



IEEE Standard Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Multimodal Conversation—Version 2

Developed by the
Collaborative Activities Governance Board

IEEE SA Industry Affiliated Network (IAN)
Base specification contributed by MPAI community



IEEE Std 3300™-2024

IEEE Standard Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Multimodal Conversation—Version 2

Developed by the

Collaborative Activities Governance Board
of the
IEEE SA Board of Governors

Approved 26 September 2024

IEEE SA Standards Board

Abstract: This standard adopts MPAA Technical Specification Version 2 as an IEEE Standard. Multimodal Conversation (MPAA-MMC) specifies use cases, all of which share the use of artificial intelligence (AI) to enable a complete and intense form of human-machine conversation.

Keywords: adoption, AI modules, bidirectional speech translation, BST, conversation with emotion, CWE, IEEE 3300™, MQA, MST, multimodal conversation, multimodal question answering, one-to-many speech translation, unidirectional speech translation, use cases, UST

The Institute of Electrical and Electronics Engineers, Inc.
3 Park Avenue, New York, NY 10016-5997, USA

Copyright © 2024 by The Institute of Electrical and Electronics Engineers, Inc.
All rights reserved. Published 14 November 2024. Printed in the United States of America.

IEEE is a registered trademark in the U.S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated.

PDF: ISBN 979-8-8557-1417-3 STD27456
Print: ISBN 979-8-8557-1418-0 STDPD27456

IEEE prohibits discrimination, harassment, and bullying.

For more information, visit <https://www.ieee.org/about/corporate/governance/p9-26.html>.

No part of this publication may be reproduced in any form, in an electronic retrieval system or otherwise, without the prior written permission of the publisher.

Important Notices and Disclaimers Concerning IEEE Standards Documents

IEEE Standards documents are made available for use subject to important notices and legal disclaimers. These notices and disclaimers, or a reference to this page (<https://standards.ieee.org/ipr/disclaimers.html>), appear in all IEEE standards and may be found under the heading “Important Notices and Disclaimers Concerning IEEE Standards Documents.”

Notice and Disclaimer of Liability Concerning the Use of IEEE Standards Documents

IEEE Standards documents are developed within IEEE Societies and subcommittees of IEEE Standards Association (IEEE SA) Board of Governors. IEEE develops its standards through an accredited consensus development process, which brings together volunteers representing varied viewpoints and interests to achieve the final product. IEEE standards are documents developed by volunteers with scientific, academic, and industry-based expertise in technical working groups. Volunteers involved in technical working groups are not necessarily members of IEEE or IEEE SA and participate without compensation from IEEE. While IEEE administers the process and establishes rules to promote fairness in the consensus development process, IEEE does not independently evaluate, test, or verify the accuracy of any of the information or the soundness of any judgments contained in its standards.

IEEE makes no warranties or representations concerning its standards, and expressly disclaims all warranties, express or implied, concerning all standards, including but not limited to the warranties of merchantability, fitness for a particular purpose and non-infringement. IEEE Standards documents do not guarantee safety, security, health, or environmental protection, or compliance with law, or guarantee against interference with or from other devices or networks. In addition, IEEE does not warrant or represent that the use of the material contained in its standards is free from patent infringement. IEEE Standards documents are supplied “AS IS” and “WITH ALL FAULTS.”

Use of an IEEE standard is wholly voluntary. The existence of an IEEE standard does not imply that there are no other ways to produce, test, measure, purchase, market, or provide other goods and services related to the scope of the IEEE standard. Furthermore, the viewpoint expressed at the time a standard is approved and issued is subject to change brought about through developments in the state of the art and comments received from users of the standard.

In publishing and making its standards available, IEEE is not suggesting or rendering professional or other services for, or on behalf of, any person or entity, nor is IEEE undertaking to perform any duty owed by any other person or entity to another. Any person utilizing any IEEE Standards document should rely upon their own independent judgment in the exercise of reasonable care in any given circumstances or, as appropriate, seek the advice of a competent professional in determining the appropriateness of a given IEEE standard.

IN NO EVENT SHALL IEEE 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.

Translations

The IEEE consensus balloting process involves the review of documents in English only. In the event that an IEEE standard is translated, only the English language version published by IEEE is the approved IEEE standard.

Use by artificial intelligence systems

In no event shall material in any IEEE Standards documents be used for the purpose of creating, training, enhancing, developing, maintaining, or contributing to any artificial intelligence systems without the express, written consent of IEEE SA in advance. “Artificial intelligence” refers to any software, application, or other system that uses artificial intelligence, machine learning, or similar technologies, to analyze, train, process, or generate content. Requests for consent can be submitted using the Contact Us form.

Official statements

A statement, written or oral, that is not processed in accordance with the IEEE SA Standards Board Operations Manual is not, and shall not be considered or inferred to be, the official position of IEEE or any of its committees and shall not be considered to be, or be relied upon as, a formal position of IEEE or IEEE SA. At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that the presenter’s views should be considered the personal views of that individual rather than the formal position of IEEE, IEEE SA, the Standards Committee, or the Working Group. Statements made by volunteers may not represent the formal position of their employer(s) or affiliation(s). News releases about IEEE standards issued by entities other than IEEE SA should be considered the view of the entity issuing the release rather than the formal position of IEEE or IEEE SA.

Comments on standards

Comments for revision of IEEE Standards documents are welcome from any interested party, regardless of membership affiliation with IEEE or IEEE SA. However, **IEEE does not provide interpretations, consulting information, or advice pertaining to IEEE Standards documents.**

Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments. Since IEEE standards represent a consensus of concerned interests, it is important that any responses to comments and questions also receive the concurrence of a balance of interests. For this reason, IEEE and the members of its Societies and subcommittees of the IEEE SA Board of Governors are not able to provide an instant response to comments or questions, except in those cases where the matter has previously been addressed. For the same reason, IEEE does not respond to interpretation requests. Any person who would like to participate in evaluating comments or revisions to an IEEE standard is welcome to join the relevant IEEE SA working group. You can indicate interest in a working group using the Interests tab in the Manage Profile & Interests area of the [IEEE SA myProject system](#).¹ An IEEE Account is needed to access the application.

Comments on standards should be submitted using the [Contact Us](#) form.²

¹ Available at: <https://development.standards.ieee.org/myproject-web/public/view.html#landing>.

² Available at: <https://standards.ieee.org/about/contact/>.

Laws and regulations

Users of IEEE Standards documents should consult all applicable laws and regulations. Compliance with the provisions of any IEEE Standards document does not constitute compliance to any applicable regulatory requirements. Implementers of the standard are responsible for observing or referring to the applicable regulatory requirements. IEEE does not, by the publication of its standards, intend to urge action that is not in compliance with applicable laws, and these documents may not be construed as doing so.

Data privacy

Users of IEEE Standards documents should evaluate the standards for considerations of data privacy and data ownership in the context of assessing and using the standards in compliance with applicable laws and regulations.

Copyrights

IEEE draft and approved standards are copyrighted by IEEE under U.S. and international copyright laws. They are made available by IEEE and are adopted for a wide variety of both public and private uses. These include both use by reference, in laws and regulations, and use in private self-regulation, standardization, and the promotion of engineering practices and methods. By making these documents available for use and adoption by public authorities and private users, neither IEEE nor its licensors waive any rights in copyright to the documents.

Photocopies

Subject to payment of the appropriate licensing fees, IEEE will grant users a limited, non-exclusive license to photocopy portions of any individual standard for company or organizational internal use or individual, non-commercial use only. To arrange for payment of licensing fees, please contact Copyright Clearance Center, Customer Service, 222 Rosewood Drive, Danvers, MA 01923 USA; +1 978 750 8400; <https://www.copyright.com/>. Permission to photocopy portions of any individual standard for educational classroom use can also be obtained through the Copyright Clearance Center.

Updating of IEEE Standards documents

Users of IEEE Standards documents should be aware that these documents may be superseded at any time by the issuance of new editions or may be amended from time to time through the issuance of amendments, corrigenda, or errata. An official IEEE document at any point in time consists of the current edition of the document together with any amendments, corrigenda, or errata then in effect.

Every IEEE standard is subjected to review at least every 10 years. When a document is more than 10 years old and has not undergone a revision process, it is reasonable to conclude that its contents, although still of some value, do not wholly reflect the present state of the art. Users are cautioned to check to determine that they have the latest edition of any IEEE standard.

In order to determine whether a given document is the current edition and whether it has been amended through the issuance of amendments, corrigenda, or errata, visit [IEEE Xplore](#) or [contact IEEE](#).³ For more information about the IEEE SA or IEEE's standards development process, visit the IEEE SA Website.

Errata

Errata, if any, for all IEEE standards can be accessed on the [IEEE SA Website](#).⁴ Search for standard number and year of approval to access the web page of the published standard. Errata links are located under the Additional Resources Details section. Errata are also available in [IEEE Xplore](#). Users are encouraged to periodically check for errata.

Patents

IEEE standards are developed in compliance with the [IEEE SA Patent Policy](#).⁵

Attention is called to the possibility that implementation of this standard may require use of subject matter covered by patent rights. By publication of this standard, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. If a patent holder or patent applicant has filed a statement of assurance via an Accepted Letter of Assurance, then the statement is listed on the IEEE SA Website at <https://standards.ieee.org/about/sasb/patcom/patents.html>. Letters of Assurance may indicate whether the Submitter is willing or unwilling to grant licenses under patent rights without compensation or under reasonable rates, with reasonable terms and conditions that are demonstrably free of any unfair discrimination to applicants desiring to obtain such licenses.

Essential Patent Claims may exist for which a Letter of Assurance has not been received. The IEEE is not responsible for identifying Essential Patent Claims for which a license may be required, for conducting inquiries into the legal validity or scope of Patents Claims, or determining whether any licensing terms or conditions provided in connection with submission of a Letter of Assurance, if any, or in any licensing agreements are reasonable or non-discriminatory. Users of this standard are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. Further information may be obtained from the IEEE Standards Association.

IMPORTANT NOTICE

Technologies, application of technologies, and recommended procedures in various industries evolve over time. The IEEE standards development process allows participants to review developments in industries, technologies, and practices, and to determine what, if any, updates should be made to the IEEE standard. During this evolution, the technologies and recommendations in IEEE standards may be implemented in ways not foreseen during the standard's development. IEEE standards development activities consider research and information presented to the standards development group in developing any safety recommendations. Other information about safety practices, changes in technology or technology implementation, or impact by peripheral systems also may be pertinent to safety considerations during implementation of the standard. Implementers and users of IEEE Standards documents are responsible for determining and complying with all appropriate safety, security, environmental, health, data privacy, and interference protection practices and all applicable laws and regulations.

³ Available at: <https://ieeexplore.ieee.org/browse/standards/collection/ieee>.

⁴ Available at: <https://standards.ieee.org/standard/index.html>.

⁵ Available at: <https://standards.ieee.org/about/sasb/patcom/materials.html>.

Participants

At the time this standard was completed, the Multimodal Conversation Working Group had the following membership:

Stephen Dukes, *Chair*
Miran Choi, *Vice Chair*
Yu Yuan, *Secretary*

<i>Organization Represented</i>	<i>Name of Representative</i>
0xSenses Corporation	Yu Yuan
Beijing Tsinghua Changung Hospital	Bin Yang
China University of Petroleum Beijing.....	Xiaohe Chen
Electronic and Telecommunications Research Institute (ETRI)	Xiaohan Liao
Imaginary Universes, LLC.....	Stephen Dukes
Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI)	Marina Bosi
Tongyuan Software Co., Ltd.....	Johnny Lin

The following members of the entity Standards Association balloting group voted on this standard. Balloters may have voted for approval, disapproval, or abstention.

0xSenses Corporation	Electronics and	Intel
China Electronic Standardization	Telecommunications Research	Moving Picture, Audio and Data
Institute	Institute (ETRI)	Coding by Artificial
CRRC Zhuzhou Institute Co.,	Guangdong Yunbai Technology	Intelligence (MPAI)
Ltd.	Co., Ltd.	PetroChina Planning and
Eaton	Imaginary Universes, LLC	Engineering Institute
	Institute of Biomedical	Tongyuan Software Co., Ltd.
	Engineering	

When the IEEE SA Standards Board approved this standard on 26 September 2024, it had the following membership:

David J. Law, *Chair*
Jon Walter Rosdahl, *Vice Chair*
Gary Hoffman, *Past Chair*
Alpesh Shah, *Secretary*

Sara R. Biyabani	Hao Hu	Paul Nikolich
Ted Burse	Yousef Kimiagar	Robby Robson
Stephen Dukes	Joseph L. Koepfinger*	Lei Wang
Doug Edwards	Howard Li	F. Keith Waters
J. Travis Griffith	Xiaohui Liu	Sha Wei
Guido R. Hiertz	John Haiying Lu	Philip B. Winston
Ronald W. Hotchkiss	Kevin W. Lu	Don Wright
	Hiroshi Mano	

*Member Emeritus



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

MPAI Technical Specification

Multimodal Conversation MPAI-MMC

V2

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	Introduction (Informative)	4
2	Scope of Standard	6
3	Terms and Definitions.....	8
4	References.....	10
4.1	Normative References.....	10
4.2	Informative References	11
5	Use Cases	11
5.1	General	11
5.2	Conversation with Personal Status (MMC-CPS).....	12
5.2.1	Scope of Conversation with Personal Status.....	12
5.2.2	Reference Architecture of Conversation with Personal Status	13
5.2.3	I/O Data of Conversation with Personal Status	14
5.2.4	Functions of AI Modules of Conversation with Personal Status	14
5.2.5	I/O Data of AI Modules of Conversation with Personal Status.....	14
5.2.6	JSON Metadata of Conversation with Personal Status.....	15
5.3	Conversation About a Scene (MMC-CAS)	15
5.3.1	Scope of Conversation About a Scene.....	15
5.3.2	Reference Architecture of Conversation About a Scene	16
5.3.3	I/O Data of Conversation About a Scene.....	17
5.3.4	Functions of AI Modules of Conversation About a Scene	17
5.3.5	I/O Data of AI Modules of Conversation About a Scene	18
5.3.6	JSON Metadata of Conversation About a Scene	18
5.4	Virtual Meeting Secretary (MMC-VMS)	19
5.4.1	Scope of Virtual Meeting Secretary.....	19
5.4.2	Reference Architecture of Virtual Meeting Secretary	19
5.4.3	I/O Data of Virtual Meeting Secretary.....	20
5.4.4	Functions of AI Modules of Virtual Meeting Secretary	20
5.4.5	I/O Data of AI Modules of Virtual Meeting Secretary	20
5.4.6	JSON Metadata of Virtual Meeting Secretary	21
5.5	Human-Connected Autonomous Vehicle (CAV) Interaction (MMC-HCI)	21
5.5.1	Scope of Human-CAV Interaction Subsystem	21
5.5.2	Reference Architecture of Human-CAV Interaction Subsystem.....	21
5.5.3	I/O Data of Human-CAV Interaction.....	23
5.5.4	Functions of AI Modules of Human-CAV Interaction	24
5.5.5	I/O Data of AI Modules of Human-CAV Interaction	24
5.5.6	JSON Metadata of Human-CAV Interaction	25
5.6	Conversation with Emotion (MMC-CWE).....	25
5.6.1	Scope of Conversation with Emotion	25
5.6.2	Reference Architecture of Conversation with Emotion.....	26
5.6.3	I/O Data of Conversation with Emotion	26
5.6.4	Functions of AI Modules of Conversation with Emotion.....	27
5.6.5	I/O Data of AI Modules of Conversation with Emotion.....	27
5.6.6	JSON Metadata of Conversation with Emotion.....	28
5.7	Multimodal Question Answering (MMC-MQA).....	28
5.7.1	Scope of Multimodal Question Answering.....	28

5.7.2	Reference Architecture of Multimodal Question Answering	28
5.7.3	I/O Data of Multimodal Question Answering	29
5.7.4	Functions of AI Modules of Multimodal Question Answering	29
5.7.5	I/O Data of AI Modules of Multimodal Question Answering	29
5.7.6	JSON Metadata of Multimodal Question Answering	30
5.8	Unidirectional Speech Translation (MMC-UST)	30
5.8.1	Scope of Unidirectional Speech Translation	30
5.8.2	Reference Architecture of Unidirectional Speech Translation	30
5.8.3	I/O Data of Unidirectional Speech Translation	31
5.8.4	Functions of AI Modules of Unidirectional Speech Translation	31
5.8.5	I/O Data of AI Modules of Unidirectional Speech Translation	32
5.8.6	JSON Metadata of Unidirectional Speech Translation	32
5.9	Bidirectional Speech Translation (MMC-BST)	32
5.9.1	Scope of Bidirectional Speech Translation	32
5.9.2	Reference Architecture of Bidirectional Speech Translation	32
5.9.3	I/O Data of Bidirectional Speech Translation	33
5.9.4	Functions of AI Modules of Bidirectional Speech Translation	33
5.9.5	I/O Data of AI Modules of Bidirectional Speech Translation	34
5.9.6	JSON Metadata of Bidirectional Speech Translation	34
5.10	One-to-Many Speech Translation (MMC-MST)	34
5.10.1	Scope of One-to-Many Speech Translation	34
5.10.2	Reference Architecture of One-to-Many Speech Translation	35
5.10.3	I/O Data of One-to-Many Speech Translation	35
5.10.4	Functions of AI Modules of One-to-Many Speech Translation	35
5.10.5	I/O Data of AI Modules of One-to-Many Speech Translation	35
5.10.6	JSON Metadata of One-to-Many Speech Translation	36
6	Composite AI Modules	36
6.1	Personal Status Extraction (MMC-PSE)	36
6.1.1	Scope of Personal Status Extraction	36
6.1.2	Reference Architecture of Personal Status Extraction	37
6.1.3	I/O Data of Personal Status Extraction	37
6.1.4	Functions of AI Modules of Personal Status Extraction	38
6.1.5	I/O Data of AI Modules of Personal Status Extraction	38
6.1.6	JSON Metadata of Personal Status Extraction and its AIMS	38
6.2	Speech and Text Translation (MMC-TST)	39
6.2.1	Functions of Speech and Text Translation	39
6.2.2	Reference Architecture of Text-and-Speech Translation	39
6.2.3	I/O Data of Text-and-Speech Translation	39
6.2.4	I/O Data of AI Modules of Text-and-Speech Translation	40
6.2.5	I/O Data of AI Modules of Text-and-Speech Translation	40
6.2.6	JSON Metadata of Speech-and-Text Translation and its AIMS	40
6.3	Personal Status Display (MMC-PSD)	40
6.3.1	Scope of Personal Status Display	40
6.3.2	Reference Architecture of Personal Status Display	41
6.3.3	I/O Data of Personal Status Display	41
6.3.4	JSON Metadata of Personal Status Display and its AIMS	42
6.4	Visual Object Identification (OSD-VOI)	42
6.4.1	Functions of Visual Object Identification	42
6.4.2	Reference Architecture of Visual Object Identification	42
6.4.3	Input/output data of Visual Object Identification	42

6.4.4	JSON Metadata of Visual Object Identification and its AIMS	43
7	Data Formats	43
7.1	Media	44
7.1.1	Audio File	44
7.1.2	Text	45
7.1.3	Video	45
7.1.4	Video File.....	45
7.2	Descriptors	45
7.2.1	Audio Scene Descriptors.....	45
7.2.2	Face Descriptors.....	45
7.2.3	Gesture Descriptors.....	45
7.2.4	Speech Descriptors.....	45
7.2.5	Speech Features.....	45
7.2.6	Text Descriptors	47
7.2.7	Visual Scene Descriptors	47
7.3	Personal Status	47
7.3.1	Factors and Modalities	47
7.3.2	Personal Status Data.....	49
7.3.3	Cognitive State.....	51
7.3.4	Emotion.....	53
7.3.5	Social Attitude.....	56
7.4	Miscellanea	61
7.4.1	Instance Identifier	61
7.4.2	Intention	62
7.4.3	Language Preference.....	63
7.4.4	Meaning	63
7.4.5	Spatial Attitude	65
7.4.6	Query Format of Video of Faces KB	65
	Annex 1 - MPAI Basics (Informative).....	66
1	General	66
2	AI Framework	66
3	Space, Objects, and Scenes	67
4	Avatar-Based Videoconference	67
5	Connected Autonomous Vehicle	68
	Annex 2 - MPAI-wide terms and definitions.....	71
	Annex 3 - Notices and Disclaimers Concerning MPAI Standards (Informative).....	74
	Annex 4 - Patent declarations (Informative).....	76
	Annex 5 - Personal Status (Informative).....	77
	Annex 6 - Communication Among AIM Implementors (Informative)	80

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, un-affiliated, 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 [1] enables dynamic configuration, initialisation, and control of AIWs in a standard environment called AI Framework (AIF). Figure 1 depicts the AI Framework.

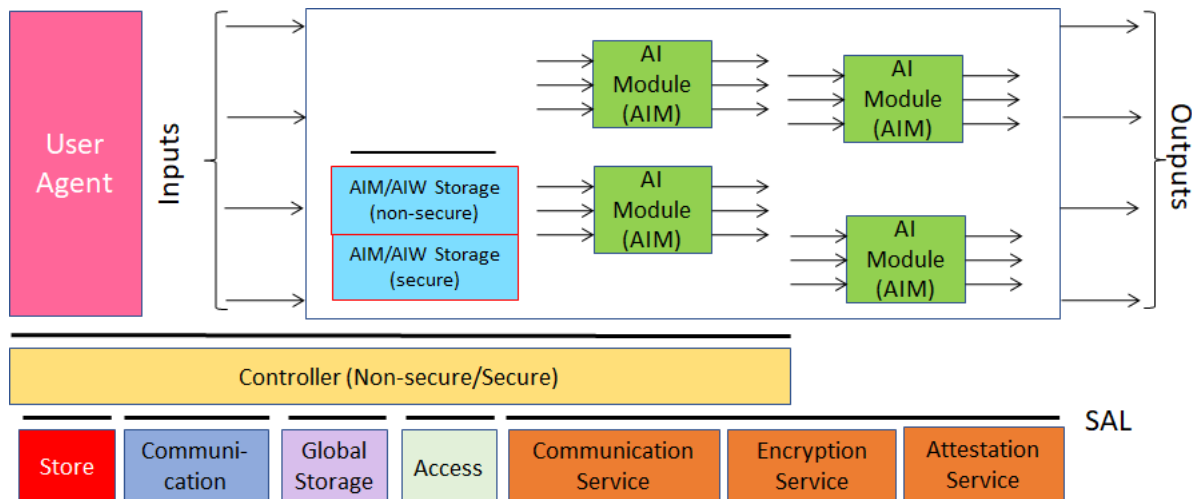


Figure 1 - The AI Framework (MPAI-AIF) V2 Reference Model

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.

MPAI Application Standards are designed to enable a User to obtain, via standard protocols, an Implementation of an AIW and of the set of corresponding AIMs and execute it in an AIF Implementation. The Store in *Figure 8* is the entity from which Implementations are downloaded. MPAI Standards assume that the AIF, AIW, and AIM Implementations may have been developed by independent implementers. A necessary condition for this to be possible, is that any AIF, AIW, and AIM implementations be uniquely identified. MPAI has appointed an ImplementerID Registration Authority (IIDRA) to assign unique ImplementerIDs (IID) to Implementers.¹

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 61* 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.

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

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	
- <i>Social</i>	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”.
- <i>Spatial</i>	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.
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.
- <i>Digital</i>	A Digitised or a Virtual Human in a Virtual Space.
- <i>Digitised</i>	An Object in a Virtual Space that has the appearance of a specific human when rendered.
- <i>Virtual</i>	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 Descriptor	An individual 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.
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 the MPai links below.

1. Technical Specification; AI Framework (MPai-AIF) V2; <https://mpai.community/standards/mpai-aif/>.
2. Technical Specification: Portable Avatar Format (MPai-PAF) V1.1; <https://mpai.community/standards/mpai-paf/>.
3. Technical Specification: Context-based Audio Enhancement (MPai-CAE) V2.1; <https://mpai.community/standards/mpai-cae/>.
4. Technical Specification: Connected Autonomous Vehicle (MPai-CAV) – Architecture V1; <https://mpai.community/standards/mpai-cav/>.
5. Technical Specification: Visual Object and Scene Description (MPai-OSD) V1; <https://mpai.community/standards/mpai-osd/>.

6. Technical Specification: Portable Avatar Format (MPAI-PAF) V1; <https://mpai.community/standards/mpai-paf/>.
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.

4.2 Informative References

The references provided here are for information purpose.

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/>.
19. Technical Specification; MPAI Ecosystem Governance (MPAI-GME) V1.1; <https://mpai.community/standards/mpai-gme/>.
20. Ekman, Paul (1999), "Basic Emotions", in Dalgleish, T; Power, M (eds.), Handbook of Cognition and Emotion (PDF), Sussex, UK: John Wiley & Sons.
21. Emotion Markup Language (EmotionML) 1.0; <https://www.w3.org/TR/2010/WD-emotionml-20100729/diffmarked.html>.
22. 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	<i>Media</i>	
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	<i>Descriptors</i>	
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	<i>Personal Status</i>	
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	<i>Miscellanea</i>	
7.4.1	Instance Identifier	MPAI-MMC
7.4.2	Intention	MPAI-MMC
7.4.3	Language Identifier	MPAI-MMC
7.4.4	Meaning	MPAI-MMC
7.4.5	Spatial Attitude	MPAI-MMC
7.4.6	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
2. Reference Architecture of the AIW
3. Input and Output Data of the AIW
4. Functions of the AIMs
5. Input and Output Data of the AIMs
6. JSON Metadata of the AIW and its AIMs.

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 Architecture 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.

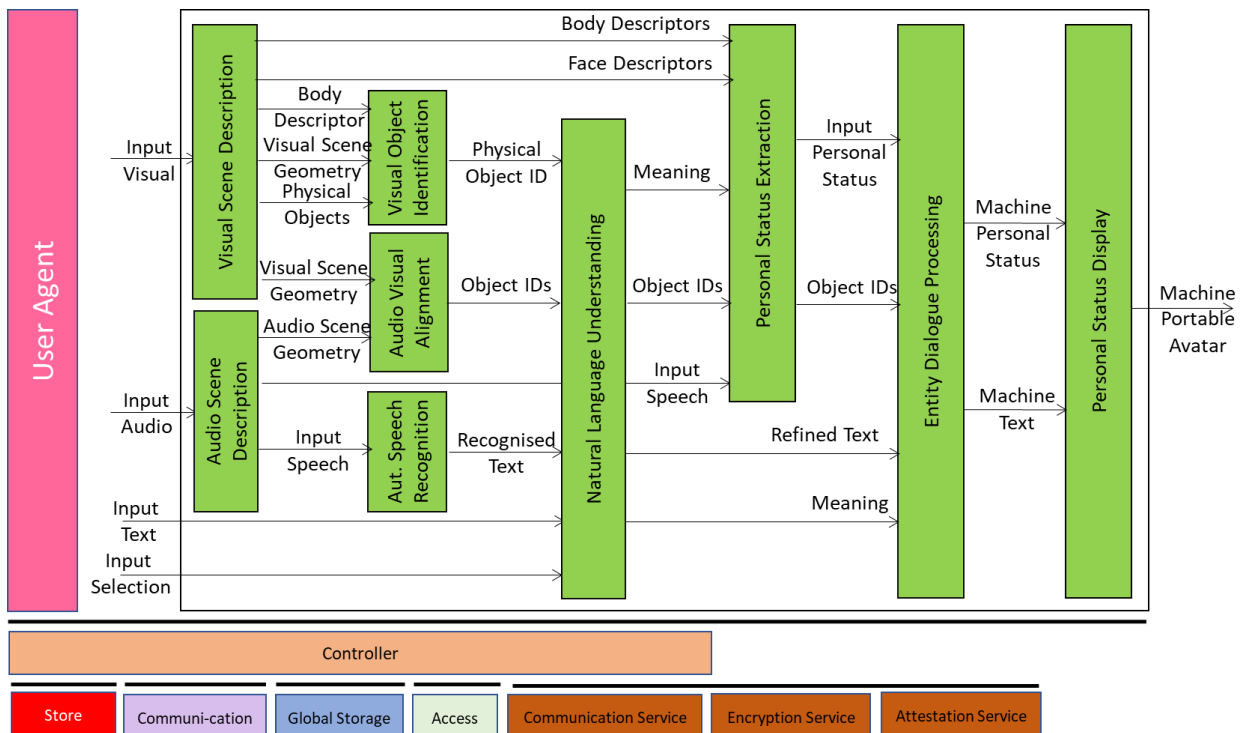


Figure 2 – Reference Model of Conversation with Personal Status

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 JSON Metadata of Conversation with Personal Status

Table 6 - Acronyms and URs of JSON Metadata

MMC-CPS	https://schemas.mpai.community/MMC/V2.1/AIW/ConversationWithPersonalStatus.json
OSD-VSD	https://schemas.mpai.community/OSD/V1.0/AIMs/VisualSceneDescription.json
CAE-ASD	https://schemas.mpai.community/CAE/V2.1/AIMs/AudioSceneDescription.json
OSD-VOI	https://schemas.mpai.community/OSD/V1.0/AIMs/VisualObjectIdentification.json
OSD-AVA	https://schemas.mpai.community/OSD/V1.0/AIMs/AudioVisualAlignment.json
MMC-ASR	https://schemas.mpai.community/MMC/V2.1/AIMs/AutomaticSpeechRecognition.json
MMC-NLU	https://schemas.mpai.community/MMC/V2.1/AIMs/NaturalLanguageUnderstanding.json
MMC-EDP	https://schemas.mpai.community/MMC/V2.1/AIMs/EntityDialogueProcessing.json
PAF-PSD	https://schemas.mpai.community/PAF/V1.1/AIMs/PersonalStatusDisplay.json

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 Architecture 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.

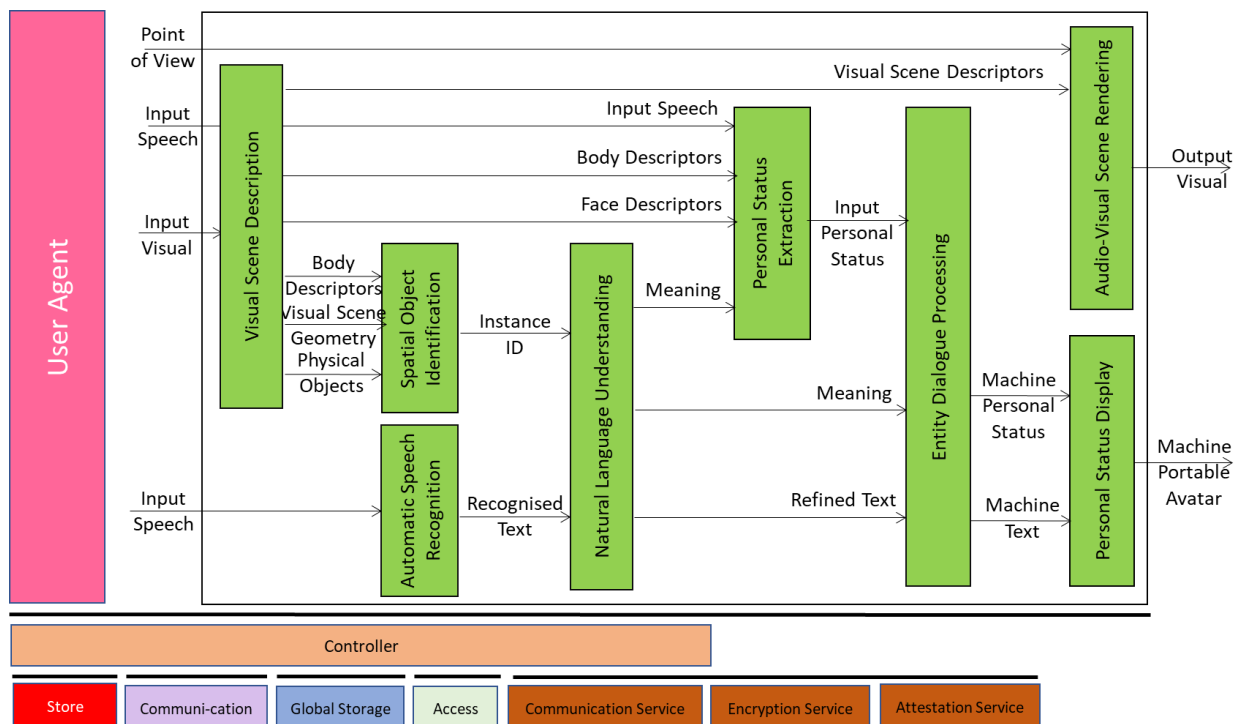


Figure 3 – Reference Model of Conversation About a Scene

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	Provides Visual Objects and their Spatial Attitudes.
Visual Object Identification	Provides the Visual Object Instance ID using Visual Scene Geometry and Visual Objects.
Automatic Speech Recognition	Recognises Speech.

Natural Language Understanding	Refines Text and extracts Meaning.
Personal Status Extraction	Extracts Personal Status from Meaning, Speech, Body, and Face.
Entity Dialogue Processing	1. Processes Refined Text and Personal Status. 2. Produces Machine's Text and Personal Status.
Audio-Visual Scene Rendering	Renders the Visual Scene as perceived by the Machine from the Point of View selected by human.
Personal Status Display	Provides Machine Portable Avatar from Machine Text and Machine Personal Status.

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 JSON Metadata of Conversation About a Scene

Table 10 - Acronyms and URs of JSON Metadata

MMC-CAS	https://schemas.mpai.community/MMC/V2.1/AIWs/ConversationAboutAScene.json
OSD-VSD	https://schemas.mpai.community/OSD/V1.0/AIMs/VisualSceneDescription.json
OSD-VOI	https://schemas.mpai.community/OSD/V1.0/AIMs/VisualObjectIdentification.json
MMC-ASR	https://schemas.mpai.community/MMC/V2.1/AIMs/AutomaticSpeechRecognition.json

MMC-NLU	https://schemas.mpai.community/MMC/V2.1/AIMs/NaturalLanguageUnderstanding.json
MMC-PSE	https://schemas.mpai.community/MMC/V2.1/AIMs/PersonalStatusExtraction.json
MMC-EDP	https://schemas.mpai.community/MMC/V2.1/AIMs/EntityDialogueProcessing.json
PAF-AVR	https://schemas.mpai.community/PAF/V1.1/AIMs/AudioVisualSceneRendering.json
PAF-PSD	https://schemas.mpai.community/PAF/V1.1/AIMs/PersonalStatusDisplay.json

5.4 Virtual Meeting Secretary (MMC-VMS)

5.4.1 Scope of Virtual Meeting Secretary

In the Avatar Based Videoconference Use Case of Portable Avatar Format [2], 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 4 of Annex 1 - MPAI Basics provides additional information on the Avatar-Based Videoconference Use Case.

5.4.2 Reference Architecture 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.

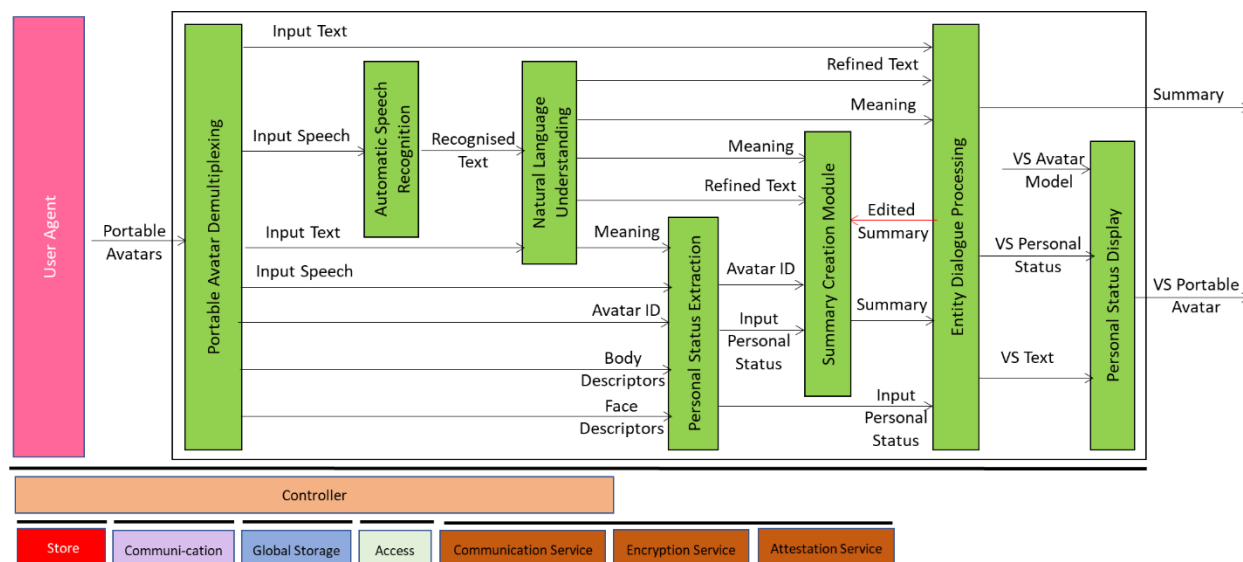


Figure 4 – Reference Model of the Virtual Meeting Secretary Use Case

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: