



IEEE Standard Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Multimodal Conversation Version 1.2

STANDARDS

Developed by the
Collaborative Activities Governance Board

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



IEEE Std 3300™-2022

IEEE Standard Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Multimodal Conversation Version 1.2

Developed by the

Collaborative Activities Governance Board
of the
IEEE SA Board of Governors

Approved 3 December 2022

IEEE SA Standards Board

Abstract: This standard adopts MPAA Technical Specification Version 1.2 as an IEEE Standard. Multimodal Conversation (MPAA-MMC) is an MPAA Standard comprising five use cases, all sharing the use of artificial intelligence (AI) to enable a form of human-machine conversation in completeness and intensity.

Keywords: 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 © 2023 by The Institute of Electrical and Electronics Engineers, Inc.
All rights reserved. Published 28 April 2023. Printed in the United States of America.

SECOND PRINTING: 9 October 2023. A title correction on the Front cover, Title page, Participants list, and Introduction.

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

The MPAA specification, Copyright ©MPAA 2021-2022. All rights reserved.

PDF: ISBN 978-1-5044-9330-7 STD25879
Print: ISBN 978-1-5044-9331-4 STDPD25879

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 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 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 this standard, including but not limited to the warranties of merchantability, fitness for a particular purpose and non-infringement. 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 his or her 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 development process involves the review of documents in English only. In the event that an IEEE standard is translated, only the English version published by IEEE is the approved IEEE standard.

Official statements

A statement, written or oral, that is not processed in accordance with the IEEE SA Standards Board Operations Manual 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, nor be relied upon as, a formal position of IEEE. 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).

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 in revisions to an IEEE standard is welcome to join the relevant IEEE 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

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.

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

² Available at: <https://standards.ieee.org/content/ieee-standards/en/about/contact/index.html>.

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 US 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.

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

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

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.

This IEEE standard is based on a technical specification developed by a consortium. Patent declarations may have been provided to the consortium as a membership requirement, or in compliance with the consortium's intellectual property rights (IPR) policies and procedures. Normative references in this standard may contain material subject to patent claims that are also subject to the consortium's or other organizations' IPR policies and procedures. Additional information about declarations may be available from the consortium's website:

MOVING PICTURE, AUDIO AND DATA CODING BY ARTIFICIAL INTELLIGENCE (MPAI)—
<https://mpai.community/standards/mpai-mmc/>

IMPORTANT NOTICE

IEEE Standards do not guarantee or ensure safety, security, health, or environmental protection, or ensure against interference with or from other devices or networks. 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, and interference protection practices and all applicable laws and regulations.

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*

<i>Organization Represented</i>	<i>Name of Representative</i>
0xSenses Corporation.....	Yu Yuan
1stCycle Corporation.....	Johnny Lin
Beijing Tsinghua Changgung Hospital.....	Bin Yang
Dreamerse.....	Stephen Dukes
Electronic and Telecommunications Research Institute (ETRI).....	Miran Choi
MPAI/Standford University	Marina Bosi

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	Institute of Biomedical Engineering, Chinese
1stCycle Corporation	Academy of Medical Sciences & Peking Union
Dreamerse	Medical Colledge
Electronics and Telecommunications Research Institute (ETRI)	PetroChina Planning and Engineering Institute

When the IEEE SA Standards Board approved this standard on 3 December 2022, it had the following membership:

David J. Law, *Chair*
Ted Burse, *Vice Chair*
Gary Hoffman, *Past Chair*
Konstantinos Karachalios, *Secretary*

Edward A. Addy	Johnny Daozhuang Lin	Mark Siira
Ramy Ahmed Fathy	Kevin Lu	Dorothy V. Stanley
J. Travis Griffith	Daleep C. Mohla	Lei Wang
Guido R. Hiertz	Andrew Myles	F. Keith Waters
Yousef Kimiagar	Damir Novosel	Karl Weber
Joseph L. Koepfinger*	Annette D. Reilly	Sha Wei
Thomas Koshy	Robby Robson	Philip B. Winston
John D. Kulick	Jon Walter Rosdahl	Daidi Zhong

*Member Emeritus

Introduction

This introduction is not part of IEEE Std 3300-2022, IEEE Standard Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Multimodal Conversation Version 1.2.

This IEEE standard is based on a technical specification developed by a consortium. Patent declarations may be provided to the consortium as a membership requirement, or in compliance with the consortium's intellectual property rights (IPR) policies and procedures. Normative references in this standard may contain material subject to patent claims that are also subject to the consortium's other organizations' IPR policies and procedures. Additional information about declaration may be available from consortium's website.

MOVING PICTURE, AUDIO AND DATA CODING BY ARTIFICIAL INTELLIGENCE (MPAI)—
<http://mpai.community/standards/mpai-mmvc/>.



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

MPAI Technical Specification

Multimodal Conversation MPAI-MMC

V1.2

WARNING

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

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

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

© Copyright MPAI 2021-2022. All rights reserved

Multimodal Conversation

V1.2

1	Introduction (Informative).....	4
2	Scope of Standard.....	5
2.1	Conversation With Emotion (CWE).....	6
2.2	Multimodal Question Answering (MQA)	6
2.3	Unidirectional Speech Translation (UST)	6
2.4	Bidirectional Speech Translation (BST).....	6
2.5	One-to-Many Speech Translation (MST).....	7
2.6	Normative content of the Use Cases.....	7
3	Terms and Definitions	7
4	References	8
4.1	Normative References	8
4.2	Informative References.....	9
5	Use Case Architectures	9
5.1	Conversation with Emotion (CWE)	9
5.1.1	Scope of Use Case.....	9
5.1.2	Input/Output Data.....	9
5.1.3	Implementation Architecture.....	9
5.1.4	AI Modules.....	10
5.1.5	AIW Metadata	11
5.2	Multimodal Question Answering (MQA)	11
5.2.1	Scope of standard	11
5.2.2	Input/output data	11
5.2.3	Implementation Architecture.....	11
5.2.4	AI Modules.....	12
5.2.5	AIW Metadata	12
5.3	Unidirectional Speech Translation (UST)	12
5.3.1	Scope of Use Case.....	12
5.3.2	Input/output data	13
5.3.3	Implementation Architecture.....	13
5.3.4	AI Modules.....	13
5.3.5	AIW Metadata	14
5.4	Bidirectional Speech Translation (BST).....	14
5.4.1	Scope of Use Case.....	14
5.4.2	Input/output data	14
5.4.3	Implementation Architecture.....	14
5.4.4	AI Modules.....	15
5.4.5	AIW Metadata	15
5.5	One-to-Many Speech Translation (MST).....	15
5.5.1	Scope of Use Case.....	15
5.5.2	Input/output data	16
5.5.3	Implementation Architecture.....	16
5.5.4	AI Modules.....	16
5.5.5	AIW Metadata	17

© Copyright MPAI 2021-2022. All rights reserved

6	AI Modules.....	17
6.1	AI Module Interoperability.....	17
6.2	MPAI-MMC AIMs and their data	17
6.2.1	Conversation with Emotion (CWE)	17
6.2.2	Multimodal Question Answering (MQA)	18
6.2.3	Unidirectional Speech Translation (UST).....	18
6.2.4	Bidirectional Speech Translation (BST)	19
6.2.5	One-to-Many Speech Translation (MST)	19
6.3	Data Formats.....	20
6.3.1	Emotion	20
6.3.2	Intention	24
6.3.3	Language identifier	26
6.3.4	Meaning.....	26
6.3.5	Object Identifier	28
6.3.6	Speech	29
6.3.7	Speech Features.....	29
6.3.8	Text.....	31
6.3.9	Text with Emotion.....	31
6.3.10	Video	32
6.3.11	Video File.....	32
6.3.12	Video of Faces KB Query Format.....	33
Annex 1	MPAI-wide terms and definitions.....	34
Annex 2	Notices and Disclaimers Concerning MPAI Standards (Informative)	37
Annex 3	Patent declarations	39
Annex 4	AIW and AIM Metadata of MMC-CWE.....	40
A4.1	AIW metadata for CWE	40
A4.2	AIM metadata	47
A4.2.1	SpeechRecognition.....	47
A4.2.2	Video Analysis	48
A4.2.3	Language Understanding.....	49
A4.2.4	Emotion Fusion	50
A4.2.5	Dialog Processing.....	51
A4.2.6	Speech Synthesis	53
A4.2.7	Lips Animation.....	54
Annex 5	AIW and AIM Metadata of MMC-MQA	56
A5.1	AIW metadata for MQA.....	56
A5.2	AIM metadata	61
A5.2.1	SpeechRecognition.....	61
A5.2.2	Video Analysis	62
A5.2.3	Language Understanding.....	63
A5.2.4	Question Analysis	64
A5.2.5	Question Answering.....	65
A5.2.6	Speech Synthesis (Text).....	66
Annex 6	AIW and AIM Metadata of MMC-UST	68
A6.1	AIW metadata for UST.....	68
A6.2	AIM metadata	72
A6.2.1	SpeechRecognition.....	72
A6.2.2	Translation.....	73
A6.2.3	Speech Feature Extraction.....	74

© Copyright MPAI 2021-2022. All rights reserved

A6.2.4 Speech Synthesis	75
Annex 7 AIW and AIM Metadata of MMC-BST	77
A7.1 AIW metadata for BST	77
A7.2 AIM metadata	82
A7.2.1 SpeechRecognition.....	82
A7.2.2 Translation.....	84
A7.2.3 Speech Feature Extraction.....	85
A7.2.4 Speech Synthesis	86
Annex 8 AIW and AIM Metadata of MMC-MST	88
A8.1 AIW metadata for MST	88
A8.2 AIM metadata	93
A8.2.1 SpeechRecognition.....	93
A8.2.2 Translation.....	94
A8.2.3 Speech Feature Extraction.....	95
A8.2.4 Speech Synthesis	96
Annex 9 Communication Among AIM Implementors (Informative)	98

1 Introduction (Informative)

In recent years, Artificial Intelligence (AI) and related technologies have been introduced in a broad range of applications, have started affecting the life of millions of people and are expected to do so even more in the future. As digital media standards have positively influenced industry and billions of people, so AI-based data coding standards are expected to have a similar positive impact. Indeed, research has shown that data coding with AI-based technologies is generally *more efficient* than with existing technologies for, e.g., compression and feature-based description.

However, some AI technologies may carry inherent risks, e.g., in terms of bias toward some classes of users. Therefore, the need for standardisation is more important and urgent than ever.

The international, unaffiliated, not-for-profit MPAI – Moving Picture, Audio and Data Coding by Artificial Intelligence Standards Developing Organisation has the mission to develop *AI-enabled data coding standards*. MPAI Application Standards enable the development of AI-based products, applications, and services.

As a rule, MPAI standards include four documents: Technical Specification, Reference Software Specifications, Conformance Testing Specifications, and Performance Assessment Specifications. The last type of Specification includes standard operating procedures to enable users of MPAI Implementations to make informed decision about their applicability based on the notion of Performance, defined as a set of attributes characterising a reliable and trustworthy implementation. In the following, Terms beginning with a capital letter are defined in *Table 1* if they are specific to this Standard and in *Table 21* if they are common to all MPAI Standards.

In general, MPAI Application Standards are defined as aggregations – called AI Workflows (AIW) – of processing elements – called AI Modules (AIM) – executed in an AI Framework (AIF). MPAI defines Interoperability as the ability to replace an AIW or an AIM Implementation with a functionally equivalent Implementation.

MPAI also defines 3 Interoperability Levels of an AIF that executes an AIW. The AIW and its AIMs may have 3 Levels:

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 offers Users access to the promised benefits of AI with a guarantee of increased transparency, trust and reliability as the Interoperability Level of an Implementation moves from 1 to 3. Additional information on Interoperability Levels is provided in reference [12].

Figure 1 depicts the MPAI-AIF Reference Model under which Implementations of MPAI Application Standards and user-defined MPAI-AIF Conforming applications operate.

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

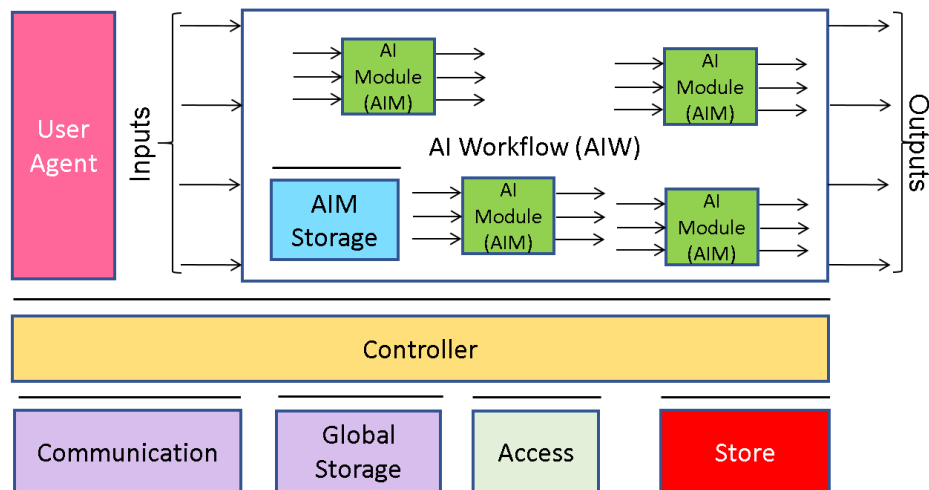


Figure 1 – The AI Framework (AIF) Reference Model and its Components

In particular, an AIM is defined by its Function and data, but not by its internal architecture, which may be based on AI or data processing, and implemented in software, hardware or hybrid software and hardware technologies.

MPAI 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 1 is an 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

A necessary condition to make possible the operations described in the paragraph above is the existence of an ecosystem composed of Conformance Testers, Performance Assessors, and an instance of the IIDRA and of the Store. Reference [12] provides an informative example of such ecosystem.

The chapters and the annexes of this Technical Specification are Normative, unless they are labelled as Informative.

2 Scope of Standard

Multimodal Conversation (MPAI-MMC) is an MPAI Standard comprising five Use Cases, all sharing the use of AI to enable a form of human-machine conversation that emulates human-human conversation in completeness and intensity:

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

1. “Conversation with Emotion” (CWE), supporting audio-visual conversation with a machine impersonated by a synthetic voice and an animated face.
2. “Multimodal Question Answering” (MQA), supporting request for information about a displayed object.
3. Three Use Cases supporting conversational 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:
 - a. “Unidirectional Speech Translation” (UST).
 - b. “Bidirectional Speech Translation” (BST).
 - c. “One-to-Many Speech Translation” (MST).

The current Version of MPAI-MMC has been developed by the MPAI Multimodal Conversation Development Committee (MM-DC). Future versions of the standard may extend the scope of the Use Cases and/or add new Use Cases in the scope of Multimodal Conversation.

MPAI expects to produce future MPAI-MMC Versions supporting enhanced current and new Use Cases.

2.1 Conversation With Emotion (CWE)

When humans have a conversation with other humans, they use speech and, in constrained cases, text. Their interlocutors may use video. Text, speech and video may convey the emotional state of a conversing human. Retaining “emotion” information in human-machine conversation is a key feature in understanding the meaning of speakers’ utterances. Therefore, a conversation system capable of recognising and conveying emotion can improve understanding of the human’s utterance and help the machine produce better replies.

The Conversation with Emotion (MMC-CWE) Use Case handles conversation with emotion. It offers a human-machine conversation system where the machine can recognise emotion in the human speech and/or text, while also using the information conveyed by video of the human’s face to produce a reply consistent with the human’s emotional state.

2.2 Multimodal Question Answering (MQA)

In a Question Answering (QA) System, a machine provides answers to a user’s question presented in natural language. Current QA systems, however, can only deal with inputs having the form of text or speech, while their effectiveness could be improved if the requesting human can present to the system also both speech and images. For example, a user might ask “Where can I buy this tool?” while showing the picture of the tool. In the Multimodal Question Answering (MMC-MQA) Use Case, a machine responds to a question expressed by a user in text or speech while showing an object. The machine’s response may use text and synthetic speech.

2.3 Unidirectional Speech Translation (UST)

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

2.4 Bidirectional Speech Translation (BST)

In the Bidirectional Translation (MMC-BST) Use Case, two people converse, each speaking a different language. The machine translates each input speech into the selected language as speech or text. If the desired output is speech, users can specify whether their speech features (voice colour, emotional charge, etc.) should be preserved in the translated speech.

© Copyright MPAI 2021-2022. All rights reserved

2.5 One-to-Many Speech Translation (MST)

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

2.6 Normative content of the Use Cases

Each Use Case normatively defines:

1. The Functions of the AIW and of the AIMs.
2. The Connections between and among the AIMs
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 Subsection of Section 6.3.
2. An *AIM*, shall:
 - a. Perform the AIM function specified by the appropriate section of Chapter 5.
 - b. Receive and produce the data specified in the appropriate Subsection of Section 6.1.
 - c. Receive as input and produce as output data having the format specified in Section 6.3.
3. A data *Format*, the data shall have the format specified in Section 6.3.

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 the notices and disclaimers of Annex 2.

3 Terms and Definitions

The terms used in this standard whose first letter is capital have the meaning defined in *Table 1*.

Table 1 – Table of terms and definitions

Term	Definition
Degree (of Emotion)	The intensity of an Emotion as one of “Low,” “Medium,” and “High.”
Emotion	An attribute indicating a member of a finite set of Emotions.
Emotion Recognition	An AIM that recognises the best member, or members, of a finite set of Emotions, as conveyed by a Text or Speech Segment or Video.
Face	A video representing human face
Image Analysis	An AIM that extracts pre-specified Features from videos.

© Copyright MPAI 2021-2022. All rights reserved

Intention	The result of analysis of the goal of an input question.
Language Understanding	An AIM that analyses a natural language Text and returns its Meaning and Emotions.
Meaning	Information – semantic, but also syntactic and structural – extracted from input data, i.e., Text, Speech, and Video.
Object Identifier	A unique ID enabling identification of a physical object, whether animate or inanimate, so that appropriate language and programs can refer to it.
Pitch	The fundamental frequency of Speech. Pitch is the attribute that makes it possible to judge sounds as "higher" and "lower."
Question Analysis	An AIM that analyses the Meaning of a question sentence and determines its Intention.
Question Answering	An AIM that analyses the user's question and produces a reply based on the user's Intention.
Speech 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.
Speech Rate	The number of Speech Units per second.
Speech Recognition	An AIM that converts Speech to Text.
Speech Synthesis	An AIM that converts Text or concept to Speech.
Speech Unit	Phoneme, syllable, or word as a segment of Speech.
Text	A collection of characters drawn from a finite alphabet.
Text Translation	An AIM that converts Text in an input language to Text in an output language.
Text with Emotion	Text marked up with, or associated with, labels or tags indicating emotional or related effects. In this standard, only one emotion is associated with a Text.
Vocal Gesture	Vocal but non-verbal elements of an utterance, e.g., laughs, coughs, hesitation syllables, etc.

4 References

4.1 Normative References

This standard normatively references the following documents, both from MPAI and other standards organisations. MPAI standards are publicly available at <https://bit.ly/30vp63g>.

1. MPAI Technical Specification; AI Framework (MPAI-AIF), V1
2. MPAI Technical Specification: Context-based Audio Enhancement (MPAI-CAE), V1
3. ISO 639; Codes for the Representation of Names of Languages — Part 1: Alpha-2 Code.
4. ISO/IEC 10646; Information technology – Universal Coded Character Set
5. ISO/IEC 14496-10; Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding
6. ISO/IEC 14496-12; Information technology – Coding of audio-visual objects – Part 12: ISO base media file format
7. ISO/IEC 23008-2; Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 2: High Efficiency Video Coding
8. ISO/IEC 23094-1; Information technology – General video coding – Part 1: Essential Video Coding

© Copyright MPAI 2021-2022. All rights reserved

9. MPAI; The MPAI Statutes; <https://mpai.community/statutes/>
10. MPAI; The MPAI Patent Policy; <https://mpai.community/about/the-mpai-patent-policy/>.
11. Framework Licence of the Multimodal Conversation Technical Specification (MPAI-MMC); <https://mpai.community/standards/mpai-mmc/framework-licence/>

4.2 Informative References

The references provided here are for information purpose.

12. MPAI Standard; The governance of the MPAI ecosystem (MPAI-GME), V1
13. Ekman, Paul (1999), "Basic Emotions", in Dalglish, T; Power, M (eds.), Handbook of Cognition and Emotion (PDF), Sussex, UK: John Wiley & Sons
14. Emotion Markup Language (EmotionML) 1.0; <https://www.w3.org/TR/2010/WD-emotionml-20100729/diffmarked.html>
15. 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 Case Architectures

5.1 Conversation with Emotion (CWE)

5.1.1 Scope of Use Case

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

5.1.2 Input/Output Data

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

Table 2 – I/O Data of Conversation with Emotion

Input	Comments
Input Selection	Data determining the use of Speech vs Text.
Input Text	Text typed by the human as additional information stream or as a replacement of the speech depending on the value of Input Selection.
Input Speech	Speech of the human having a conversation with the machine.
Input Video	Video of the Face of the human having a conversation with the machine.
Output	Comments
Output Text	Text of the Speech produced by the machine.
Output Speech	Synthetic Speech produced by the machine.
Output Video	Video of a Face whose lip movements are synchronised with the Output Speech and the synthetic machine emotion.

5.1.3 Implementation Architecture

The operation of Conversation with Emotion develops as follows:

© Copyright MPAI 2021-2022. All rights reserved

1. Emotion is recognised in the following way and later reflected in production of synthetic speech by the machine:
 - a. A set of Emotion-related cues are extracted from Input Text, Input Speech, and Input Video.
 - b. Each AIM (Language Understanding AIM, Speech Recognition AIM, and Video Analysis AIM) independently extracts Emotion from the corresponding input media.
 - c. The Emotion Fusion AIM fuses all Emotions into the Fused Emotion.
2. The Dialog Processing AIM produces a reply based on the Fused Emotion and Meaning from the text analysis.
3. The Speech Synthesis (Emotion) AIM produces Output Speech from Text with Emotion.
4. The Lips Animation AIM animates the lips of a Face drawn from the Video of Faces KB consistently with the Output Speech and the Output Emotion.

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

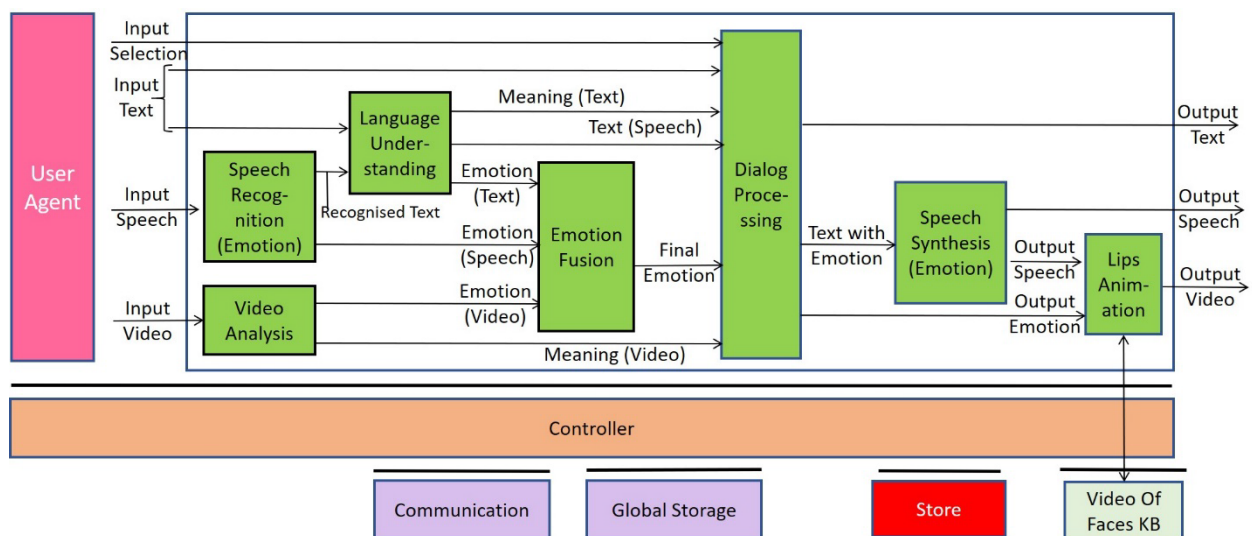


Figure 2 – Reference Model of Conversation With Emotion

5.1.4 AI Modules

The AI Modules of Conversation with Emotion perform the Functions specified in Table 3.

Table 3 – AI Modules of Conversation with Emotion

AIM	Function
Language Understanding	Extracts Meaning (Text) and Emotion (Text) included in Input Text containing natural language.
Speech Recognition	Analyses Input Speech and produces Recognised Text and Emotion (Speech).
Video Analysis	Extracts Emotion (Video) carried by Input Video.
Emotion Fusion	Determines the Fused Emotion from Emotion (Text), Emotion (Speech), and Emotion (Video).

Dialog Processing	Produces Output Emotion and Output Speech by analysing Meaning (Text), and Text (Speech) or Input Text, depending on the value of Input Selection.
Speech Synthesis	Produces Output Speech from Text with Emotion.
Lips Animation	Animates the lips of a video obtained by querying the Video Faces KB, using the Output Emotion.

5.1.5 AIW Metadata

Specified in Annex 4 Section A4.1.

5.2 Multimodal Question Answering (MQA)

5.2.1 Scope of standard

A human asks a question in natural language expressed as Text or Speech while showing an object to which the question refers. The machine responds to the question with Text and synthetic Speech.

5.2.2 Input/output data

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

Table 4 – I/O Data of Multimodal Question Answering

Input	Comments
Input Selection	Data determining the use of Speech or Text.
Input Text	Text typed by the human as a replacement for Input Speech.
Input Speech	Speech of the human asking the machine a question.
Input Video	Video of the human showing a held object.
Output	Comments
Output Text	The Text generated by MQA in response to human inputs.
Output Speech	The Speech generated by MQA in response to human inputs.

5.2.3 Implementation Architecture

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

1. Human asks a question in the form of text or speech depending on the value of the Input Selection.
2. Language Understanding extracts the Meaning of the question from Input Text or Input Speech depending on the value of Input Selection.
3. Video Analysis identifies the object.
4. Question Analysis determines the Intention of the question.
5. Question Answering uses Intention and Meaning to produce the answer as Output Text.
6. Speech Synthesis (Text) produces the Output Speech from Output Text.

Figure 3 depicts the input/output data, the AIMs and the data exchanged between the AIMs.

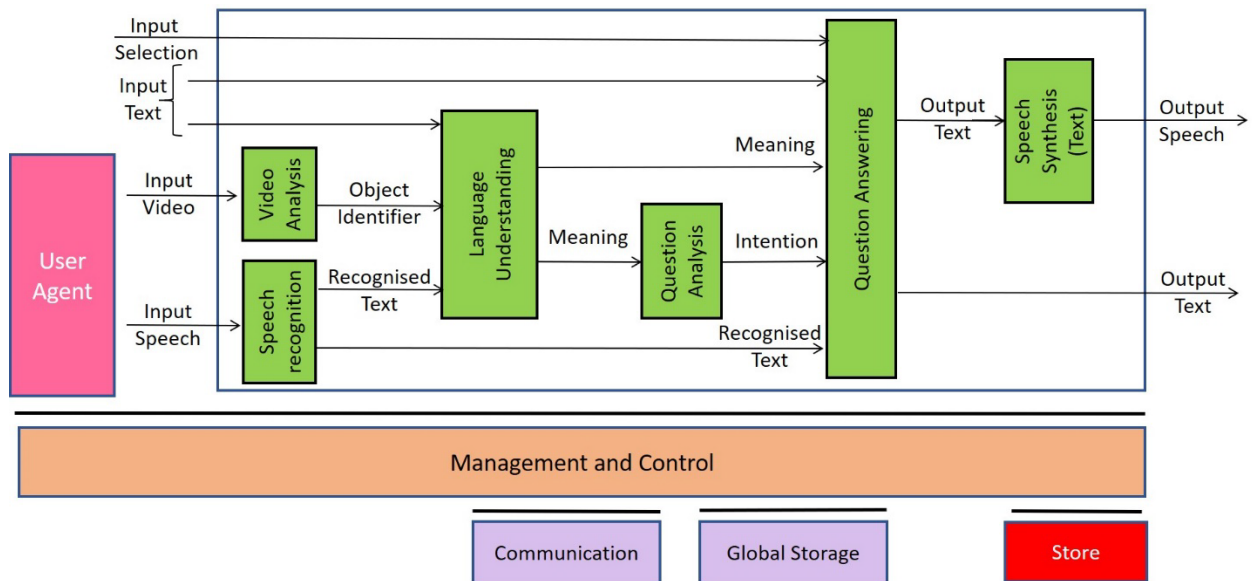


Figure 3 – Reference Model of Multimodal Question Answering

5.2.4 AI Modules

The AI Modules of Multimodal Question Answering are given in Table 5.

Table 5 – AI Modules of Multimodal Question Answering

AIM	Function
Video Analysis	Produces the name of the object in focus by analysing Input Video.
Speech Recognition	Produces Recognised Text by recognising Input Speech.
Language Understanding	Produces Meaning by analysing Input Text or Input Speech depending on the value of Input Selection.
Question Analysis	Determines Intention by analysing Meaning.
Question Answering	Produces Output Text by analysing Input Text or Recognised Text depending on the value of Input Selection, Intention and Meaning.
Speech Synthesis (Text)	Converts Output Text to Output Speech.

5.2.5 AIW Metadata

Specified in Annex 5 Section A4.1.

5.3 Unidirectional Speech Translation (UST)

5.3.1 Scope of Use Case

In Unidirectional Speech Translation (UST), Speech in Language A is translated into Speech in Language B. The flow of control is from Input Speech or Input Text to Translated Text, and then to Output Speech and Output Text. Depending on the value of Input Selection:

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

5.3.2 Input/output data

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

Table 6 – I/O Data of Unidirectional Speech Translation

Input	Comments
Input Selection	Determines whether 1. The input will be in Text or Speech 2. The Input Speech features are preserved in the Output Speech.
Requested Languages	User-specified input Language (A) and output Language (B).
Input Speech	Speech produced in Language A by a human desiring translation into language B.
Input Text	Alternative textual source information to be translated into and pronounced in language B depending on the value of Input Selection.
Output	Comments
Translated Speech	Input Speech translated into language B preserving the Input Speech features in the Output Speech, depending on the value of Input Selection.
Translated Text	Text of Input Speech or Input Text translated into language B, depending on the value of Input Selection.

5.3.3 Implementation Architecture

Figure 4 describes the input/output data, the AIMs and the data exchanged between AIMs.

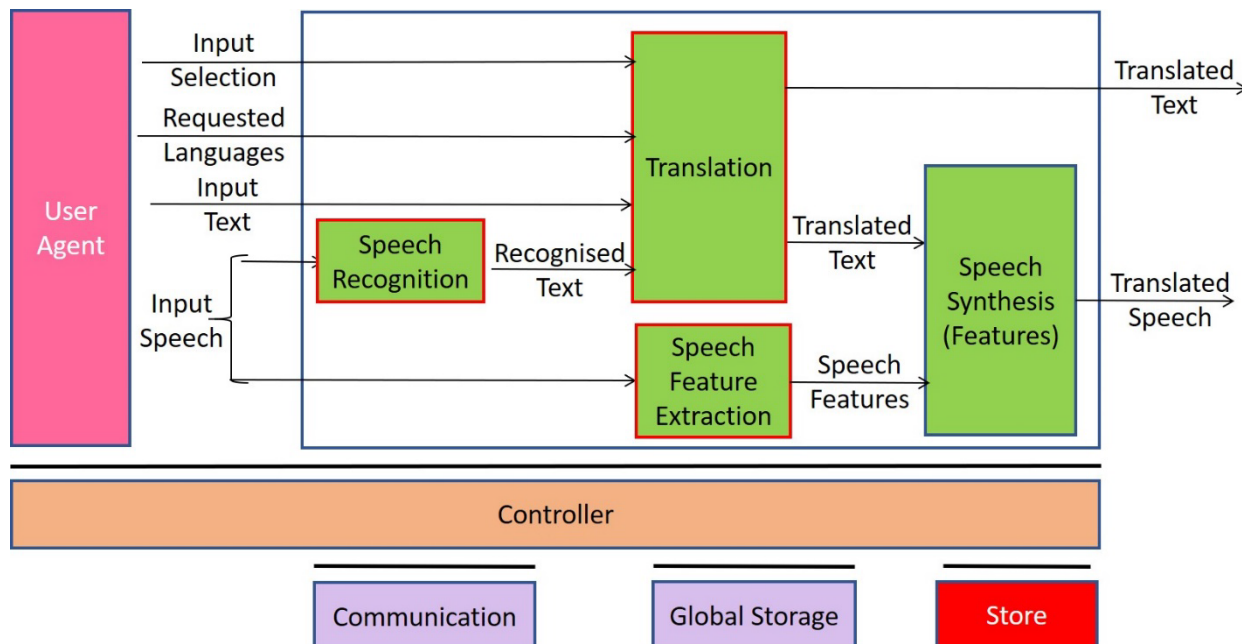


Figure 4 – Reference Model of Unidirectional Speech Translation (UST)

5.3.4 AI Modules

The AI Modules of Unidirectional Speech Translation are given in Table 7.

Table 7 – AI Modules of Unidirectional Speech Translation

© Copyright MPAI 2021-2022. All rights reserved

AIM	Function
Speech Recognition	Converts one Input Speech Segment into Recognised Text
Translation	Translates Input Text or Recognised Text (depending on the value of Input Selection) into the target language.
Speech Feature Extraction	Extracts Speech Features specific to the speaker (e.g., tones, intonation, intensity, pitch, emotion, speed) from Input Speech.
Speech Synthesis (Features)	Produces Output Speech from Translated Text using the Speech Features, depending on the value of Input Selection.

5.3.5 AIW Metadata

Specified in Annex 6 Section A4.1.

5.4 Bidirectional Speech Translation (BST)

5.4.1 Scope of Use Case

In Bidirectional Speech Translation, two people converse, one speaking language A and the other language B. The flow of control (from Input Speech to Translated Text to Output Speech) is identical to that of the Unidirectional case. The difference is that, rather than one such flow, two flows are provided in two different channels – the first from language A to language B, and the second from language B to language A.

Depending on the value of Input Selection

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

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

5.4.2 Input/output data

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

Table 8 – I/O Data of Bidirectional Speech Translation

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

5.4.3 Implementation Architecture

Figure 5 depicts the AIMs and the data exchanged between AIMs.

© Copyright MPAI 2021-2022. All rights reserved

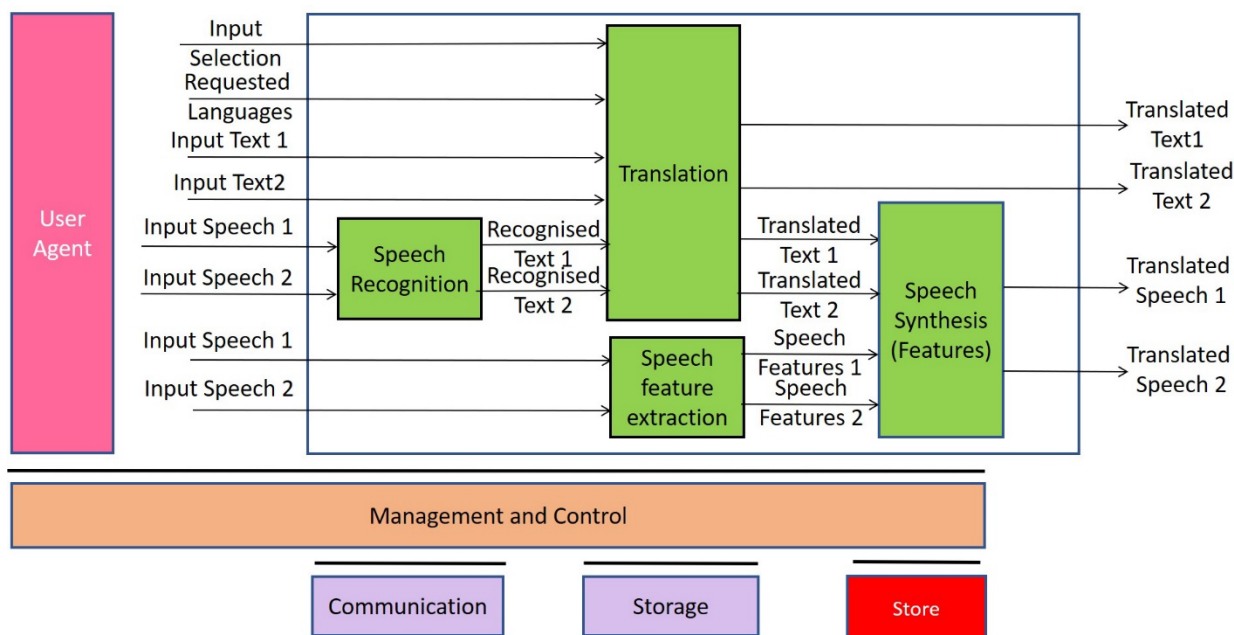


Figure 5 – Reference Model of Bidirectional Speech Translation (BST)

5.4.4 AI Modules

The AI Modules are given in Table 9.

Table 9 – AI Modules of Bidirectional Speech Translation

AIM	Function
Speech Recognition	Converts Input Speech 1 Segment and Input Speech 2 Segment into Recognised Text 1 and Recognised Text 2.
Translation	Translates Input Text1 and Input Text 2 (or Recognised Text 1 and Recognised Text 2) – depending on the value of Input Selection – into Translated Text 1 and Translated Text 2.
Speech Feature Extraction	Extracts Speech Features 1 and Speech Features 2 from Input Speech 1 and Input Speech 2.
Speech Synthesis (Features)	Produces Translated Speech 1 and Translated Speech 2 from Translated Text 1 and Translated Text 2, making use of the Speech Features extracted from the corresponding Input Speech segments, depending on the value of Input Selection.

5.4.5 AIW Metadata

Specified in Annex 7 Section A4.1.

5.5 One-to-Many Speech Translation (MST)

5.5.1 Scope of Use Case

In One-to-Many Speech Translation, any person speaking his or her preferred language broadcasts to two or more audience members, each listening, in a different language. The flow of control (from Recognised Text to Translated Text to Output Speech) is identical to that of the

© Copyright MPAI 2021-2022. All rights reserved

Unidirectional case. However, rather than one such flow, multiple paired flows are provided – the first pair from language A to language B and B to A; the second from A to C and C to A; and so on.

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

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

5.5.2 Input/output data

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

Table 10 – I/O Data of One-to-Many Speech Translation

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

5.5.3 Implementation Architecture

Figure 6 depicts the AIMs and the data exchanged between AIMs.

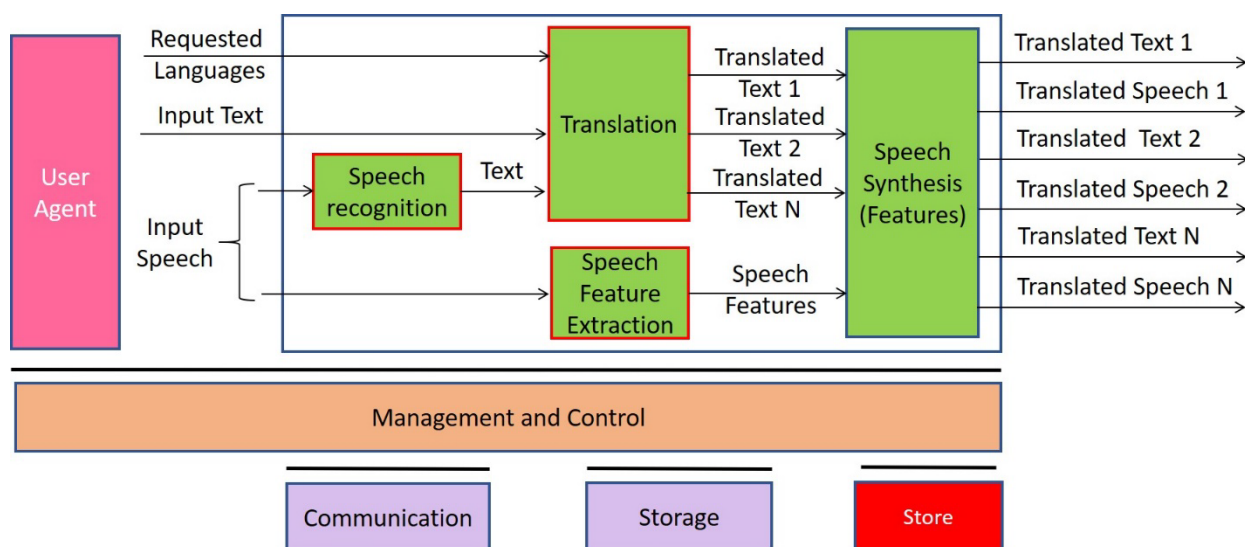


Figure 6 – Reference Model of One-to-Many Speech Translation (MST)

5.5.4 AI Modules

The AI Modules of Personalised Automatic Speech Translation are given in Table 11.

© Copyright MPAI 2021-2022. All rights reserved

Table 11 – AI Modules of One-to-Many Speech Translation

AIM	Function
Speech Recognition	Converts one Input Speech Segment into Recognised Text
Speech Feature Extraction	Extracts speaker-specific Speech Features from the Input Speech.
Translation	Translates one Text input into a set of Translated Texts in the Requested Languages.
Speech Synthesis (Features)	Uses the set of Translated Texts and the Speech Features to produce a set of Speech Segments in the Desired Languages.

5.5.5 AIW Metadata

Specified in Annex 8 Section A4.1.

6 AI Modules

This Chapter specifies the AIMs and their input and output data employed by all Use Cases specified in this Standard.

Section 6.1 contains general information about the meaning of Interoperability between two AIMs when the two neural networks are cascaded.

Section 6.2 lists the AIMs and their data in tabular form.

Section 6.3 specifies the Formats of the input and output data used in this Standard. The reader is alerted that some data Formats in this Standard are shared with the Context-based Audio Enhancement (MPAI-CAE) Standard [2]. At the current date, however, the specification of such data Formats is repeated verbatim in both Standards.

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

6.1 AI Module Interoperability

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

6.2 MPAI-MMC AIMs and their data

6.2.1 Conversation with Emotion (CWE)

Table 12 gives the AIMs and the input/output data of the Conversation with Emotion Use Case.

Table 12 – Conversation with Emotion AIMs and data formats

AIM	Input Data	Output Data
Video analysis	Input Video	Emotion (Video)

© Copyright MPAI 2021-2022. All rights reserved

Speech recognition	Input Speech	Recognised Text Emotion (Speech)
Language understanding	Input Text Recognised Text	Meaning (Text) Text (Speech) Emotion (Text)
Emotion Fusion	Emotion (Text) Emotion (Speech) Emotion (Video)	Fused Emotion
Dialog processing	Input Selection Input Text Meaning (Text) Text (Speech) Fused Emotion Meaning (Speech)	Output Text Text with Emotion Output Emotion
Speech Synthesis (Emotion)	Text with Emotion	Output Speech
Lips animation	Output Speech Output Emotion Video of Face	Output Video

The AIM Metadata are given in Annex 4 Section A4.2.

6.2.2 Multimodal Question Answering (MQA)

Table 13 gives the AIMs and the input/output data of the Multimodal Question Answering Use Case.

Table 13 – Multimodal Question Answering AIMs and data formats

AIM	Input Data	Output Data
Speech Recognition	Input Speech	Recognised Text
Video Analysis	Video	Object Identifier
Language understanding	Input Text Recognised Text	Meaning Meaning
Question analysis	Meaning	Intention
Question Answering	Input Selection Input Text Meaning Intention Recognised Text	Output Text
Speech Synthesis (Text)	Output Text	Output Speech

The AIM Metadata are given in Annex 5 Section A4.2.

6.2.3 Unidirectional Speech Translation (UST)

Table 14 gives the AIMs and the input/output data of the Unidirectional Speech Translation Use Case.

Table 14 – Unidirectional Speech Translation AIMs and data formats

© Copyright MPAI 2021-2022. All rights reserved

AIM	Input Data	Output Data
Speech Recognition	Input Speech	Recognised Text
Translation	Input Selection Requested Language Input Text	Translated Text
Speech feature extraction	Input Speech	Speech Features
Speech synthesis (Features)	Translated Text Speech features	Translated Speech

The AIM Metadata are given in Annex 6 Section A4.2.

6.2.4 Bidirectional Speech Translation (BST)

Table 15 gives the AIMs and the input/output data of the Bidirectional Speech Translation Use Case.

Table 15 – Bidirectional Speech Translation AIMs and data formats

AIM	Input Data	Output Data
Speech Recognition	Input Speech 1 Input Speech 2	Recognised Text 1 Recognised Text 2
Translation	Input Selection Requested Languages Input Text 1 Input Text 2	Translation Text1 Translation Text2
Speech feature extraction	Input Speech 1 Input Speech 2	Speech Features 1 Speech Features 2
Speech synthesis (Features)	Translation Text1 Translation Text2 Speech Features 1 Speech Features 2	Output Speech 1 Output Speech 2

The AIM Metadata are given in Annex 7 Section A4.2.

6.2.5 One-to-Many Speech Translation (MST)

Table 16 gives the input/output data of the One-to-many Speech Translation Use Case.

Table 16 – One-to-many Speech Translation AIMs and data formats

AIM	Input Data	Output Data
Speech Recognition	Input Speech	Recognised Text
Translation	Requested languages Input Text Input Speech	Translated Text 1 ... Translated Text N
Speech feature extraction	Input Speech	Speech Features
Speech synthesis (Features)	Text (Translation result) Speech Features	Translated Text 1 Translated Speech 1 ... Translated Text N Translated Speech N

© Copyright MPAI 2021-2022. All rights reserved

The AIM Metadata are given in Annex 8 Section 2.

6.3 Data Formats

The data Formats specified in this Technical Specification are listed in *Table 17*.

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

Table 17 – Data formats

Name of Data Format	Subsection	Use Case
Emotion	6.3.1	CWE
Intention	6.3.2	MQA
Language identifier	6.3.3	UST
		BST
		MST
Meaning	6.3.4	CWE
Object identifier	6.3.5	MQA
Speech	6.3.6	CWE
		MQA
		UST
		BST
		MST
Speech Features	6.3.7	UST
	6.3.7	UST
	6.3.7	UST
Text	6.3.8	CWE
		MQA
		UST
		BST
		MST
Text with Emotion	6.3.9	CWE
Video	6.3.10	CWE
Video File	6.3.11	ARP
Video Of Faces KB Query Format	6.3.12	CWE

6.3.1 Emotion

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

6.3.1.1 Syntax

Human Emotion is represented by.

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

© Copyright MPAI 2021-2022. All rights reserved

```

        "type": "number"
      },
      "emotionSetName": {
        "type": "string"
      }
    }
  },
  "type": "object",
  "properties": {
    "primary": {
      "$ref": "#/definitions/emotionType"
    },
    "secondary": {
      "$ref": "#/definitions/emotionType"
    }
  }
}

```

6.3.1.2 Semantics

Name	Definition
<i>emotionType</i>	Specifies the Emotion that the input carries.
<i>emotionDegree</i>	Specifies the Degree of Emotion as one of “Low,” “Medium,” and “High.”
<i>emotionName</i>	Specifies the ID of an Emotion listed in <i>Table 19</i> .
<i>emotionSetName</i>	Specifies the name of the Emotion set which contains the Emotion. Emotion set of <i>Table 19</i> is used as a baseline, but other sets are possible.

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.).

Table 18 gives the MPAI standardised three-level Basic Emotion Set partly based on Paul Eckman [12]:

1. The EMOTION CATEGORIES column specifies the categories using nouns.
2. The GENERAL ADJECTIVAL column gives adjectival labels for general or basic emotions within a category.

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.

Table 15 provides the semantics for each label in the GENERAL ADJECTIVAL and SPECIFIC ADJECTIVAL columns.

An Implementer wishing to extend or replace *Table 18* is requested to do the following:

1. Create a new *Table 18* where:
 - a. Proposed additions are clearly marked (in case of extension).
 - b. All Emotions and levels (up to 3) are listed (in case of replacement).
2. Create a new *Table 15* where:
 - a. the semantics of the Emotions is added to the semantics of the existing emotions (in case of extension).
 - b. is provided (in case of replacement).

© Copyright MPAI 2021-2022. All rights reserved

The semantics provided should have a level of details comparable to the semantics given in the current Table 15

3. Submit both tables to the MPAI Secretariat.

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

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

Table 18 – Basic Emotion Set

EMOTION CATEGORIES	GENERAL ADJECTIVAL	SPECIFIC ADJECTIVAL
ANGER	anger	furious irritated frustrated
APPROVAL, DISAPPROVAL	admiring/approving disapproving indifferent	awed contemptuous
AROUSAL	aroused/excited/energetic	cheerful playful lethargic sleepy
ATTENTION	attentive	expectant/anticipating thoughtful distracted/absent-minded vigilant hopeful/optimistic
BELIEF	credulous	Sceptical
CALMNESS	calm	peaceful/serene resigned
DISGUST	disgust	Loathing
FEAR	fearful/scared	terrified anxious/uneasy
HAPPINESS	happy	joyful content delighted amused
HURT	hurt jealous	
INTEREST	interested	fascinated curious bored
PRIDE/SHAME	proud ashamed	guilty/remorseful/sorry embarrassed
SADNESS	sad	lonely grief-stricken

© Copyright MPAI 2021-2022. All rights reserved

		discouraged depressed disappointed
SOCIAL DOMINANCE, CONFIDENCE	arrogant confident submissive	
SURPRISE	surprised	astounded startled
UNDERSTANDING	comprehending	uncomprehending bewildered/puzzled

Table 19 – Semantics of the Basic Emotion Set

ID	Emotion	Meaning
1	admiring/approving	emotion due to perception that others' actions or results are valuable
2	amused	positive emotion combined with interest (cognitive)
3	anger	emotion due to perception of physical or emotional damage or threat
4	anxious/uneasy	low or medium degree of fear, often continuing rather than instant
5	aroused/excited/energetic	cognitive state of alertness and energy
6	arrogant	emotion communicating social dominance
7	astounded	high degree of surprised
8	attentive	cognitive state of paying attention
9	awed	approval combined with incomprehension or fear
10	bewildered/puzzled	high degree of incomprehension
11	bored	not interested
12	calm	relative lack of emotion
13	cheerful	energetic combined with and communicating happiness
14	comprehending	cognitive state of successful application of mental models to a situation
15	confident	emotion due to belief in ability
16	contemptuous	high degree of disapproval
17	content	medium or low degree of happiness, continuing rather than instant
18	credulous	cognitive state of conformance to mental models of a situation
19	curious	interest due to drive to know or understand
20	delighted	high degree of happiness, often combined with surprise
21	depressed	high degree of sadness, continuing rather than instant, combined with lethargy (see AROUSAL)
22	disappointed	sadness due to failure of desired outcome
23	disapproving	not approving
24	discouraged	sadness combined with frustration
25	disgust	emotion due to urge to avoid, often due to unpleasant perception or disapproval
26	distracted/absent-minded	not attentive to present situation due to competing thoughts
27	embarrassed	shame due to consciousness of violation of social conventions

© Copyright MPAI 2021-2022. All rights reserved

28	expectant/anticipating	attentive to (expecting) future event or events
29	fascinated	high degree of interest
30	fearful/scared	emotion due to anticipation of physical or emotional pain or other undesired event or events
31	frustrated	angry due to failure of desired outcome
32	furious	high degree of anger
33	grief-stricken	sadness due to loss of an important social contact
34	guilty/remorseful/sorry	shame due to consciousness of hurting or damaging others
35	happy	positive emotion, often continuing rather than instant
36	hopeful/optimistic	expectation of good outcomes
37	hurt	emotion due to perception that others have caused social pain or embarrassment
38	indifferent	neither approving nor disapproving
39	interested	cognitive state of attentiveness due to salience or appeal to emotions or drives
40	irritated	low or medium degree of anger
41	jealous	emotion due to perception that others are more fortunate or successful
42	joyful	high degree of happiness, often due to a specific event
43	lethargic	not aroused
44	loathing	high degree of disgust
45	lonely	sadness due to insufficient social contact
46	peaceful/serene	calm combined with low degree of happiness
47	playful	energetic and communicating willingness to play
48	proud	emotion due to perception of positive social standing
49	resigned	calm due to acceptance of failure of desired outcome, often combined with low degree of sadness
50	sad	negative emotion, often continuing rather than instant, often associated with a specific event
51	sceptical	not credulous
52	sleepy	not aroused due to need for sleep
53	startled	surprised by a sudden event or perception
54	submissive	emotion communicating lack of social dominance
55	surprised	cognitive state due to violation of expectation
56	terrified	high degree of fear
57	thoughtful	attentive to thoughts
58	uncomprehending	not comprehending
59	vigilant	high degree of expectation or attentiveness

6.3.2 Intention

This subclause specifies data formats to describe intention, the outputs of Question analysis AIM. The “intention” consists of the following elements.

- qtopic
- qfocus
- qLAT
- qSAT