic speech from text and emotion descriptors”.
3. The format of the output data, e.g., “speech”.

An AIM is defined by its function and interfaces, but not by its internal architecture, which may be based on AI or data processing, and implemented in software, hardware or hybrid software and hardware technologies.

MPAI normatively specifies the process, the tools and the data or the characteristics of the data to be used by a Performance Assessor to Assess the Grade of Performance of an AIM or an AIW. MPAI defines Performance of an implementation as having the attributes of Reliability

Robustness, Replicability and Fairness.

MPAI standards offers implementers 3 different Levels of compliance to MPAI Standards:

|  |  |
| --- | --- |
| Level 1 | An Implementation running an AIW made up of AIMs performing any function and exposing any proprietary interface but exposing the interfaces that make AIMs executable in an AIF. |
| Level 2 | An Implementation having the additional feature of executing AIWs made up of AIMs whose functions and interfaces are specified by an MPAI Application Standard. |
| Level 3 | An Implementation with the additional feature of executing AIWs made up of AIMs certified to possess the attributes of Reliability, Robustness, Replicability and Fair­ness – collectively called Performance. |

The MPAI offers users access to the promised benefits of AI with a guarantee of increased transparency, trust and reliability as the Interoperability Level moves from 1 to 3. More informative details are provided by Annex 3.

# Terms and definitions

*Table 1 – Definition of Terms used in this document*

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Basic World Representation | A description of the Environment based on the CAV sensors and an Offline Map. |
| Command | High-level instructions whose execution allows a CAV to reach a Goal. |
| Connected Autonomous Vehicle | A vehicle capable to autonomously reach an assigned target by understan­ding human utterances, planning a route, sensing and interpreting the environment, exchanging information with other CAVs and acting on the CAV’s motion subsystem. |
| Decision horizon | The estimated time between the current state and the Goal. |
| Emotion | An attribute that indicates an emotion out of a finite set of Emotions |
| Emotion Grade | The intensity of an Emotion |
| Emotion Recognition | An AIM that decides the final Emotion out of Emotions from different sources |
| Environment | The portion of the world within the Decision horizon. |
| Full World Representation | A representation of the Environment using the CAV’s and other CAVs’ Basic World Representations. |
| Goal | The planned State at a future time.  |
| Intention | Intention is the result of a question analysis that denotes information on the input question. |
| Language Understanding | An AIM that analyses natural language as Text to produce its meaning and emotion included in the text |
| Meaning | Information extracted from the input text such as syntactic and semantic information |
| Offline Map | An offline-created map of a location and associated metadata. |
| Question Analysis | An AIM that analyses the meaning of a question sentence and determines its Intention  |
| Question Answering | An AIM that analyses the user’s question and produces a reply based on the user’s Intention  |
| Speech Recognition  | An AIM that converts speech to Text |
| Speech Synthesis | An AIM that converts Text or concept to speech |
| State | Pose, Velocity and Acceleration of a CAV at a given time. |
| Text | A collection of characters drawn from a finite alphabet |
| Translation | An AIM that converts Text in a source language to Text in a target language |
| Video analysis | An AIM that extracts features from video |
| Way Point | A point 𝑤𝑖 given as a coordinate pair (𝑥𝑖, 𝑦𝑖), in an Offline Map |

# The Human-CAV Interaction AIW

## Functionalities

The human-CAV interaction is based on the principle that the CAV is impersonated by an avatar, selected by the CAV right-holder, who has the capability to animate head and face and emit speech that inlude features that display as much as possible the features, e.g., emotion, that would be displayed by a human driver.

Examples are:

1. The CAV’s avatar is reactive to the Environment shows, e.g., it shows an angry face because a driver has made an improper motion.
2. The CAV’s avatar is reactive to a Human, e.g., it shows an appropriate face to a human who has made a joke.

Other forms of interaction are:

1. CAV authenticates human.
2. A human issues commands to a CAV, e.g.,
	1. Commands to Autonomous Motion Subsystem, e.g.: go to a Way point, display Full World Representation (see 5.3), etc.
	2. Other commands, e.g.: turn off air conditioning, turn on radio, call a person, open window or door, search for information etc.
3. A human entertains a dialogue with a CAV, e.g.,
	1. Information requests, e.g.: time to destination, route conditions, weather at destination etc.
	2. Casual conversation.
4. A CAV monitors the passenger compartment, e.g.,
	1. Physical conditions, e.g.: temperature level, media being played, sound level, noise level, anomalous noise, etc.
	2. Passenger data, e.g.: number of passengers, ID, estimated age, destination of passengers.
	3. Passenger activity, e.g.: level of passenger activity, level of passenger-generated sound, level of passenger movement, emotion on face of passengers.
	4. Passenger-to-passenger dialogue, two passengers shake hands, or passengers hold everyday conversation.

The Human-CAV Interaction collects a variety of data generated by humans inside the vehicle for possible action. This issue if part of the more general problem of data privacy in a CAV that is handled in a dedicated chapter of the Connected Automotive Vehicle document [1].

## Reference Human-CAV Interaction architecture

*Figure 6* is the of Human-CAV Interaction (HCI) reference model The following is noted:

1. A combination of Conversation with Emotion and Multimodal Question Answering AIMs with gesture recognition capabilities covers most Human-CAV Interaction needs.
2. New requirements to existing data format are added where required.
3. Additional AIMs are added when new HCI data formats are required.



*Figure 6 – Human-CAV Interaction Reference Model*

Depending on the technology used (data processing or AI), the AIMs in *Figure 6* may need to access external information, such as Knowledge Bases, to perform their functions. While not represented in *Figure 6*, they will be identified, if required, in the AI Modules subsection.

## Input and output data

*Table 1* lists the input and output data of the HCI AIMs Interaction depicted in *Figure 6*.

*Table 1 – HCI input and output data*

|  |  |  |
| --- | --- | --- |
| **Input data** | **From** | **Comment** |
| Audio | User outdoor | User authenticationUser command |
| Text | User outdoor | User authenticationUser command |
| Text | Passenger compartment | Social life of userCommands or interaction with CAV |
| Audio | Passenger compartment | Social life of userCommands or interaction with CAV |
| Video | Passenger compartment | Social life of userCommands or interaction with CAV |
| Full World Representation | Autonomous Motion SS | For processing by FWR Viewer |
| **Output data** | **To** | **Comments** |
| Text | Autonomous Motion Subsystem | Commands to be executed |
| Synthetic speech | Passenger compartment | CAV’s response to passengers |
| Synthetic face | Passenger compartment | CAV’s response to passengers |
| Full World Representation | Passenger compartment | For passengers to view external world  |

# Human-CAV Interaction AIMs

*Table 2* lists the AIMs depicted in *Figure 6*.

*Table 2* *– AI Modules of* *Human-CAV interaction*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Speech detection and separation** | 1. Separates relevant speech vs non-speech signals
2. Detects request for dialogue.
 |
| **Speaker identification** | Recognises speaker. |
| **Speech recognition** | 1. Analyses the speech input
2. Generates text and emotion output.
 |
| **Object and gesture analysis**  | 1. Analyses video to identify object
2. Produces the ID of the object in focus
3. Analyses video
4. Produces motion and mean­ing of gesture.
 |
| **Face recognition** | 1. Analyses the video of the face of a human
2. Recognise the human’s identity.
 |
| **Face analysis** | 1. Analyses the video of the face of a human
2. Extracts emotion and meaning.
 |
| **Language understanding** | 1. Uses a language model (embedded in AIM)
2. Analyses natural language expressed as text
3. Produces the meaning of the text.
4. Produces text related to Object ID
 |
| **Emotion recognition** | 1. Fuses Emotions from Speech, Face and Gesture.
2. Produces Final Emotion.
 |
| **Question analysis** | 1. Fuses Meanings of Speech, Face and Gesture
2. Analyses the meaning of the sentence
3. Determines the Intention.
4. Outputs Final Meaning
 |
| **Question & dialog processing** | 1. Receives Speaker ID and Face ID
2. If speaker ID and face ID match, then
	1. Produces a command to Autonomous Motion SS
	2. Analyses user’s emotion, intention, meaning and/or ques­tion, text
	3. Produces Reply (speech) and Reply (face).
3. Else, responds appropriately
 |
| **Speech synthesis** | Converts Reply (speech) to speech. |
| **Face animation** | Converts Reply (face) to animated face |
| **Full World Representation Viewer** | 1. Receives Full World Representation
2. Presents a view as instructed by human via Commands.
 |

# Technologies and Functional Requirements

The Functional Requirements refer to the individual technologies identified as necessary to implement MPAI-CAV Use Cases belonging using AIMs operating in an MPAI-AIF AI Framework. The Functional Requirements developed in this document adhere to the following guidelines:

AIMs are defined to allow implementations by multiple technologies (AI, ML, DP)

DP-based AIMs need interfaces, e.g., to a Know­ledge Base. AI-based AIM will typically require a learning process, however, support for this process is not included in the document. MPAI may develop further requirements covering that process in a future document.

AIMs can be aggregated in larger AIMs. Consequently, some data flows of aggregated AIMs may no longer be accessible.

## I/O Data summary

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

*Table 3 – I/O data of Human-CAV Interaction AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| **Object and gesture analysis** | Video | Object IDGesture IDEmotion |
| **Face analysis** | Video | EmotionMeaning |
| **Face identification** | Video | Face ID |
| **Speech separation** | Audio | Speech |
| **Speech recognition** | Audio | TextEmotion |
| **Speaker identification** | Audio | Speaker ID |
| **Language understanding** | Object identifierText | Text Meaning |
| **Emotion recognition** | Emotion (speech)Emotion (face)Emotion (gesture) | Final emotion |
| **Question analysis** | Meaning (speech)Meaning (face)Meaning (gesture) | Intention |
| **Question and dialogue processing** | Speaker IDFace IDIntentionTextFinal emotion | Synthetic speechFace animationText |
| **Speech synthesis** | Reply (speech) | Speech |
| **Face animation** | Reply (face) | Face animation |

MPAI has acquired a set of first-generation technologies related to the data types listed below in a previous Call for Technologies for the MPAI-MMC standard [1]. MPAI is ready, however, to consider new technologies related to the data below to support new requirements and/or to enhance capabilities.

## Data format requirements

### Text

As there is a need to support most languages in use, Text representation conforms to ISO/IEC 10646, Information technology – Universal Coded Character Set (UCS).

**To respondents**

Respondents are invited to comment on this choice.

### Audio

Audio is sampled from an analogue source (passenger compartment) at a frequency in the 44.1-96 kHz range with at least 16 and at most 24 bits/sample.

**To respondents**

Respondents are invited to comment on this choice.

### Video

Video is intended for use in the passenger compartment.

Video has adopted the following characteristics of 2D Video.

1. Pixel shape: square
2. Bit depth: 8-10 bits/pixel
3. Aspect ratio: 4/3 and 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 and BT2020
9. Colour format: RGB and YUV
10. Compression: uncompressed, if compressed AVC, HEVC

**To respondents**

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

Respondents are requested to propose a data format for 3D Video having video+depth as the baseline format.

### Object identifier

Include choice of MQA.

### Emotion

Include choice of CAE and MMC.

### Meaning

Include choice of MMC.

### Intention

Include choice of MMC.

### Reply (speech)

Include choice of MMC.

### Reply (face)

MPAI has already adopted a format for lips animation (MMC).

Here MPAI is looking for a technology that enables head and face animation capable to represent

1. Motion of head when speaking
2. Motion of face muscles and eyeballs
3. Turning of gaze to a particular person

### Human commands

The commands given to the Autonomous Motion Subsystem are:

1. Go to a Waypoint
2. Park close to a Waypoint
3. Go faster
4. Drive slowly
5. Display FWR

**To respondents**

Respondents are requested to propose a coded representation of the above commands. The addition of coded representation of new commands is welcome.

### AMS response

The responses of the Autonomous Motion Subsystem are:

1. The desired Waypoint has been reached

**To respondents**

Respondents are requested to propose a coded representation of the above commands. The addition of coded representation of new responses is welcome.

### Face identity

The Face Identity AIM shall be able to recognise the identity of a limited number of faces.

**To respondents**

Respondents are requested to propose a face identification system suitable for a limited number of faces.

### Speaker identity

The Speaker Identity AIM shall be able to recognise the identity of a limited number of speakers.

**To respondents**

Respondents are requested to propose a face identification system suitable for a limited number of speakers.

### Full World Representation

The requirements of the FWR AIM are developed in the context of CAV Autonomous Motion Subsystem requirements.

### FWR interaction

The requirements of FWR interaction should be developed once the FWR requirements will be defined.

# References

Connected Automotive Vehicles

MPAI-MMC Call for Technologies, N154, https://mpai.community/standards/mpai-mmc/#CfT