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

|  |  |
| --- | --- |
| N97 | 2020/12/16 |
| Source | General Assembly #3 (MPAI-3) |
| Title | Use Case-Requirements-candidate technologies for MPAI-MMC CfT |
| Target | MPAI Members |

# Summary

This document deals the following MPAI-MMC issues

1. Use Cases
   1. Conversation with emotion
   2. Multimodal Question Answering
   3. Personalized Automatic Speech Translation
2. Candidate technologies to appear in the expected MPAI-MMC Call for Technologies
3. Initial requirements of the said candidate technologies.

# 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)* that can be combined and *executed* in an MPAI-specified *AI-Framework* that MPAI calls MPAI-AIF.

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

This document includes 4 additional chapters

|  |  |
| --- | --- |
| Chapter 2 | introduces the The MPAI AI Framework. |
| Chapter 3 | presents the 4 MPAI-MMC Use Cases with the following structure   1. Use Case description 2. Implementation architecture 3. AI Modules and External data 4. Workflow 5. Functions and potential technologies for CfT 6. An initial analysis of the requirements associated to each identified technology. |
| Chapter 4 | presents the technologies likely to be common across MPAI-CAE and MPAI-MMC. |
| Chapter 5 | gives information on the next steps of the process that will lead to the development of the MPAI-MMC standard. |

# 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-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
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.

# Initial requirements of MPAI-MMC candidate technologies

## Introduction

So far, MPAI has identified the following Multimodal Conversation Use Cases benef­iting from MPAI standardisation:

1. **Conversation with emotion**: a human-machine conversation system where the computer can recognize emotion in the user’s speech to produce a reply
2. **Multimodal Question Answering**: a human-machine Question Answering system where the human asks questions to the computer presenting an image
3. **Personalized Automatic Speech Translation**: a system 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.

## Conversation with emotion

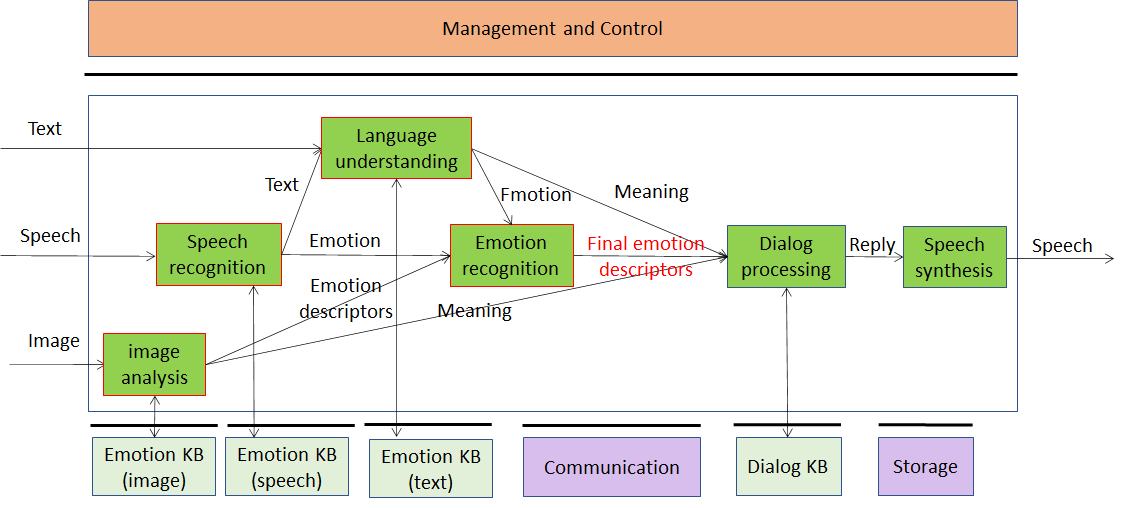
### Use Case description

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: speech, facial expression, text, sign languages and gesture. 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.

### Implementation architecture

The architecture of *Figure 2* supports the case in which the user can or cannot use speech. Text information is fed into Language understanding through either speech recognition or text input by the user.

The following AIMs can be implemented either as AI or legacy modules: Image analysis, Speech recognition, Language understanding and dialog KB. If any of these AIMs are implemented as a neural network, access to the corresponding KB may not be needed.



*Figure 2 – Conversation with emotion*

### AI Modules and External data

|  |  |
| --- | --- |
| **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 |
| **Image analysis** | Analyses image and produces the emotion it carries |
| **Emotion recognition** | Determines the final emotion from multi-source emotions |
| **Dialog processing** | Analyses user’s utterance and produces a reply based on user’s intention and emotion |
| **Speech synthesis** | Produces speech from the input text |
| **Emotion KB (text)** | Contains words/phases with associate emotion.  Is queried by Language understanding 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  Is queried by Speech recognition to obtain frequency-domain and time-domain features correlated to the emotions of the speech. |
| **Emotion KB (image)** | Contains image descriptors with associate emotion.  Is queried by Image analisis to obtain the emotion associated with an image |
| **Dialog KB** | Contains sentences with associated dialogue acts  I queried by Dialog processing to obtain dialogue acts with associated sentences |

### Workflow

1. Speech enters Speech recognition
2. Speech recognition
   1. Queries Emotion KB (Speech)
   2. Computes Emotion descriptors
   3. Sends Emotion descriptors to Emotion recognition
   4. Produces text
   5. Sends text to Language understanding
3. Text enters Language understanding (if user cannot speak)
4. Language understanding
   1. Queries Emotion KB (Text)
   2. Produces Meaning and Emotion descriptors
   3. Sends Emotion descriptors to Emotion recognition
   4. Sends Meaning to Dialog processing
5. Image enters Image analysis
6. Image analysis
   1. Computes Image features
   2. Queries Emotion KB (Image)
   3. Computes Emotion descriptors
   4. Sends Emotion descriptors to Emotion recognition
   5. Sends Meaning to Dialog processing
7. Emotion recognition
   1. Computes Final emotion descriptors
   2. Sends Final emotion descriptors to Dialog processing
8. Dialog processing
   1. Queries dialog KB
   2. Computes Reply (text and Emotion descriptors)
   3. Sends Reply to Speech synthesis
9. Speech synthesi module
   1. Produces Speech
   2. Outputs Speech

### Functions and potential technologies for CfT

|  |  |
| --- | --- |
| **Function** | **Potential CfT items** |
| Digital Speech | 16-24 bit/s, 22.05-96 kHz |
| Emotion metadata | 1. Coded representation of Basic Emotion metadata 2. Coded representation of cultures 3. Coded representation of Specific Emotion metadata |
| Metadata attached to speech | File format of speech with sync’ed metadata |
| Speech features | Coded speech features |
| Image features | Coded image features |
| Emotion KB (text) query format | Format of query to Emotion KB (text)  Format of response from Emotion KB (text) |
| Emotion KB (speech) query format | Format of query to Emotion KB (speech)  Format of response from Emotion KB (speech) |
| Emotion KB (image) query format | Format of query to Emotion KB (image)  Format of response from Emotion KB (image) |
| Emotion descriptors | Coded representation of Emotion descriptors |
| Text with emotion | Coded representation of Text with emotion |
| Meaning | Coded representation of Meaning |

### Digital Speech

MPAI should not be too prescriptive. It should allow use of speech sampled in the 22.05-96 kHz with 16-24 bit/sample.

The CfT should not call for Digital speech technologies. However, it might be useful to collect comments on the choice made by MPAI.

### Emotion metadata

By Emotion metadata we mean attributes that classify emotion.

The most basic emotions are fear, happiness, sadness, anger. These can be taken as “universal” in the sense that they are common to all cultures. Hower, in the literature other universal emotions are proposed. Therefore, the CfT should call for universal emotion metadata and their grades, their semantics and digital representation.

We need the requirements derived from the intended application: how is emotion added to em-otion-less speech.

It may be too early to ask for metadata that represent culture-dependent emotions. If we do, we should probably call for metadata that describe cultures and emotions within a culture. Even if we do not call for them, however, we should make it clear in the CfT that the basic emotion metadata are expected to be extensible, e.g. capable to represent culture-dependent emotions by adding appropriate metadata.

We should pay particular attention to this Emotion metadata technology because its use is not restricted to speech. A machine could produce images conveying a particular emotion. Conver-sation with emotion needs Emotion metadata for text and images (videos?).

### Speech features

Speech features are extracted by the speech recognition module using the input speech and emotion of the input speech is determined based on the speech features

The following features might be considered because they have information about emotion.

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.
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.
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.

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

### Image features

Image features are extracted by the Image analysis module using the input image and the emotion that the input image carries is determined using emotion KB (image).

Image features should represent face properties of single images, such as the opening of the mouth or rising of the eye-brows as a feature vector. Then these features can be used to determine the emotion that the facial expression is associated with.

### Emotion KB (text) query format

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

The Language understanding module queries the Emotion KB by giving a text as input. the Emotion KB responds by giving emotions corelated with the text input.

### 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 Speech recognition module queries the Emotion KB (speech) by giving a list of speech features as input. The Emotion KB responds with frequency-domain and time-domain features correlated with the emotions of the Speech.

### Emotion KB (image) query format

Emotion KB (image) contains features extracted from the different images of facial expressions from diverse group of people with an agreed set of emotions.

The Image analysis module queries the Emotion KB (image) by giving a list of image features as input. The Emotion KB (image) responds by giving the emotions the image conveys.

### Emotion descriptors

Emotion descriptors are produced by Emotion recognition module and sent to Dialog processing to produce the reply to the user input based on the emotion and the meaning of the input sentence.

### Text with emotion

We need a standard format for text with emotions assigned to different portions of the text. Text should be encoded according to ISO/IEC 10646, Information technology – Universal Coded Character Set (UCS) to support most languages in use, but the CfT should ask for comments. An example of how emotion in the text should be represented is offered by emoticons.

### Meaning

Meaning is structured data produced by Language understanding ­­­on the user input in text. The meaning and emotion of the user input will be sent to the Dialog processing module to produce the response of the conversation system. The format of the meaning and what it contains among many language analysis results are the target of CfT.

## Multimodal Question Answering

### Use case description

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.

### Implementation

Question and image are recognised and analysed in the following way and answers are produced in the output speech:

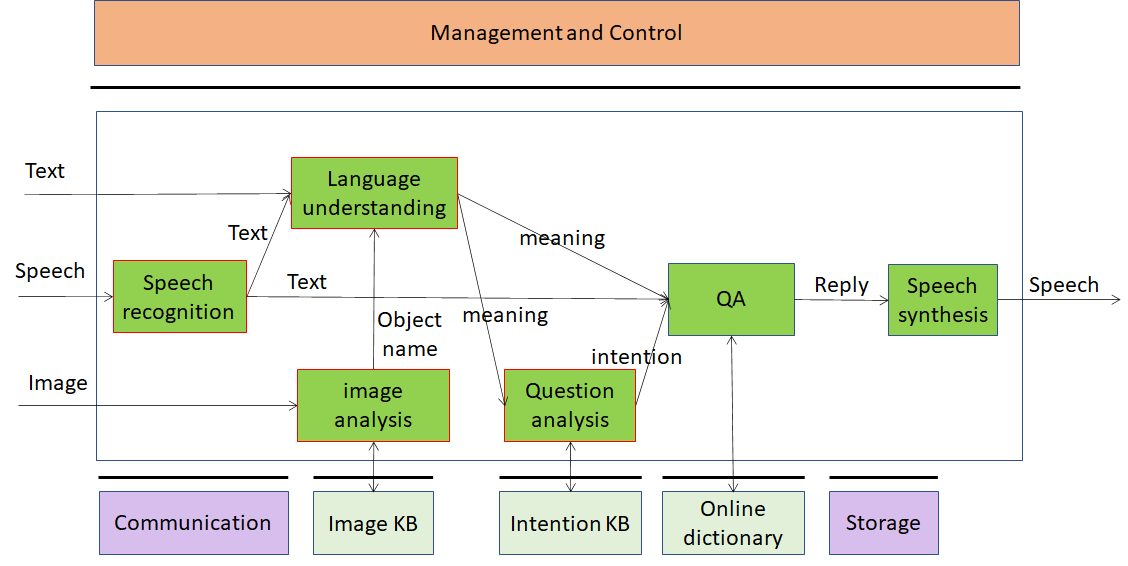
1. The meaning of the question is recognised in the form of text or voice
2. Image is analysed and the object name of the image is sent to the language understanding module.
3. The integrated meaning from the multimodal inputs are generated from the language understanding module.
4. Intention analysis module determines the intention of the question and the intention is sent to the QA module.
5. The QA module produces the answer based on the intention of the question, meaning from the Language understanding module.
6. The speech synthesis module produces the speech from the answer in text.

The architecture of *Figure 2* supports the case in which the user can or cannot use speech. Text information is fed into Language understanding through either speech recognition or text input by the user.

The following AIMs can be implemented either as AI or legacy modules: Image analysis, Speech recognition, Language understanding and dialog KB. If any of these AIMs are implemented as a neural network, access to the corresponding KB is not needed.

An MPAI implementation of this use case can be depicted by *Figure 4*.

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



*Figure 5* – *Multimodal Question Answering*

### AI Modules and External data

|  |  |
| --- | --- |
| **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 |
| **Questin analysis** | Analyses the the meaning of the sentence and determines the Intention |
| **QA** | Analyses user’s question and produces a reply based on user’s Intention |
| **Intention KB** | Responds to Intention analysis’s 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 QA module to find answers to the question |

### Workflow

1. Speech enters Speech recognition
2. Speech recognition module
   1. Converts Speech into Text
   2. Sends Text to Language understanding and QA
3. Image enters Image analysis
4. Image analysis
   1. Computes Image features
   2. Queries Image KB
   3. Computes Image descriptors
5. Text enter Language understanding
6. Language understanding
   1. Generates meaning
   2. Sends meaning to Intention analysis and QA
   3. Sends Image descriptors to Language understanding
7. Intention analysis module
   1. Queries Intention KB
   2. Computes Intention
   3. Sends Intention to QA
8. QA module
   1. Queries Online dictionary
   2. Provides Reply
   3. Sends Reply (text) to Speech synthesis
9. Speech synthesis module
   1. Produces Speech
   2. Outputs Speech

### Functions and potential technologies for CfT

|  |  |
| --- | --- |
| **Function** | **Potential CfT items** |
| Digital Speech | 24 bit/s, 22.05-96 kHz |
| Text | ISO/IEC 10646 |
| Digital image | Digital Image format |
| Image features | Coded representation of image features |
| Image KB query format | Input query format of Image KB  Output query format of Image KB |
| Image descriptors | Coded representation of Image descriptors |
| Meaning | Format of the meaning of the question |
| Intention KB query format | Input query format of Intention KB,  Output query format of Intention KB |
| Intention | Format of Intention of the question |
| Online dictionary query format | Input query format of online dictionary  Output query format of online dictionary |

### Digital Speech

MPAI should not be too prescriptive. It should allow use of speech sampled in the 22.05-96 kHz with 16-24 bit/sample.

The CfT should not call for Digital speech technologies. However, it might be useful to collect comments on the choice made by MPAI.

### Text

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

### Digital image

The CfT should not call for Digital image technologies. However, it might be useful to collect comments on the choice made by MPAI.

### Image features

Image features are extracted by Image analysis from the input image and used to query the Image KB.

### Image KB query format

Image KB contains image features and corresponding image descriptors. Image analysis queries the Image KB giving image features as input. Image KB responds with object name.

### Image descriptors

Text representing the name of the object produced by Image analysis and sent to Language understanding. Language understanding interprets the input sentence based on the image descriptor and the text to obtain the correct meaning of the input sentence.

### Meaning

Meaning is structured data produced by Language understanding on the user question. The format of the meaning and what it contains among many language analysis results are the target of CfT.

### Intention KB query format

Intention KB contains intention types of the user question and the keywords that denote those intention types.

Question Analysis queries the Intention KB by giving text as input. Intention KB responds with the type of question intention correlated with the keywords in the query.

### Intention

The CfT should call for a standard and extensible classification and applicable format of all types of intention of the user question.

### Online dictionary query format

Online dictionary contains the structured data that include topics and related information. QA queries Online dictionary by giving text as input. Online dictionary responds with paragraphs that contain answers that have the highest correlation with the user question.

## Personalized Automatic Speech Translation

### Use case description

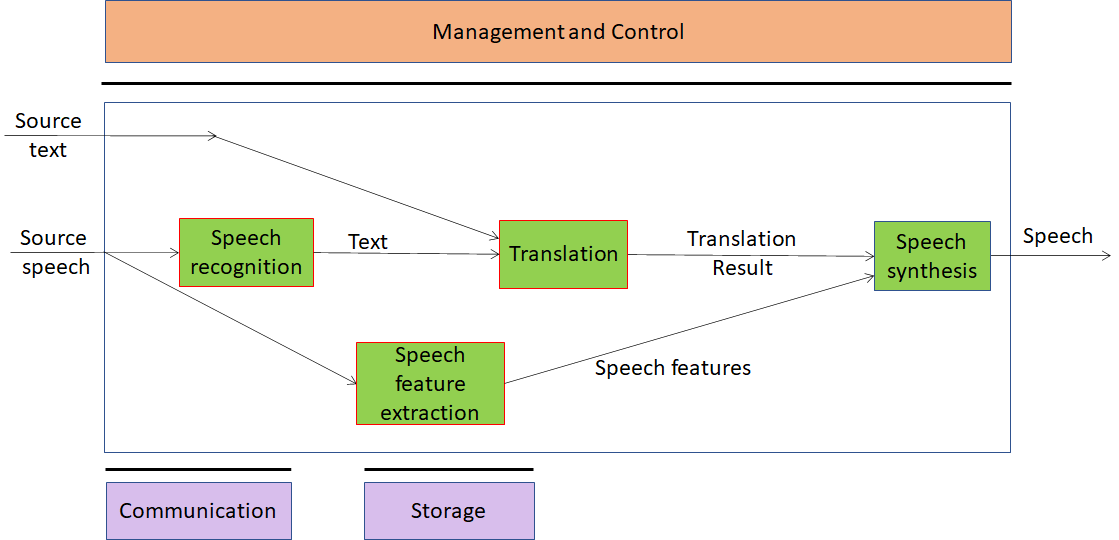
Automatic speech translation technology denotes technology that recognizes a voice uttered in a language by a speaker, converts the recognized voice into another language through automatic translation, and outputs a converted voice as text-type subtitles or as a synthesized voice preserving the speaker’s 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.

### Implementation

The personalized speech translation is working in the following way and the translation results are produced in the output speech:

1. The input text or voice which is recognised as text through Speech recognition module is translated in the Translation module.
2. The input voice is analysed through speech feature extraction module and speech features are sent to the speech synthesis module.
3. The speech synthesis module produces the speech from the translation result and speech features of the user input.

The AI Modules implied by a personalized automatic speech translation system look approxim­ately as presented in *Figure 6*. The interaction between different AIMs are described including a speech recognition module, a speech feature extraction module, a translation module and a speech synthesis module.



*Figure 6* – *Personalized Automatic Speech Translation*

### AI Modules and External data

|  |  |
| --- | --- |
| **Speech Recognition** | Converts Speech into Text |
| **Speech synthesis** | Produces Speech from the text resulting from translation with the speech features extracted from the speaker of the source language |
| **Translation** | Translates the user 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  Encodes personal voice features |
| **Speech feature KB** | Allows Speech feature extraction module to obtain the Speech features |

### Workflow

1. Source speech enters Speech recognition
2. Speech recognition
   1. Produces Text
   2. Sends Text to Translation module
3. Speech feature extraction
   1. Extract Speech features such as tones, intonation, intensity, pitch, emotion, intensity or speed
   2. Sends Speech features to Speech synthesis
4. Source text enters Translation
5. Translation
   1. Translates Source text into text in the target language
   2. Sends Translation results to Speech synthesis
6. Speech synthesis module
   1. Produces Speech that reflects Speech features of Source speech
   2. Outputs Speech

### Functions and potential technologies for CfT

|  |  |
| --- | --- |
| **Function** | **Potential CfT items** |
| Digital speech | 16-24 bit/s, 22-96 kHz |
| Speech features | Coded speech features |
| Text | ISO/IEC 10646 |
| Language identification | Language classification, ISO 639 |

### Digital speech

MPAI should not be too prescriptive. It should allow use of speech sampled in the 22.05-96 kHz with 16-24 bit/sample.

The CfT should not call for Digital speech technologies. However, it might be useful to collect comments on the choice made by MPAI.

### Speech features

Speech features such as tones, intonation, intensity, pitch, emotion, intensity or speed are extracted by the speech extraction module using the input speech. The speech features are used to encode personal voice features for the translation results in the Speech synthesis module.

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.

### Text

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

### Language identification

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

# Potential common technologies

The following acronyms have been introduced

|  |  |  |
| --- | --- | --- |
| **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 |

The following technologies are potentially applicable to different Use Cases.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Function** | **EES** | **ARP** | **EAE** | **AOG** | **CWE** | **MQA** | **PST** |
| Digital speech | X |  | X |  |  |  |  |
| Digital Audio |  | X |  | X |  |  |  |
| Digital image |  | x |  |  | X | X |  |
| Emotion metadata | X |  |  |  | X | X | X |
| Speech features | X |  |  |  | X | X | X |
| Text |  | X |  |  | X | X | X |
| Image features |  |  |  |  |  |  |  |
| Image descriptors |  | X |  |  | X | X |  |

# Conclusions

The document in its current form is work in progress. MPAI intends to add more details to the Use Cases add more Use Cases to enable MPAI before issuing a Call for Technologies.

When the document will be considered sufficiently mature, MPAI will issue a Call for Technol-ogies requesting MPAI members and the industry members to submit proposals for:

1. Data formats suitable as inputs and outputs of the identified AIMs
2. Additions or removal of input/output signals to the identified AIMs with identification of data formats required by the new input/output signals
3. Possible alternative partitioning of the AIMs implementing the example cases providing
   1. Arguments in support of the proposed partitioning
   2. Detailed specifications of the inputs and outputs of the proposed AIMs
4. New Use Cases fully described as in the final version of this document.

Respondents will be asked to state in their submissions their intention to adhere to the Framework Licence developed for MPAI-MMC when licencing their technologies if included in the MPAI-MMC standard. Please note that “a Framework Licence is the set of conditions of use of a licence without the values, e.g., currency, percent, dates etc.”. The Framework Licence will give the MPAI-MMC standard a clear IPR licensing framework.

The MPAI-MMC Framework Licence will be developed, as for all other MPAI Framework Licences, in compliance with the generally accepted principles of competition law.