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

|  |  |
| --- | --- |
| N133 | 2021/01/20 |
| Source | MPAI Audio-Events-Data |
| Title | Draft MPAI-MMC Use Cases and Functional Requirements |
| Target | MPAI Members |

# Introduction

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

The MPAI approach to developing AI data coding standards is based on the definition of *standard interfaces* of *AI Modules (AIM).* AIMs operate on input data having a standard format to provide output data having a standard format. AIMs can be *combined* and *executed* in an MPAI-specified *AI-Framework* called MPAI-AIF. A [Call for MPAI-AIF Technologies](https://mpai.community/standards/mpai-aif/) [1] is currently open.

While AIMs must expose standard interfaces to be able operate in an MPAI AI Framework, their performance may differ depending on the technologies used to implement them. MPAI believes that *competing* developers striving to provide more performing *proprietary* and *interoperable* AIMs will promote *horizontal markets* of *AI solutions* that build on and further promote AI *innovation*.

This document is a collection of Use Cases and Functional Requirements for the MPAI Multimodal Conversation (MPAI-MMC) work area. The Use Cases in the MPAI-MMC standard enable human-machine conversation that emulates human-human conversation in com­pleteness and intensity. Currently MPAI has identified three Use Cases falling in the Multimodal Communication area:

1. Conversation with emotion (CWE)
2. Multimodal Question Answering (MQA)
3. Personalized Automatic Speech Translation (PST)

This document is to be read in conjunction with the MPAI-CAE Call for Technologies (CfT) [2] as it provides the functional requirements of all the technologies that have been identified as required to implement the current MPAI-MMC Use Cases. Respondents to the MPAI-MMC CfT should make sure that their responses are aligned with the functional requirements expressed in this document.

In the future, MPAI may issue other Calls for Technologies falling in the scope of MPAI-MMC to support identified Use Cases.

It should also be noted that some technologies identified in this document are the same, similar, or related to technologies required to implement some of the Use Cases of the companion document MPAI-CAE Use Cases and Functional Requirements [3]. Readers of this document are advised that being familiar with the content of the said companion document is a prerequisite for a proper understanding of this document.

This document is structured in 7 chapters, including this Introduction.

|  |  |
| --- | --- |
| Chapter 2 | briefly introduces the AI Framework Reference Model and its six Components |
| Chapter 3 | briefly introduces the 3 Use Cases. |
| Chapter 4 | presents the 4 MPAI-MMC Use Cases with the following structure1. Reference architecture
2. AI Modules
3. I/O data of AI Modules
4. Technologies and Functional Requirements
 |
| Chapter 5 | identifies the technologies likely to be common across MPAI-MMC and MPAI-CAE, a companion standard project whose Call for Technologies is issued simul­taneously with MPAI-MMC’s. |
| Chapter 6 | gives suggested references. Respondents are advised to become familiar with the references |
| Chapter 7 | gives a basic list of relevant terms and their definition |

# The MPAI AI Framework (MPAI-AIF)

Most MPAI applications considered so far can be implemented as a set of AIMs – AI, ML and even traditional Data Processing (DP)-based units with standard interfaces assembled in suitable topol­ogies to achieve the specific goal of an application and executed in an MPAI-defined AI Frame­work. MPAI is making all efforts to identify processing modules that are re-usable and upgradable without necessarily changing the inside logic. MPAI plans on completing the development of a 1st generation AI Framework called MPAI-AIF in July 2021.

The MPAI-AIF Architecture is given by *Figure 1*.



*Figure 1 – The MPAI-AIF Architecture*

Where:

1. *Management and Control* manages and controls the AIMs, so that they execute in the correct order and at the time when they are needed.
2. *Execution* is the environment in which combinations of AIMs operate. It receives external inputs and produces the requested outputs both of which are application specific interfacing with Management and Control and with Communication, Storage and Access.
3. *AI Modules* (AIM) are the basic processing elements receiving processing specific inputs and producing processing specific outputs.
4. *Communication* is required in several cases and can be implemented, e.g., by means of a service bus and may be used to connect with remote parts of the framework
5. *Storage* encompasses traditional storage and is used to e.g., store the inputs and outputs of the individual AIMs, data from the AIM’s state and intermediary results, shared data among AIMs.
6. *Access* represents the access to static or slowly changing data that are required by the application such as domain knowledge data, data models, etc.

# Use Cases

## Conversation with emotion

## Conversation with emotion

A human-machine conversation system where the computer can recognize emotion in the user’s speech to produce a reply. This MPAI-MMC Use Case handles conversation with emotion. When people talk, they use multiple modalities. Emotion is one of the key features to understand the meaning of the utterances made by the speaker. Therefore, a conversation system should have the capability to recognize emotion to understand the user’s speech and produce the reply as the output.

Emotion is recognised in the following way and reflected in the speech production side. First, a set of emotion related cues are extracted from text, voice and video. Then, each recognition module for text, voice and video, recognises emotion independently. The emotion recognition module determines the final emotion based on each emotion. The emotion will be transferred to dialog processing module. Then the dialog processing module produces the reply based on the final emotion and meaning from the text and video analysis. Finally, the speech synthesis module produces the speech from the reply in text.

### Multimodal Question Answering

Question Answering Systems (QA) answer a user’s question presented in natural language. Cur­rent QA system only deals with the case where input is in “text” form or “speech” form. However, more attention is paid these days to the case where mixed inputs such as speech with a image are presented to the system. For example, a user can ask a question about a picture which contains some specific tool as in “Where can I buy this tool?” showing the picture of the tool. In that case, the QA system should process the question in a text along with the image and should find out the answer to the question.

Question and image are recognised and analysed in the following way and answers are produced in the output speech: The meaning of the question is recognised in the form of text or voice. Image is analysed to find the object name which is sent to the language understanding module. Then, the integrated meaning from the multimodal inputs is generated from the language understanding module. The Intention analysis module determines the intention of the question and the intention is sent to the QA module. The QA module produces the answer based on the intention of the question, and the meaning from the Language understanding module. The speech synthesis module produces the speech from the answer in text.

### Personalised Automatic Speech Translation

Automatic speech translation technology denotes technology that recognizes a voice uttered in a language by a speaker, converts the recognized voice into another language through automatic translation, and outputs a converted voice as text-type subtitles or as a synthesized voice preserving the speaker’s features in the translated speech. Recently, as interest in voice synthesis among main technologies for automatic interpretation increases, research concentrates on personalized voice synthesis, a technology that outputs a target language through voice recognition and automatic translation, as a synthesis voice similar to a tone (or an utterance style) of a speaker.

The automatic interpretation system for generating a synthetic sound having characteristics similar to those of an original speaker’s voice includes a speech recognition to generate text data for an original speech signal of an original speaker and extract characteristic information such as pitch information, vocal intensity information, speech speed information, and vocal tract characteristic information of the original speech. Then the text data produced by the speech recognition module go through the automatic translation module to generate a synthesis-target translation and a speech synthesis module to generate a synthetic sound that resembles the original speaker using the extracted characteristic information.

# Functional Requirements

## Conversation With Emotion

### Implementation architecture

The architecture of *Figure 2* supports the case in which the user either uses or cannot use speech.

The Speech recognition and Language understanding AIMs required by this Use Case can be implemented either using AI-based or legacy technology. If these AIMs are implemented using AI technologies, access to the corresponding KB may not be needed.



*Figure 2 – Conversation with emotion*

### AI Modules

The AI Modules of Conversation with Emotion are given in *Table 1*

*Table 1 – AI Modules of Conversation With Emotion*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Language understanding** | Analyses natural language in a text format to produce its meaning and emotion included in the text |
| **Speech Recognition** | Analyses the voice input and generates text output and emotion carried by it |
| **Video analysis** | Analyses the video and recognises the emotion it carries |
| **Emotion recognition** | Determines the final emotion from multi-source emotions |
| **Dialog processing** | Analyses user’s utterance/text and produces Reply based on the meaning and emotion implied by the user’s text |
| **Speech synthesis** | Produces speech from Reply (the input text) |
| **Face animation** | Produces an animated face consistent with the Reply generated by the machine |
| **Emotion KB (text)** | Contains words/phases with associate emotion. Language understanding queries Emotion KB (text) to obtain the emotion associated with a text |
| **Emotion KB (speech)** | Contains features extracted from speech recordings of different speakers reading/reciting the same corpus of texts with an agreed set of emotions and without emotion, for a set of languages and for different genders.Speech recognition queries Emotion KB (speech) to obtain emotions corresponding to the features provided as input. |
| **Emotion KB (video)** | Contains features extracted from video recordings of different people speaking with an agreed set of emotions and without emotion for different genders.Video analysis queries Emotion KB (video) to obtain emotions corres­ponding to the features provided as input. |
| **Dialog KB** | Contains sentences with associated dialogue acts. Dialog processing queris Dialog KB to obtain dialogue acts with associated sentences. |

### I/O interfaces of AI Modules

The I/O data of Conversation with Emotion are given in *Table 2*.

*Table 2 – I/O data of Conversation With Emotion AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| **Language understanding** | Input TextRecognised TextResponse from Emotion KB (Text) | EmotionMeaning Query to Emotion KB (Text) |
| **Speech Recognition** | Input SpeechResponse from Emotion KB (Speech) | TextEmotionQuery to Emotion KB (Speech) |
| **Video analysis** | Digital video | Emotion |
| **Emotion recognition** | Emotion (from text)Emotion (from speech)Emotion (from image) | Final Emotion |
| **Dialog processing** | MeaningFinal emotionMeaningResponse from Dialogue KB | ReplyQuery to Dialogue KB |
| **Speech synthesis** | Reply | Speech |
| **Face animation** | Animation parameters | Video |
| **Emotion KB (text)** | Query | Response |
| **Emotion KB (speech)** | Query | Response |
| **Emotion KB (video)** | Query | Response |
| **Dialog KB** | Query | Response |

### Technologies and Functional Requirements

#### Digital Speech

Conversation with Emotion (CWS) requires that speech be sampled at a frequency between 22.05 kHz and 96 kHz and digitally represented between 16 bits/sample and 24 bits/sample.

**To Respondents**

Respondents are invited to comment on these two choices.

#### Digital Video

Digital video has the following features.

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

**To Respondents**

Respondents are invited to comment on these choices.

#### Emotion

By Emotion we mean an attribute that indicates an emotion out of a finite set of Emotions.

Emotion is extracted and digitally represented as Emotion from text, speech and video.

The most basic emotions are described by the set: “anger, disgust, fear, happiness, sadness, and surprise” [4], or “joy versus sadness, anger versus fear, trust versus disgust, and surprise versus anticipation” [5]. One of these sets can be taken as “universal” in the sense that they are common across all cultures. An Emotion may have different Grades [6,7].

**To Respondents**

Respondents are invited to propose

1. A minimal set of Emotions whose semantics are shared across cultures
2. A set of Grades that can be associated to Emotions
3. A digital representation of Emotions and their Grades [8]

This CfT does not specifically address culture-specific Emotions. However, the proposed digital representation of Emotions and their grades should either be capable to accommodate or be extensible to support culture-specific Emotions.

#### Speech features

Speech features are extracted from the input speech. Emotion of the input speech is determined based on the speech features.

Examples of features that have information about emotion are:

1. Features to detect the arousal level of emotions: sequences of short-time prosody acoustic features (features estimated on a frame basis), e.g., short-term speech energy [12].
2. Features related to the **pitch signal** (i.e., the glottal waveform) that depends on the tension of the vocal folds and the subglottal air pressure. Two parameters related to the pitch signal can be considered: pitch frequency and glottal air velocity. E.g., high velocity indicates a speech emotion like happiness. Low velocity is in harsher styles such as anger [14].
3. The shape of the vocal tract is modified by the emotional states. The formants (characterized by a center frequency and a bandwidth) could be a representation of the vocal tract resonances. Features related to the number of harmonics due to the non-linear airflow in the vocal tract. E.g., in the emotional state of anger, the fast air flow causes additional excitation signals other than the pitch. Teager Energy Operator-based (TEO) features, could be an example of measure of the harmonics and cross-harmonics in the spectrum [15].

An example solution of the features could be the Mel-frequency cepstrum (MFC) [16].

**To Respondents**

Respondents are requested to propose an extensible set of speech features that satisfy the following requirements

1. Be suitable for extracting Emotion information from natural speech containing Emotion.
2. Be suitable as input to query the Emotion (speech) KB

#### Emotion KB (speech) query format

Emotion KB (speech) contains features extracted from the speech recordings of different speakers reading/reciting the same corpus of texts with an agreed set of emotions and without emotion, for a set of languages and for different genders.

The Emotion KB (speech) is queried with a list of speech features. The Emotion KB responds with the emotions of the speech.

**To Respondents**

Respondents are requested to propose an Emotion KB (speech) query format that satisfies the following requirements

1. Capable of querying by specific speech features.
2. Extensible, i.e., capable to include additional speech features.

Note: An AI-based implementation may not need Emotion KB (Speech).

#### Text features

Text features considered are: grammatical features, e.g., part of speech; named entities, places, people, organisations; semantic features, e.g., roles, such as agent [18].

**To Respondents**

Respondents are requested to propose Text features satisfying the following requirements

1. Suitable for extracting Emotion information from natural language text containing Emotion.
2. Suitable as input to query the Emotion (text) KB.

#### Emotion KB (text) query format

Emotion KB (text) contains text features extracted from a text corpus with an agreed set of Emotions, for a set of languages and for different genders.

The Emotion KB (text) is queried with a list of text features. The Emotion KB (text) responds by giving emotions correlated with the text features provided as input.

**To Respondents**

Respondents are requested to propose an Emotion KB (text) query format that satisfies the fol­lowing requirements

1. Capable of querying by specific text features.
2. Extensible, i.e., capable to include additional text features.

Note: An AI-based implementation may not need Emotion KB (Text).

#### Video features

Video features are extracted from video for the purpose of querying the Emotion KB (Video).

**To Respondents**

Respondents are requested to propose Video features satisfying the following requirements

1. Suitable for extracting Emotion information from a video containing the face of a human expressing Emotion.
2. Suitable as input to query the Emotion (video) KB.

#### Emotion KB (video) query format

Emotion KB (video) contains features extracted from the video recordings of different speakers reading/reciting the same corpus of texts with an agreed set of emotions and without emotion, for for different genders.

Emotion KB (video) is queried with a list of video features. Emotion KB responds with the emotion of the video.

**To Respondents**

Respondents are requested to propose an Emotion KB (video) query format that satisfies the following requirements:

1. Capable of querying by specific video features.
2. Extensible, i.e., capable of including additional video features.

Note: An AI-based implementation may not need Emotion KB (video).

#### Input to speech synthesis

Respondents should propose suitable technology for driving the speech synthesiser. Here we consider “text with emotion to speech” and “concept to speech”.

**To Respondents**

*Text with emotion to speech*

A standard format for text with Emotions attached to different portions of the text. An example of how emotion in the text could be added to text is offered by emoticons.

Text should be encoded according to ISO/IEC 10646, Information technology – Universal Coded Character Set (UCS) to support most languages in use.

Respondents are requested to comment on the choice of character set and to propose a solution for emotion added to a text satisfying the following requirements

1. A scheme for annotating text with emotion should be proposed either as text with emotion expressed with text or with additional characters.
2. The emotion annotation representation scheme should include the basic emotions and be extensible.
3. The emotion annotation representation scheme should be language independent.

*Concept to speech*

Respondents are requested to propose technology that enables to go straight from meaning and emotion to “concept to speech synthesiser”, as in [25]. Therefore, we request digital representation of concept.

#### Meaning

Meaning is information extracted from the input text such as question, statement, exclamation, expression of doubt, request, invitation [18].

**To Respondents**

Respondents are requested to propose a solution for an extensible list of meanings and their digital representation satisfying the following requirements

1. The meaning extracted from the input text shall have a structure that includes grammatical information and semantic information.
2. The digital representation of meaning shall allow for the addition of new features to be used in different applications.

## Multimodal Question Answering

### Implementation Architecture

The architecture of *Figure 3* supports the case in which the user either uses or cannot use speech. Therefore, Text information is fed into Language understanding either through speech recognition or through text input by the user.

The Image analysis, Intention KB and Question Answering AIMs can be implemented either using AI or legacy technologies. If any of these AIMs are implemented as a neural network, access to the corresponding KB may not be needed.



*Figure 3* – *Multimodal Question Answering*

### AI Modules

The AI Modules of Multimodal Question Answering are given in *Table 3*.

*Table 3* *– AI Modules of Multimodal Question Answering*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Language understanding** | Analyses natural language expressed as text using a language model to produce the meaning of the text |
| **Speech Recognition** | Analyse the voice input and generate text output |
| **Speech synthesis** | Converts input text to speech  |
| **Image analysis** | Analyses image and produces the object name in focus |
| **Question analysis** | Analyses the meaning of the sentence and determines the Intention  |
| **Question Answering** | Analyses user’s question and produces a reply based on user’s Inten­tion  |
| **Intention KB** | Responds to queries using a question ontology to provide the features of the question |
| **Image KB** | Responds to Image analysis’s queries providing the object name in the image |
| **Online dictionary** | Allows Question Answering AIM to find answers to the question |

### I/O interfaces of AI Modules

The AI Modules of Multimodal Question Answering are given in *Table 4*.

*Table 4 – I/O data of* *Multimodal Question Answering AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| **Speech Recognition** | Digital Speech | Text |
| **Image analysis** | ImageImage KB response | Image KB queryText |
| **Language understanding** | TextText | Meaning Meaning |
| **Question analysis** | MeaningIntention KB response | Intention Intention KB query |
| **QA** | MeaningTextIntentionOnline dictionary query | TextOnline dictionary response |
| **Speech synthesis** | Text | Digital speech |
| **Intention KB** | Query | Response |
| **Image KB** | Query | Response |
| **Online dictionary** | Query | Response |
| **Dialog KB** | Query | Response |

### Technologies and Functional Requirements

#### Digital Speech

Multimodal QA (MQA) requires that speech be sampled at a frequency between 22.05 kHz and 96 kHz and digitally represented between 16 bits/sample and 24 bit/sample.

**To Respondents**

Respondents are invited to comment on these two choices.

#### Text

Text should be encoded according to ISO/IEC 10646, Information technology — Universal Coded Character Set (UCS) to support most languages in use.

**To Respondents**

Respondents are invited to comment on this choice.

#### Digital Image

A Digital image is an uncompressed or compressed picture. If compressed, the JPEG format should be used [19].

**To Respondents**

Respondents are invited to comment on this choice.

#### Image features

Image features are extracted from the input image representing an object [21].

A vector of image features extracted from object is used to identify the object.

**To Respondents**

Respondents are requested to propose a set of image features that satisfy the following requirements

1. Suitable for extracting the object name from an Image.
2. Suitable for querying a KB that contains representative object features.
3. Extensible to include objects to be added in the future.

#### Image KB query format

Image KB contains feature vectors extracted from different images of objects [26].

The Image KB is queried with a list of image features. The Image KB responds by giving the identifier of the object.

**To Respondents**

Respondents are requested to propose an Image KB query format that satisfies the following requirements

1. Capable of query by specific image features.
2. Extensible to include additional image features.

An AI-Based implementation may not need Image KB.

#### Object identifier

The object must be uniquely identified.

**To Respondents**

Respondents are requested to propose a universally applicable object classification scheme.

#### Meaning

Meaning is information extracted from the input text such as question, statement, exclamation, expression of doubt, request, invitation [18].

**To Respondents**

Respondents are requested to propose a solution for an extensible list of meanings and their digital representation satisfying the following requirements

1. The meaning extracted from the input text shall have a structure that includes grammatical information and semantic information.
2. The digital representation of meaning shall allow for the addition of new features to be used in different applications.

#### Intention

Intention is the result of the question analysis. For instance, what, where, for whom, how... [22]

**To Respondents**

Respondents are requested to propose an extensible classification of Intentions and their digital representation satisfying the following requirements

1. The intention of the question shall be represented as including question types, question focus and question topics.
2. The digital representation of intention shall be extensible, i.e., allow for the addition of new features to be used in different applications.

#### Intention KB query format

Intention KB contains features extracted from the user questions and the keywords that denote those intention types.

The Intention KB is queried by giving text as input. Intention KB responds with the type of question intention.

**To Respondents**

Respondents are requested to propose and Intention KB query format satisfying the following requirements

1. Capable of querying by specific question features.
2. Extensible, i.e., capable to include additional intention features.

An AI-Based implementation may not need Intention KB.

#### Online dictionary query format

Online dictionary contains structured data that include topics and related information in the form of summaries, table of contents and natural language text [23].

The Online dictionary is queried by giving text as input. The Online dictionary responds with paragraphs where to find answers that have high correlation with the user question.

**To Respondents**

Respondents are requested to propose an Online dictionary KB query format satisfying the following requirements

1. Capable of querying by text as keywords.
2. Extensible, i.e., capable to include additional text features.

## Personalized Automatic Speech Translation

### Implementation Architecture

The AI Modules implied by a personalized automatic speech translation system are configured as in *Figure 4*. This Use Case does not envisage the use of KBs.



*Figure 4* –*Personalized Automatic Speech Translation*

### AI Modules

The AI Modules of Personalized Automatic Speech Translation are given in *Table 5*.

*Table 5 – AI Modules of Personalized Automatic Speech Translation*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **Speech Recognition** | Converts Speech into Text |
| **Translation** | Translates the user text input in source language to the target language |
| **Speech feature extraction** | Extracts Speech features such as tones, intonation, intensity, pitch, emotion, intensity or speed from the input voice specific of the speaker. |
| **Speech synthesis** | Produces Speech from the text resulting from translation with the speech features extracted from the speaker of the source language |

### I/O interfaces of AI Modules

The AI Modules of Personalized Automatic Speech Translation are given in *Table 6*.

*Table 6 – I/O data of Personalized Automatic Speech Translation AIMs*

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input Data** | **Output Data** |
| **Speech Recognition** | Digital Speech | Text |
| **Image analysis** | ImageImage KB response | Image KB queryText |
| **Translation** | TextSpeech | Text |
| **Speech feature extraction** | Digital speech | Speech features |
| **Speech synthesis** | TextSpeech features | Digital speech |

### Technologies and Functional Requirements

#### Digital Speech

Personalized Automatic Speech Translation (PST) requires that speech be sampled at a frequency between 22.05 kHz and 96 kHz and digitally represented between 16 bits/sample and 24 bit/sample.

**To Respondents**

Respondents are invited to comment on these two choices.

#### Speech features

Speech features such as tones, intonation, intensity, pitch, emotion or speed are extracted by the speech extraction module. The speech features are used to encode speech features of the speaker.

The following features should be included in the speech features to describe the speaker’s voice: pitch, prosodic structures per intonation phrase, vocal intensity, speed of the utterance per word/sentence/intonation phrase, vocal tract characteristics of the speaker of the source language, and additional speech features associated with hidden variables. The vocal tract characteristics can be expressed as characteristic parameters of Mel-frequency cepstral coefficient (MFCC) and glottal wave.

**To Respondents**

Respondents are requested to propose a set of speech features that shall be suitable for

1. Extracting voice characteristic information from natural speech containing personal features.
2. Producing synthesized speech reflecting the original user’s voice characteristics.

#### Text

Text should be encoded according to ISO/IEC 10646, Information technology — Universal Coded Character Set (UCS) to support most languages in use.

**To Respondents**

Respondents are requested to comment on this choice.

#### Language identification

ISO 639 – Codes for the Representation of Names of Languages — Part 1: Alpha-2 Code.

**To Respondents**

Respondents are requested to comment on this choice.

# Potential common technologies

*Table 7* introduces the MPAI-CAE and MPAI-MMC acronyms.

*Table 7 – Acronyms of MPAI-CAE and MPAI-MMC Use Cases*

|  |  |  |
| --- | --- | --- |
| **Acronym** | **App. Area** | **Use Case** |
| EES | MPAI-CAE | Emotion-Enhanced Speech |
| ARP | MPAI-CAE | Audio Recording Preservation |
| EAE | MPAI-CAE | Enhanced Audioconference Experience |
| AOG | MPAI-CAE | Audio-on-the-go |
| CWE | MPAI-MMC | Conversation with emotion |
| MQA | MPAI-MMC | Multimodal Question Answering |
| PST | MPAI-MMC | Personalized Automatic Speech Translation |

*Table 8* gives all MPAI-CAE and MPAI-MMC technologies in alphabetical order.

Please note the following acronyms

|  |  |
| --- | --- |
| KB | Knowledge Base |
| QF  | Query Format |

*Table 8 – Alphabetically ordered MPAI-CAE and MPAI-MMC technologies*

|  |  |  |
| --- | --- | --- |
| **UC** | **Technology** | **Description** |
| AOG | Delivery | Speech transport format |
| EAE | Delivery | Speech transport format |
| AOG | Digital Audio | PCM Audio 48-96 kHz/16-24 bit |
| ARP | Digital Audio | PCM Audio 48-96 kHz/16-24 bit |
| ARP | Digital Image | A (un)compressed digital video frame |
| MQA | Digital Image | (un)compressed image |
| CWE | Digital Speech | PCM speech 22.05-96kHz/16-24 bit |
| EAE | Digital Speech | PCM speech 22.05-96kHz/16-24 bit |
| EES | Digital Speech | PCM speech 22.05-96kHz/16-24 bit |
| MQA | Digital Speech | PCM speech 22.05-96kHz/16-24 bit |
| PST | Digital Speech | PCM speech 22.05-96kHz/16-24 bit |
| ARP | Digital Video | Digital Video |
| CWE | Digital Video | Digital Video |
| CWE | Emotion | Digital representation of emotion |
| EES | Emotion | Digital representation of emotion |
| EES | Emotion descriptors | Derivations of Speech features |
| CWE | Emotion KB (speech) QF | Provides emotion from speech features |
| CWE | Emotion KB (text) QF | Provides emotion from text features |
| CWE | Emotion KB (video) QF | Provides emotion from video features |
| EES | Emotion KB QF | Provides Emotion descriptors |
| ARP | Image Features | Features characterising tape irregularities |
| MQA | Image features | Image features of object |
| MQA | Image KB QF | Provides object identifier |
| CWE | Input to speech synthesis | Plain text or concept |
| MQA | Intention | Information such as what, where, how |
| MQA | Intention KB QF | Provides Intention |
| PST | Language identification | Language identifier |
| CWE | Meaning | Information such as question, statement  |
| MQA | Meaning | Information such as question, statement |
| AOG | Microphone geometry information | Description of microphone position |
| EAE | Microphone geometry information | Description of microphone position |
| MQA | Object identifier | Identifier of a physical object  |
| MQA | Online dictionary QF | Provides paragraphs correlated with questions |
| EAE | Output device acoustic model metadata KB QF | Provides output device metadata |
| ARP | Packager | Audio/Video/Images/Text Multiplexer |
| AOG | Relevant vs non-relevant sound KB QF | Provides relevant sound |
| AOG | Sound array | Vector of extracted sounds |
| AOG | Sound categorisation KB QF | Provides sound category |
| AOG | Sounds categorisation | Identifier of a type of sound |
| EES | Speech and Emotion File Format | Multiplexed digital speech and emotion |
| CWE | Speech features | Speech features containing emotion info |
| EES | Speech features | Features associated to speech analysis |
| PST | Speech features | Features of input speech |
| ARP | Tape irregularity KB QF | Provides image features  |
| ARP | Text | Plain text |
| MQA | Text | Plain text |
| PST | Text | Plain text |
| CWE | Text features | Text features containing emotion info |
| AOG | User Hearing Profiles KB QF | Provides profile of identified user |
| CWE | Video features | Video features containing emotion info |

The following technologies are potentially applicable to different Use Cases.

*Table 9 – Technologies potentially shared by MPAI-CAE and MPAI-MMC*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Function** | **EES** | **ARP** | **EAE** | **AOG** | **CWE** | **MQA** | **PST** |
| Delivery |  |  | X | X |  |  |  |
| Digital speech | X |  | X |  |  |  |  |
| Digital audio |  | X |  | X |  |  |  |
| Digital image |  | X |  |  |  | X |  |
| Digital video |  | X |  |  | X |  |  |
| Emotion | X |  |  |  | X |  |  |
| Image features |  | X |  |  |  | X |  |
| Meaning |  |  |  |  | X | X |  |
| Microphone geometry information |  |  | X | X |  |  |  |
| Speech features  | X |  |  |  | X |  | X |
| Text |  | X |  |  | X | X | X |

The following technologies are shared or shareable across Use Cases:

1. Delivery
2. Digital speech
3. Digital audio
4. Digital image
5. Digital video
6. Emotion
7. Meaning
8. Microphone geometry information
9. Text

Image features apply to different visual objects. Speech features are different for all Use Cases.

However, respondents should consider the possibility of proposing a unified set of Speech features as proposed in [27]

# Terminology

*Table 10 –MPAI-MMC terms*

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Access | Static or slowly changing data that are required by an application such as domain knowledge data, data models, etc. |
| AI Framework (AIF) | The environment where AIM-based workflows are executed |
| AI Module (AIM) | The basic processing elements receiving processing specific inputs and producing processing specific outputs |
| Communication | The infrastructure that connects the Components of an AIF |
| Dialog processing | An AIM that |
| Digital Speech | Digitised speech as specified by MPAI |
| Emotion | An attribute that indicates an emotion out of a finite set of Emotions |
| Emotion Grade | The intensity of an Emotion |
| Emotion Recognition | An AIM that decides the final Emotion out of Emotions from different sources |
| Emotion KB (text) | A dataset of Text features |
| Emotion KB (speech) | A dataset of Speech features |
| Emotion KB (Video) | A dataset of Video features |
| Emotion KB query format | The format used to interrogate a KB |
| Execution | The environment in which AIM workflows are executed. It receives external inputs and produces the requested outputs both of which are application specific |
| Image analysis | An AIM that extracts Image features |
| Image KB | A dataset of Image features |
| Intention | Intention is the result of a question analysis |
| Intention KB | A question classification providing the features of a question |
| Language Understanding | An AIM that analyses natural language as Text to produce its meaning and emotion included in the text |
| Management and Control | Manages and controls the AIMs in the AIF, so that they execute in the correct order and at the time when they are needed |
| Meaning | Information extracted from the input text such as question, statement, exclamation, expression of doubt, request, invitation |
| Online Dictionary | A dataset that includes topics and related information in the form of summaries, table of contents and natural language text |
| Question Analysis | An AIM that analyses the meaning of a 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 | Features used to extract Emotion from Digital Speech |
| Speech feature extraction | An AIM that extracts Speech features from Digital speech |
| Speech Recognition  | An AIM that converts Digital speech to Text |
| Speech Synthesis | An AIM that converts Text or concept to Digital speech |
| Storage | Storage used to, e.g., store the inputs and outputs of the individual AIMs, data from the AIM’s state and intermediary results, shared data among AIMs |
| Text | A collection of characters drawn from a finite alphabet |
| Translation | An AIM that converts Text in a language to Text in another language |

# References

1. MPAI-AIF Call for Technologies; <https://mpai.community/standards/mpai-aif/#Technologies>
2. MPAI-MMC Call for Technologies
3. MPAI-CAE Use Cases and Functional Requirements
4. Ekman, P. (1999). Basic Emotions. In T. Dalgleish and T. Power (Eds.) The Handbook of Cognition and Emotion Pp. 45–60. Sussex, U.K.: John Wiley & Sons, Ltd.
5. Plutchik R., Emotion: a psychoevolutionary synthesis, New York Harper and Row, 1980
6. Russell, James (1980). "A circumplex model of affect". Journal of Personality and Social Psychology. 39 (6): 1161–1178. doi:10.1037/h0077714
7. Cahn, J. E., The Generation of Affect in Synthesized Speech, Journal of the American Voice I/O Society, 8, July 1990, p. 1-19
8. https://www.w3.org/TR/2014/REC-emotionml-20140522/
9. Burkhardt, F., & Sendlmeier, W. F., Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis, ISCA Workshop on Speech & Emotion, Northern Ireland 2000, p. 151-156.
10. Scherer, K. R., Ladd, D. R., & Silverman, K., Vocal cues to speaker affect: Testing two models, Journal of the Acoustic Society of America, 76(5), 1984, p. 1346-1356
11. Kasuya, H., Maekawa, K., & Kiritani, S., Joint Estimation of Voice Source and Vocal Tract Parameters as Applied to the Study of Voice Source Dynamics, ICPhS 99, p. 2505-2512
12. Mozziconacci, S. J. L., Speech Variability and Emotion: Production and Perception, PhD Thesis, Technical University Eindhoven, 1998
13. Burkhardt, F., & Sendlmeier, W. F., Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis, ISCA Workshop on Speech & Emotion, Northern Ireland 2000, p. 151-156.
14. Cahn, J. E., The Generation of Affect in Synthesized Speech, Journal of the American Voice I/O Society, 8, July 1990, p. 1-19
15. Hamed Beyramienanlou, Nasser Lotfivand, "An Efficient Teager Energy Operator-Based Automated QRS Complex Detection", Journal of Healthcare Engineering, vol. 2018, Article ID 8360475, 11 pages, 2018. https://doi.org/10.1155/2018/8360475]
16. Davis S B. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 1980, 28(4):65-74
17. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, pp. 3501–3504, May 2014. 2- Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition Journal, Elsevier, 44 (2011) 572–587
18. Mohamed Zakaria Kurdi (2017). Natural Language Processing and Computational Linguistics: semantics, discourse, and applications, Volume 2. ISTE-Wiley.
19. Semaan, P. (2012). Natural Language Generation: An Overview. Journal of Computer Science & Research (JCSCR)-ISSN, 50-57
20. Hudson, Graham; Léger, Alain; Niss, Birger; Sebestyén, István; Vaaben, Jørgen (31 August 2018). "JPEG-1 standard 25 years: past, present, and future reasons for a success". Journal of Electronic Imaging. 27 (4)
21. Hobbs, Jerry R.; Walker, Donald E.; Amsler, Robert A. (1982). "Natural language access to structured text". Proceedings of the 9th conference on Computational linguistics. 1. pp. 127–32.
22. MMP Petrou, C Petrou, Image processing: the fundamentals – 2010, Wiley
23. Suman Kalyan Maity, Aman Kharb, Animesh Mukherjee, Language Use Matters: Analysis of the Linguistic Structure of Question Texts Can Characterize Answerability in Quora, ICWSM 2017
24. [Xanh Ho](https://arxiv.org/search/?searchtype=author&query=Ho%2C+X), [Anh-Khoa Duong Nguyen](https://arxiv.org/search/?searchtype=author&query=Nguyen%2C+A+D), [Saku Sugawara](https://arxiv.org/search/?searchtype=author&query=Sugawara%2C+S), [Akiko Aizawa](https://arxiv.org/search/?searchtype=author&query=Aizawa%2C+A), Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps, COLING 2020
25. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.433.7322&rep=rep1&type=pdf
26. Mohamed Elgendy, Deep Learning for Vision Systems, Manning Publication, 2020
27. Problem Agnostic Speech Encoder; https://github.com/santi-pdp/pase