Towards a standard for human interaction with connected autonomous vehicles

Leonardo Chiariglione, Miran Choi, Gérard Chollet, Ferruccio Damiani, Jisu Kang, David Schultens, Mark Seligman, Gianluca Torta, Fathy Yassa

Towards a Standard for Human Interaction with Connected Autonomous Vehicles

Leonardo Chiariglione
President, Chairman of the Board
MPAI
Geneva, CH
president@mpai.community

Miran Choi
Language Intelligence Research Team
ETRI
Daejeon, South Korea
miranc@etri.re.kr

Gérard Chollet
Institut Mines Télécom
Évry, France
gerard.chollet@telecom-sudparis.eu

Ferruccio Damiani
Dipartimento di Informatica
Università di Torino
Turin, Italy
0000-0001-8109-1706

Jisu Kang
AI Research Center
Klleon
Seoul, South Corea
jisu.kang@klleon.io

David Schultens
Hubtropolis
Vallejo, CA, USA
davidschultens@gmail.com

Mark Seligman
Speech Morphing, Inc.
San José, CA, USA
mark@speechmorphing.com

Gianluca Torta
Dipartimento di Informatica
Università di Torino
Turin, Italy
0000-0002-4276-7213

Fathy Yassa
Speech Morphing, Inc.
San José, CA, USA
fathy@speechmorphing.com

***Abstract*—Moving Picture, Audio, and Data Coding by Artificial Intelligence (MPAI) have developed reference models of 4 subsystems in Connected Autonomous Vehicles (CAV). This paper delves into the "first" subsystem, the Human-CAV Interaction (HCI). It presents functionality, requirements, and technologies standardized in the first version of the HCI specification, functionality and requirements for the next version, that will be the target of an upcoming Call for Technologies.**

Keywords—connected autonomous vehicle, standard, reference model, human interface

I. INTRODUCTION

For several decades, Autonomous Vehicles have been the target of research and experimentation in industry and academia. Since a decade, trials on real roads are conducted, e.g., [1], and connected Vehicles are a reality today. In several countries, legislation has been enacted to allow circulation of autonomous vehicles, e.g., [2,3]. Technology continues to evolve with many research papers being produced every year on Connected Autonomous Vehicles (CAVs), e.g., [4-6].

Standardisation is a component of many industrial activities. However, its importance often depends on the mindset of the industries involved. CAVs are particularly relevant because of the different nature of the interacting technologies making up CAVs, the sheer size of the future CAV market [7] and the need for users and regulators alike to be assured of CAV safety, reliability, and explainability [8]. Another important element influencing the attitude toward standardisation is the fact that CAVs belong to a nascent industry, that will eventually be tasked to produce CAV units in the hundreds of millions p.a. using components coming from disparate sources.

The process of developing standards typically includes he phases of research - standardisation – industry deployment. The transition between the 3 phases can be facilitated by the creation of a flexible and modular CAV Reference Model focused on identification and consolidation of components and identification and definition of their interfaces. The transition from research to standardisation can then be implemented as a series of interactions between research proposing components and interfaces, and standardisation either requesting more results, or refining the results, or adopting the proposal. Eventually, the Reference Model will morph into a specification of functions and interfaces of standardised components ready to be reviewed and taken over by the industry.

Moving Pictures, Audio and Data Coding by Artificial Intelligence (MPAI) [9] is an international, unaffiliated, not-for-profit organisation developing AI-centred data coding standards. MPAI defines data coding as the transformation of data from one format to another that is more convenient to an application. An example if the transformation of the environment captured by a Lidar into an object-based visual scene for interpretation and action.

MPAI has produced a Reference Model where a CAV is subdivided in 4 subsystems. The *Human-CAV Interaction (HCI)* handles the human-CAV interface. The *Environment Sensing Subsystem* acquires information from the physical environment via a variety of sensors. The *Autonomous Motion Subsystem* interprets the sensed data, creates the Full World Representation, and issues commands to drive the CAV to the intended destination. The *Motion Actuation Subsystem* receives/actuates motion commands in and issues feedbacks from the environment.

Each of the 4 subsystems is an instantiation of the MPAI-standardised AI Framework (AIF) designed to create and execute AI Workflows (AIW) composed of AI Modules (AIM). AIMs correspond to the components introduced above. They are defined by their functions and interfaces, not by their internals which can be implemented with data

processing, AI and ML technologies in hardware, software, or hybrid.

The Reference Model identifies and specifies the requirements for the format and semantics of the data received or generated by the AIMs in the AIW corresponding to each subsystem. During the iterative process of research and standardisation described above, the data format specifications undergo a constant review as the update of an AIM may impact the AIMs it is connected to, and so on.

The Reference Model allows researchers to select test data and setups, propose updated interfaces, conduct contests, consider the influence of external components, and subdivide the workload in a way that allows unambiguous comparison of results. When the functions and requirements of a subsystem are considered mature, a Call for Technologies is issued, to acquire the technologies that are selected and integrated in a standard resulting from competition between proposals.

MPAI is aware of the difficulties encountered by those attempting to use advanced technology standards such as those likely to be required for implementing CAVs. In its standardisation process, MPAI has replaced the vague and ambiguous notion of Fair, Reasonable and Non-Discriminatory (FRAND) patent declarations [10] with a process whereby submitters of technologies for standardisation agree to license their technologies according to a standard-specific Intellectual Property Rights Framework called Framework Licence. Among other items, the Framework stipulates that the licences of the technologies issued by Standard Essential Patent Holders will be issued at a price comparable with similar standard technologies and not after products that use the technologies are on the market.

The paper is structured as follows. Sections II and III introduce the fundamentals of the AI Framework (AIF) and the 4 CAV subsystems, respectively, while Section IV adds details on the Autonomous Motion Subsystem. Section V and its subsections provide details of the current state of the specification of the Human-CAV Interaction subsystem, and Section VI presents the functions and requirements of the next version. Finally, Section VII compares our proposal with related work, and section VIII points to future work directions and concludes the paper.

II. The MPAI AI Framework

The development of MPAI standards is driven by Use Cases. This process allows MPAI to subdivide AI systems designed to implement Use Cases into the functional elements called AI Modules (AIM) introduced above. AIMs have a specified function, and their input and output data have specified syntax and semantics. *Figure 1* depicts Path Planner, an AIM whose function is the creation of a sequence of Paths by receiving a Route as input and by accessing the World Representation defined in the following.

AIMs are typically implemented with Artificial Intelligence technologies such as Neural Networks. However, AIMs can be implemented with other Data Processing technologies. AIMs can consist of software, hardware, and mixed hardware-software. MPAI defines *standard interfaces* of AIMs combined and executed as *AI Workflows* (AIW) in an MPAI-specified *AI-Framework* (AIF). AIMs operate on input data and produce output data with standard syntax and semantics.

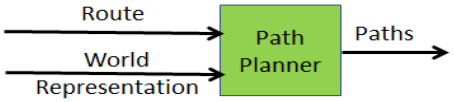

Figure 1 – The MPAI AI Module (AIM)

Figure 2 represents the AIF Reference Model [11]. The role of the AIF is to support the execution of AI Workflows (AIW) built from AIMs. The Reference Model envisions an entity called MPAI Store tasked with the handling of sub-missions and registrations of AIMs and AIWs; the unique identification of AIMs and AIWs; and the queries issued by AIFs for credentials, metadata, and URLs of specific AIMs and AIWs.

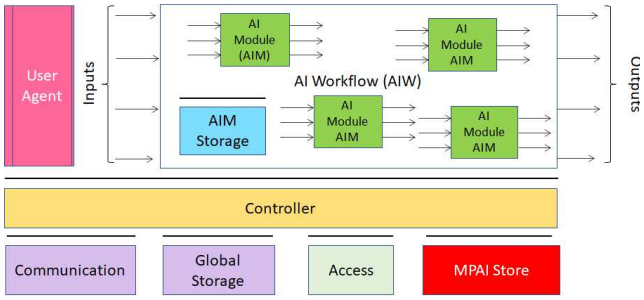

Figure 2 – The MPAI AI Framework (AIF) including an AI Workflow (AIW)

AIWs connect AIMs into computational graphs, In the simpler case these are directional and acyclic (DAGs) and express computations flowing from the inputs to the outputs of the workflow. In more complex cases they are cyclic, allowing AIMs to generate feedback to upstream AIMs.

An important characteristic of the AIF is the adoption of a zero-trust model. This advocates mutual authentication of system components, including checking identity and integrity of components irrespective of location and provides access based on the confidence on component identity and health.

The Controller is in charge of the overall control of the AIW execution (see *Figure 2*). It provides basic functionalities such as scheduling the execution of AIMs and handling communication between AIMs and other AIF Components, e.g., Internal and Shared Storage; runs one or more AIWs at a time; activates/suspends/resumes/deactivates AIWs based on user or other inputs. Moreover, it may communicate with other (remote) Controllers as explained below.

For many workflows, it is not important to consider how many instances are currently running, since each instance is completely independent of the others. In such cases, the distinction between a workflow definition and its instances is almost irrelevant. In some important scenarios, however, different instances of the same workflow (or even of different workflows) need to communicate with each other to perform their task.

III. The MPAI-CAV Reference Model

Figure 3 represents the MPAI-CAV Reference Model, based on 4 Subsystems. Each subsystem corresponds to one AIW:

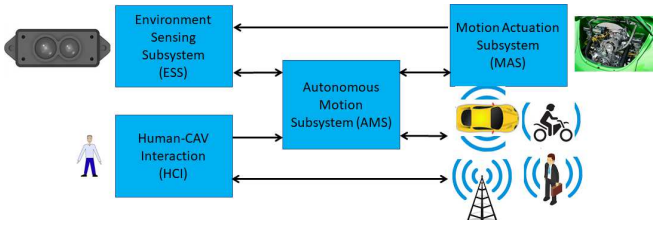

Figure 3 – The MPAI-CAV Subsystems

1. *Human-CAV interaction (HCI)* recognises the CAV rights holder (owner or tenant), responds to humans' commands and queries, provides extended environment representation (Full World Representation, see below) for humans to use, senses human activities during the travel and may activate other subsystems as required by humans or as deemed necessary by a CAV Subsystem based on identified conditions. This paper is mostly focused on this subsystem.
2. *Environment Sensing Subsystem* acquires information from the physical environment via a variety of sensors and produces a representation of the environment (Basic World Representation) that is its best estimate given the sensory data available to the CAV.
3. *Autonomous Motion Subsystem* computes the Route to destination, based on the result of human-CAV interaction, and uses different sources of information – CAV sensors, other CAVs and transmitting units – to produce a Full World Representation and give commands that drive the CAV to the intended destination.
4. *Motion Actuation Subsystem* provides non-electromagnetic environment information (position, etc.), and receives and actuates motion commands in the environment.

IV. The Autonomous Motion Subsystem

The typical series of operations carried out by the Autonomous Motion Subsystem, which is at the core of the CAV autonomy, is the following:

1. *Human-CAV Interaction* requests the Autonomous Motion Subsystem to plan and move the CAV to the human-selected waypoint.
2. The CAV requests the Environment Sensing Subsystem to provide the current Basic World Representation
3. While moving, the CAV
   a. Transmits the Basic World Representation and other data to CAVs in range.
   b. Receives Basic World Representations and other data from CAVs in range.
   c. Produces the Full World Representation by fusing its own Basic World Representation with those from other CAVs in range.
   d. Plans a Path connecting Poses.
   e. Selects the behaviour allowing it to reach intermediate Goals, considering information about the Goals other CAVs in range intend to reach.
   f. Defines a Trajectory that complies with general traffic rules and local traffic regulations and preserves passengers' comfort.
   g. Refines Trajectory to avoid obstacles.
   h. Sends to the Motion Actuation Subsystem the command(s) that take the CAV to the next Goal.

The Autonomous Motion Subsystem Reference Model is represented in Figure 4. It is out of the scope of the present paper to describe each AIM in the subsystem. For our purposes, it is important to note that the Autonomous Motion Subsystem exchanges data with the HCI through both the Route Planner AIM and the Full World Representation Fusion AIM.

Moreover, a CAV exchanges information via radio with other entities, e.g., CAVs in range and other CAV-like communicating devices such as Roadside Units and Traffic Lights, thereby improving its environment perception capabilities. One of the most important pieces of information exchanged between CAVs is the Basic World Representation mentioned above. It is a high-level description of the objects sensed by a CAV, comparable with the Cooperative Perception Messages defined by ETSI standards [12,13].

The MPAI-AIF standard supports communication between an AIM that is part of an AIW running on an AIF and its peers running on other (remote) AIFs. Such communication is made possible by the Controllers associated with the AIFs:

1. The AIM invokes its own Controller to get a list of the remote Controllers in range, with metadata about the AIWs currently executed by such controllers Using the metadata about remote Controllers, the AIM can send or receive data to/from specific AIMs running on remote Controllers, through so-called *remote ports*.
2. In particular, in the Autonomous Motion Subsystem, the Full World Representation (FWR) Fusion AIM can request to its local Controller the list of Controllers corresponding to other CAVs in range, and then receive their Basic World Representations (BWRs) to be fused with its own BWR into an FWR.

At the lower levels, the communication happens in broadcast mode when a CAV advertises its identity and when it transmits heavy messages, such as the BWR. Unicast mode may be used in other cases.

Communication is handled by a Communication Device that makes the relevant data available to the AIMs when they request it to their Controller.

The text above outlines the role of the Communication Device in connecting the Autonomous Motion Subsystems (actually, their FWR Fusion AIMs). However, it can also allow communication between other remote subsystems/ AIMs, e.g., an AIM of the Motion Actuation Subsystem could inform its peers in range of the sudden appearance of ice on the road.

V. The current Human-CAV Interaction specification

A. Reference Model of Human-CAV Interaction

Interaction of humans with vehicles with different SAE Levels of Driving Automation [14] have been the subject of several papers (e.g., [15]). This paper, however, addresses the human-CAV interaction from the viewpoint of a fully autonomous CAV where a user expects to be able to ask questions to, hold conversation with and receives information from a CAV that is perceived as a replacement of a human driver.

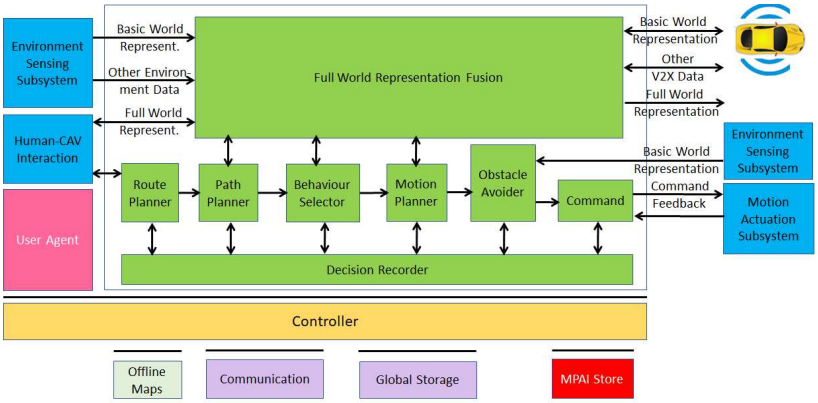

Figure 4 – Autonomous Motion Subsystem Reference Model

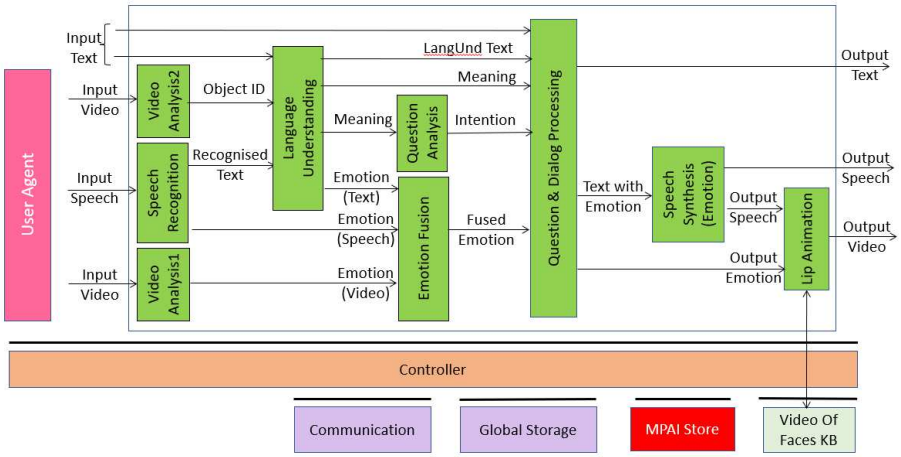

Figure 5 – Human-CAV Interaction Reference Model (Version 1)

The functionalities supported in the first version of the Human-CAV Interaction (HCI) have been motivated by the consideration that the passengers should perceive the HCI as a "personal driver" with which communication should be as natural and possible and defined with the goal to define components (AIMs) able to be interconnected with other componets an offer the requested functionalities.

In other words, the HCI should provide a set of admittedly functionalities some of which appear ambitious today but can be be expected to be reached is a sequence of versions of the technologies:

1. Recognise the identity of the owner or tenant.
2. Perceive the emotional state of a passenger communicating to it and respond with the appropriate type and level of emotion in its utterance responding to them.
3. Display a face expressing a sympathetic emotion in sync with its utterance and with eyes gazing at the passenger the speech is directed to.
4. Entertain a conversation with and respond to questions from the human on well-identified topics in the automotive context.

The following will introduce some of the standard elements specified by the MPAI Multimodal Conversation (MPAI-MMC [16]) standard and the plan for the next version of MPAI standards relevant to HCI. An example is Context-Based Audio Enhancement (MPAI-CAE [17]) which will standardise such technologies as separation of speech from environment sounds and identification of different sounds.

Figure 5 shows the Reference Model for the HCI V1. The standards developed so far support the following operations of HCI:

1. Humans interact with text and/or speech conversing with and asking questions to HCI. Currently allowed questions may involve pictures shown by the user, e.g., the human passenger may ask the HCI to take them to a shop by providing a picture of it. Obviously, questions can take forms that do not involve concrete objects held by the user.
2. HCI can observe the face and the object held by the human through the Video Analysis1 AIM (extracts emotion from face) and Video Analysis2 AIM (extracts object ID), respectively (see *Figure 5).*
3. Speech Recognition extracts both Text and Emotion from speech.
4. The text and/or the recognised speech, and the Object ID are fed to the Language Understanding AIM which extracts Meaning and Emotion and provides a Text refined from the text provided by the Speech Recognition AIM.
5. All sources of Emotion are fed to the Emotion Fusion AIM which produces Fused Emotion.
6. Dialogue Processing replies to the human utterance based on Input Text, Text from Language Understanding, Meaning, Intention and Fused Emotion. The reply takes the form of Text, Speech with embedded Emotion generated by HCI, and Emotion which is used to animate the lips of an avatar in sync with the speech.

B. Main data format specification

A Call for Technologies was issued in February 2021 and responses received. As mandated by its statutes, MPAI has developed the specifications for the main data formats used in the HCI Subsystem.

The main data formats that have been standardised concern the following data types with known semantics:

1. Emotion, an identifiable state of speech and face.
2. Intention, the intention embedded in a question.
3. Meaning, the meaning of a question.
4. "Video of Faces" KB Query Format, the format by which an external knowledge base of videos is queried by providing an emotion to obtain a matching output video.

In the following, the main aspects of the standard data type formats are described.

1) Emotion

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.).

Human Emotion is represented in the MPAI HCI standard by the following JSON schema:

```
{
  "$schema":"http://json-schema.org/draft-07/schema",
  "definitions":{
    "EmotionType":{
      "type":"object",
      "properties":{
        "emotionDegree":{
           "type":"{Enum high | Enum medium | Enum low}"
        },
        "emotionName":{
          "type":"string"
        },
        "emotionSetName":{
          "type":"string"
        }
      }
    }
  }
```

MPAI has standardised a three-level Emotion set. The EMOTION CATEGORIES column specifies the categories using nouns; the GENERAL ADJECTIVAL column gives adjectival labels for general or basic emotions within a category; and the SPECIFIC ADJECTIVAL column gives labels for more specific (sub-categorized) emotions in the relevant category, often (but not always) representing differing degrees of the basic emotion. Emotion names are given by the elements of General Adjectival and Specific Adjectival columns.

Two examples are given in Table I.

TABLE I. BASIC EMOTIONS (EXAMPLES)

EMOTION CATEGORIES	GENERAL ADJECTIVAL	SPECIFIC ADJECTIVAL
HAPPINESS	happy	joyful
		content
		delighted
		amused
SADNESS	sad	lonely
		grief-stricken
		discouraged
		depressed
		disappointed

Examples of the semantics for each label in the GENERAL ADJECTIVAL and SPECIFIC ADJECTIVAL columns are given in Table II.

TABLE II. EMOTION SEMANTICS (EXAMPLES)

Emotion	Meaning
admiring/ approving	emotion due to perception that others' actions or results are valuable
amused	positive emotion combined with interest (cognitive)
anger	emotion due to perception of physical or emotional damage or threat
anxious/uneasy	low or medium degree of fear, often continuing rather than instantaneous
aroused/excited/ energetic	cognitive state of alertness and energy
arrogant	emotion communicating social dominance

The MPAI process envisages the case of an implementor who wishes to extend the tables or produce entirely new tables and have them certified by MPAI.

2) Intention

Provides abstracts of Intention of Question. The Syntax is represented by

```
{
  "$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" },
        "qdomain":{ "type":"string"}
      }
    }
  },
}
```

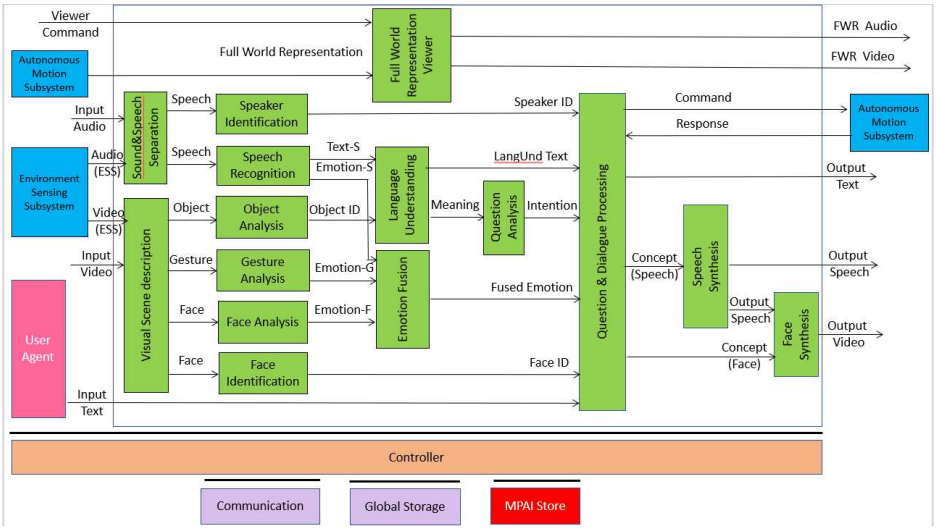

Figure 6 - Human-CAV Interaction Reference Model (V2)

The Semantics of the five properties is given by:

qtopic	The object or event that the question is about. E.g., "Town Hall" is qtopic of "When will we reach the Town Hall?"
qfocus	The question part that, replaced by the answer, makes the question a stand-alone statement. E.g., "What" is qfocus of "What is the park we are passing?"
qLAT	The lexical answer type of the question. E.g., "designer" is qLAT for "author" in "Who is the designer of the bridge we are passing?"
qSAT	The semantic answer type of the question. E.g., "person" is qSAT for "designer" in "Who is the designer of the bridge we are passing?"
qdomain	The domain of the question, e.g., "CAV internal status", "environment", "road status", etc.

3) Meaning

Provides the meaning of the question, i.e., an abstract description of natural language analysis results. The Syntax of Meaning is represented by

```
{
  "$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"}
        }
      }
    }
  }
}
```

The Semantics of the four properties is given by:

POS_tagging	The result of tagging of Parts of Speech.
NE_tagging	The result of tagging of Named Entities, e.g., person, organisation, type of object.
dependency_tagging	The result of tagging of Dependencies (i.e., the structure of the sentence, e.g., subject, object, head of the relation, etc.)
SRL_tagging	The result of tagging of Semantic Role Labelling, i.e., the semantic structure of the sentence e.g., agent, location, patient role, etc.

4) Video of Faces KB Query Format

The Video of Faces KB is queried with an Emotion. The response of the KB is a Video File of a human face with the selected Emotion.

VI. The Next Human-CAV Interaction Requirements

A. Reference Architecure

Version 1 of the Human-CAV Interaction (HCI) Reference Model is a significant achievement because it integrates disparate technologies instantiated as AI Modules (AIM) in a complete system allowing a human to interact with an HCI where both human and HCI use text, speech and video. The design assumptions made, however, were limited both functionally and technologically.

The purpose of Version 2 is to overcome some of those limitations, still preserving the modular approach, with the understanding that some of the AIMs depicted in Figure 6

could be merged when performance improvements justify it. Note that in general, however, MPAI favours the identification of AIMs with specific functions because they increase the HCI explainability: splitting the computation into meaningful steps allows to trace back the final outcome to partial outcomes obtained by the intermediate steps.

These are the additional goals pursued with Version 2:

1. Human identification via speech and face.
2. Speech is separated from the sound, acquired by the microphones both outside and inside the CAV for the purpose of obtaining a cleaner speech for speech recognition, for dialogue and speaker identification purposes.
3. A description of the visual scene is required to allow the HCI to obtain independent face, gesture, and object – including their spatial coordinates – and to allow passenger identification.
4. Emotion fusion includes emotion extracted from gesture.
5. Speech and Face Synthesis is no longer simply based on Text and Emotion, but on more complex data structures (Concept) rather than a simple sequence of words with associate emotion.
6. Humans can have verbal interactions with the Autonomous Motion Subsystem in specific domains identified in the following.
7. The Full World Representation produced by the Autonomous Motion Subsystem is made available to passengers for consumption.

As mandated by the MPAI Statutes, the technologies supporting the new or extended functionalities will be acquired with the responses to one or more Calls for Technologies.

The following subsections give a sample of the data format requirements for version 2 of the Reference Model.

B. Requirements

1) Human identification

A human identification format based on speech and face that is suitable for a limited number of individuals.

2) Sound and Speech Separation

It should be possible to provide separate audio objects, e.g., for human identification and as part of the Full World Representation. Similarly, the microphones inside the CAV should be able to provide separate speech objects for each of the passenger for efficient human-to-CAV conversation.

3) Visual scene description

The visual scene description shall support the following requirements:

1. The visual information of the human speaking to the CAV should be separated from the rest of the visual scene inside and outside of the CAV.
2. The humans inside the CAV should be individually separated and their spatial coordinates provided.

4) Speech and Face Synthesis

The "Concept to Speech" data format should be able to represent emotions and meaning varying in time and the "Concept to Face" data format should be able to represent:

1. Motion of head when speaking.
2. Motion of face muscles and eyeballs.
3. Turning of gaze to a particular person.
4. Emotion of the associated spoken sentence.
5. Meaning of the associated spoken sentence.

5) Verbal interaction with CAV

Passengers should be able to entertain a domain-specific interaction with the CAV (specifically with the Autonomous Motion Subsystem) with questions and commands such as the following:

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

6) Full World Representation

Passengers should be able to interact with the Full World Representation in the following ways:

1. Select an area whose coordinates are covered by the Full World Representation.
2. Access physical parameters of the environment: e.g., weather, temperature, air pressure, ice and water on the road.
3. Access the following parameters of each object: position, velocity, acceleration, bounding box (more, if available), semantics (if available), flags (e.g., warning).
   View road structure and local traffic signalisation.
4. View a specific object with a representation supporting a choice of the Level of Detail.
5. Access individual sounds identified by Audio and Speech Separation with their spatial coordinates and semantics.

C. Additional features

With its plans to develop Version 2, MPAI intends to revisit some of the Technologies standardised in Version 1, in particular:

1. Motivated extensions of the standard emotion sets or new technology supporting enhanced emotion representation. V1 offers a standard set of emotions enabling the creation of a market of components, emotion remains an intense area of research [18,19]. On the other hand, V2 is open to considering means to represent emotions with a finer grade or to define new ways of representing emotions.
2. Motivated extensions or new technology to express Meaning, especially if applicable to other information sources, such as face and gesture. V1 has addressed the meaning embedded in text and speech, but meaning can be conveyed by visual parts of the human body, both to understand the meaning expressed by a human and to expand the meaning expressed by the avatar materialising a CAV.

VII. Related Work

Papers proposing overall CAV Reference Models have already been published (e.g., [20-21]). The MPAI Reference Model presented in Section III, however, differs in several respects: 1) it adopts a holistic approach that includes all IT components of a CAV; 2) it uses AIF-AIW-AIM as the unifying model to determine the functionality and the data of all CAV components; 3) it benefits from AIMs whose functions and data are being or have already been specified in MPAI standards already developed; 4) it focuses on the formats of the data exchanged between AIMs rather than just on the AIMs functions. Specifically, regarding the Human-CAV Interaction (HCI) subsystem, some of the overall CAV Reference Models either ignore it [22], or give a very high-level, incomplete description of its components and data. As we have shown in this paper, the MPAI model of the HCI subsystem identifies many components, their function and, most importantly, the data they exchange.

Some other works present just the design model of the HCI component of CAVs. In [23], the authors present at a high level the design for the HCI of a CAV considering two main components: an *internal* HCI for the communication between CAV and its passengers; and an *external* HCI for the communication between the CAV and other, external participants in the traffic environment (e.g., pedestrians). The focus of the authors is on a minimal internal HCI, based on a touch screen for input/output interaction with the CAV.

On the other hand, the present paper only considers the internal HCI component envisioning a much richer multimodal interaction between CAV and passengers, including gesture-enhanced dialogue. The choice between minimality and richness depends on several factors, but we believe that, as vehicles approach full autonomy, passengers will increasingly want to interact with the CAV as the new shape of a human personal driver. The new virtual driver may even know our schedule and take us to work every morning without even asking for input from us because it will use context and historical data to make predictions about our destination [24]. Clearly, such a view would require rich, intelligent interactions between humans and CAV. This is the main reason why an avatar impersonation of the CAV, capable to exhibit emotions, gestures, and gaze at the human interlocutor during interactions with the passengers has been envisioned.

[25] proposes a multimodal HCI like the one pursued by MPAI. The authors propose a multimodal HCI where speech, gesture, and eye gaze coming from the passenger(s) are recognized and fused by the CAV in order to conduct a dialogue with the goal of understanding commands and uttering responses aloud through speech synthesis. The overall vision underlying the proposal made by [25], where the "user operates an autonomous vehicle as a taxi by speech", is remarkably like ours. As shown above, we consider a richer multimodal interface requiring, e.g., emotion recognition, passenger, and objects identification, and allowing broader and more complex dialogues (in [25] a simple dialogue handling model based on transducers is proposed). Moreover, our focus is the standardisation of the data exchanged by the HCI components, and other CAV subsystems.

VIII. Conclusions and future work

Connected Autonomous Vehicles can benefit from standardization of their software components, and MPAI has undertaken this task with an innovative process, that contemplates iterations and transitions from research, standardisation and industry. The goal is to achieve agreed standard components benefitting users and industry thus simplifying the task of regulators.

In this paper, we have touched on the subsystems of the MPAI CAV Reference Model, going into detail for the Human-CAV Interaction subsystem, because it is the one with a more mature specification, and because it best illustrates the synergy with other MPAI Use Cases. For the interested readers, [26] continuously updates the current status of the CAV standardisation effort.

There are many directions in which we would like to expand and refine our work. First of all, in the current versions of the MPAI CAV model, we are assuming fully autonomous vehicles (i.e., SAE level 5). While the field of AVs is rapidly progressing, the current vehicles are at most at level 3, and even when level 5 will be reached we will face mixed traffic scenarios in which different levels of automation will coexist. We should therefore enrich the HCI model such that it is still useful with lower levels of CAV autonomy.

A related aspect that deserves further work is the explicit consideration in our HCI model of the role played by systems that are already present in today's vehicles for the comfort and safety of drivers and passengers, such as ADAS. While it is true that fully autonomous CAV may not need to interact with humans in order to exploit such safety features, it is easy to imagine hybrid situations where the CAV has some autonomy, but humans still desire or need to intervene when some maneuvers are performed manually.

The implementation of prototypes of the proposed models is another fundamental goal that we have in the mid to long term since the MPAI standardisation process explicitly requires the development of reference implementations for each component it standardises. Since the MPAI effort has started just 18 months ago, such implementations are still in their early phases, but some of them are already on their way. Minimum viable prototypes will be released as open-source software and potentially hardware as they become available. It is also worth noting that any interested party is welcome to join the implementation effort by becoming a member of the MPAI Community [9].

References

1. Smart Mobility Projects and Trials Across the World, https://imoveaustralia.com/smart-mobility-projects-trials-list/
2. Takeyoshi Imai, Legal regulation of autonomous driving technology: Current conditions and issues in Japan, IATSS Research, Volume 43, Issue 4, December 2019, Pages 263-267
3. Dasom Lee, David J. Hess, Regulations for on-road testing of connected and automated vehicles: Assessing the potential for global safety harmonization; Transportation Research, 08 April 2020

4. David Elliott, Walter Keen, Lei Miao; Recent advances in connected and automated vehicles; Journal of Traffic and Transportation Engineering, Volume 6, Issue 2, April 2019, Pages 109-131
5. Siegel, Joshua E; Erb, Dylan Charles; Sarma, Sanjay E; A Survey of the Connected Vehicle Landscape – Architectures, Enabling Technologies, Applications, and Development Areas, IEEE Transactions on Intelligent Transportation Systems; vol. 19, no. 8, Aug. 2018, pp. 2391–406
6. Claudine Badue et al., Self-driving cars: A survey, Expert Systems with Application, 4 August 2020, 165 (2021) 1138 16
7. Global Autonomous Cars (Semi & Fully) Market Report 2021, https://www.prnewswire.com/news-releases/global-autonomous-cars-semi--fully-market-report-2021-market-is-expected-to-reach-1-383-89-billion-in-2025---forecast-to-2030--301292821.html
8. A Connected and Autonomous Vehicle Reference Architecture for Attack Surface Analysis; November 2019, Applied Sciences 9(23):5101, DOI:10.3390/app9235101
9. MPAI Community; https://mpai.community/
10. European Commission, Fair, Reasonable and Non-Discriminatory (FRAND) Licensing Terms. Research Analysis of a Controversial Concept, https://publications.jrc.ec.europa.eu/repository/handle/JRC96258 (2015)
11. MPAI Community. "Artificial Intelligence Framework (MPAI-AIF) V1": https://bit.ly/3t3SNDT
12. ETSI: Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Analysis of the Collective Perception Service (CPS), Release 2 (2019)
13. Generation of Cooperative Perception Messages for Connected and Automated Vehicles; IEEE Transactions On Vehicular Technology, Vol. 69, No. 12, December 2020
14. SAE J3016; Levels of driving automation™ https://sae.org/standards/content/13016_202104
15. Zhengyu Tan et a.; Human–Machine Interaction in Intelligent and Connected Vehicles: A Review of Status Quo, Issues and Opportunities; IEEE Transactions on Intelligent Transportation Systems; Digital Object Identifier 10.1109/TITS.2021.3127217
16. MPAI Community. "Multi-modal conversation (MPAI-MMC) V1.1": https://bit.ly/3t1cUTd
17. Marina Bosi et al., Designing A Standard For Sound And Music Using Artificial Intelligence; in Proceedings of the 18th Sound and Music Computing Conference 2021, SMC'21, Virtual Conference, 2021, DOI: 10.5281/zenodo.5045003
18. Herbert L. Meiselman; A review of the current state of emotion research in product development; Food Research International, Volume 76, Part 2, October 2015, Pages 192-199
19. Paul Ekman; What Scientists Who Study Emotion Agree About; Perspectives on Psychological Science 2016, Vol. 11(1) 31–34
20. A.C. Serban, E. Poll, J. Visser; A Standard Driven Software Architecture for Fully Autonomous Vehicles; IEEE International Conference on Software Architecture Companion (ICSA-C), 2018
21. Bijayita Thapa, Eduardo B. Fernandez; A Survey of Reference Architectures for Autonomous Cars; Pattern Languages of Programs Conference, 2020.
22. S. Behere and M. Törngren, "A functional reference architecture for autonomous driving," Information and Software Technology, vol. 73, pp. 136–150, 2016.
23. O. Benderius, C. Berger and V. Malmsten Lundgren, "The Best Rated Human–Machine Interface Design for Autonomous Vehicles in the 2016 Grand Cooperative Driving Challenge," in IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 4, pp. 1302-1307, April 2018, doi: 10.1109/TITS.2017.2749970
24. A. L. Kun, S. Boll and A. Schmidt, "Shifting Gears: User Interfaces in the Age of Autonomous Driving," in IEEE Pervasive Computing, vol. 15, no. 1, pp. 32-38, Jan.-Mar. 2016, doi: 10.1109/MPRV.2016.14
25. T. Nakagawa, R. Nishimura, Y. Iribe, Y. Ishiguro, S. Ohsuga and N. Kitaoka, "A human machine interface framework for autonomous vehicle control," 2017 IEEE 6th Global Conf. on Consumer Electronics (GCCE), 2017, pp. 1-3, doi: 10.1109/GCCE.2017.8229312.
26. MPAI Community, CAV, http://cav.mpai.community/