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

|  |  |
| --- | --- |
| N71 | 2020/11/18 |
| Source | Michelangelo Guarise, Sergio Canazza |
| Title | Proposal for MPAI-CAE Functional Requirements work programme |
| 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 AI data coding standards is by defining *AI Modules (AIM)* with standard interfaces that are combined and executed within an MPAI-specified AI-Framework. With its standards, MPAI intends to promote the development of *horizontal markets* of *competing* *proprietary* solutions with standard interfaces tapping from and further promoting AI *innovation.*

This paper describes the current MPAI plan to develop “Context-based Audio Enhancement” (MPAI-CAE), an MPAI area of work that uses AI substantially to improve the user experience for a variety of uses such as entertainment, communication, teleconferencing, gaming, post-produc­tion, restorat­ion etc. in a variety of contexts such as in the home, in the car, on-the-go, in the studio etc.

Chapter 2 explains the MPAI-CAE features, Chapter 3 provides summary information on the advanced IT environment that will execute MPAI-CAE applications and Chapter 4 identifies the items that will likely be the object of the MPAI-CAE standard.

# MPAI-CAE features

Currently, there are solutions that adapt the conditions in which the user experiences content or service for some of the contexts mentioned above. However, they tend to be vertical in nature, making it difficult to re-use possibly valuable AI-based components of the solutions for different applications.

MPAI-CAE uses context information to act on the input audio content using AI, processing such content via updatable and extensible AIMs, and finally delivering the processed output via the most appropriate protocol.

MPAI-CAE allows providers, vendors and manufacturers to deliver complex optimizations and thus superior user experience with reduced time to market as MPAI-CAE will make combinations of 3rd party components easy from a technical and licensing perspective.

So far, the AIMs required by the following application areas have been considered for possible standardisation by MPAI-CAE:

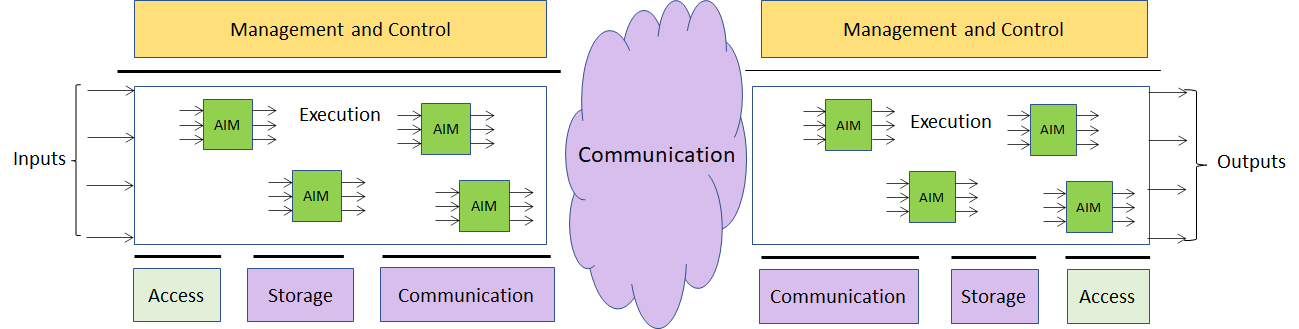
1. Enhanced audio experience in a conference call (see 4.1): Adaptive audio processing Pipeline to improve conference call experience.
2. Audio-on-the-go (see 4.2): Adaptive audio processing Pipeline to improve Sound Quality on the go without loosing contact with the acoustic surroundings.
3. Emotion enhanced synthesized voice: Expressive speech model based on the primary emotions (fear, happiness, sadness, and anger) (see 4.3)
4. AI for audio documents cultural heritage (see 4.4): **Automatic techniques** to extract information from analog audio and video tapes: **automatic analysis** (preprocessing step); and sec­ond step (classification), in which a **classifier** is used to determine the content of each image saved during pre-processing.
5. (Serious) gaming
6. Efficient 3D sound
7. Normalization of TV volume
8. Automotive
9. Audio mastering
10. Voice communication
11. Audio (post-)production

# AI Framework

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.

# MPAI-CAE work plan

In this chapter there are currently two application areas with one relevant AI Module (AIM) identified and its inputs/outputs summarily specified.

## Enhanced audioconference experience

Often, the user experience of a video/audio conference can be marginal. Too much background noise or undesired sounds can lead to participants not understanding what participants are saying.

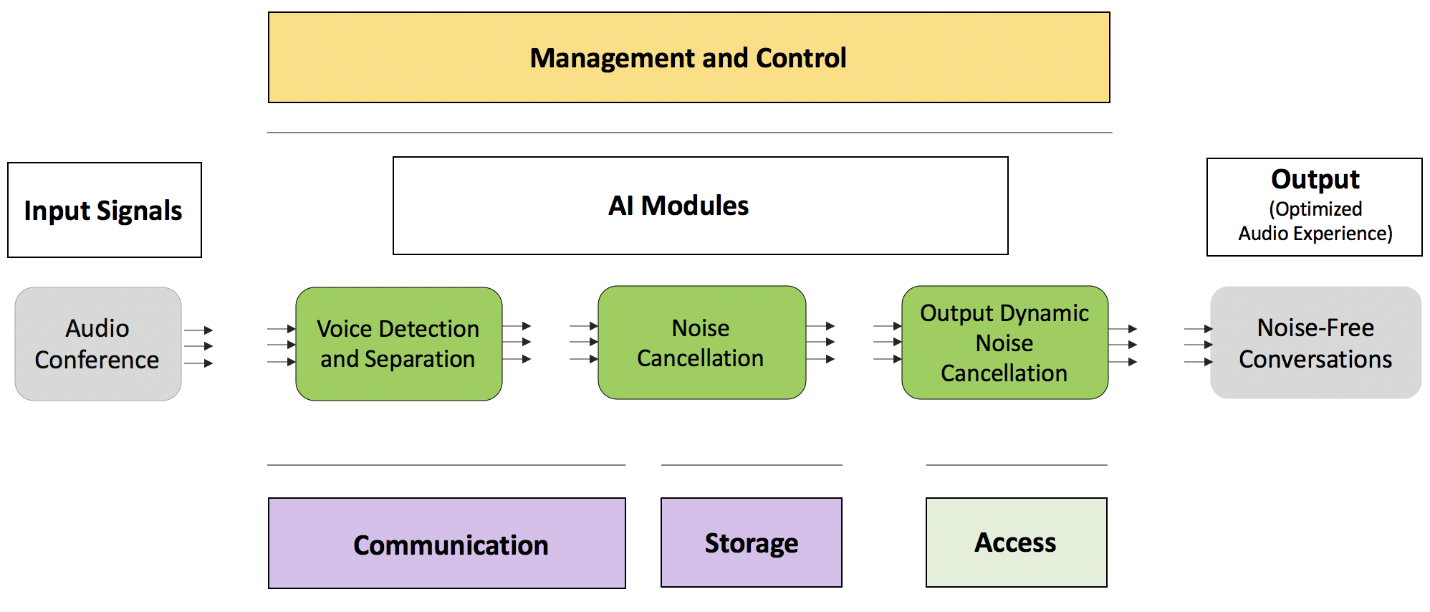
By using AI-based adaptive noise-cancellation and sound enhancement, MPAI-CAE can virtually eliminate those kinds of noise without using complex microphone systems to capture environment characteristics.

The input signal (audioconference Audio) is treated using a combination of processing by three different processing modules:

* Voice Recognition, which can discern voice vs non-voice signal, allowing removal of non-voice or not-relevant sounds from the conversation.
* Noise Cancellation Component, which further removes noise element from the conver­sation.
* Output Dynamic Noise Cancellation, which can further reduce the level of noise considering the output characteristics.

The output signal, resulting from the combined process above, will then be delivered using the most suitable delivery protocol for the current usage scenario e.g. Bluetooth low latency if suitable hardware is used.

The AI Framework for this usage example is given by the following *Figure 2*.



*Figure 2 –Audioconference*

In the following the lists of inputs and outputs of the AIMs required by the usage example are given.

### Voice detection and separation

|  |  |
| --- | --- |
| Function | Discern relevant voice vs non-voice signals |
| Inputs | 1. Single microphone with its physical characteristics 2. Geometry: 1 or more located in different places |
| Outputs | 1. Voice signal 2. Any other non-voice signal 3. Geometry |

### Noise cancellation

|  |  |
| --- | --- |
| Function | Remove Noise elements from Audio Signal |
| Inputs | 1. Voice Signals (from Voice detection and separation AIM) 2. Geometry (from Voice detection and separation AIM) |
| Outputs | 1. De-Noised Voice Signal 2. Noise signal |

### Output dynamic noise cancellation

|  |  |
| --- | --- |
| Function | Reduce the level of noise considering Output Characteristics |
| Inputs | 1. De-Noised Voice Signal (from Noise cancellation AIM) 2. Set of metadata representing Output Device Achoustic Model |
| Outputs | 1. De-Noised Voice Signal with equalisation based on Output Device Achoustic Model |

## Audio-on-the-go

While biking in the middle of city traffic, AI can process the signals from the environment captured by the microphones available in many earphones and earbuds (for active noise cancellation), adapt the sound rendition to the acoustic environment, provide an enhanced audio experience (e.g. performing dynamic signal equalization), improve battery life and selectively recognize and allow relevant environment sounds (i.e. the horn of a car).

The user enjoys a satisfactory listening experience without losing contact with the acoustic surroundings.

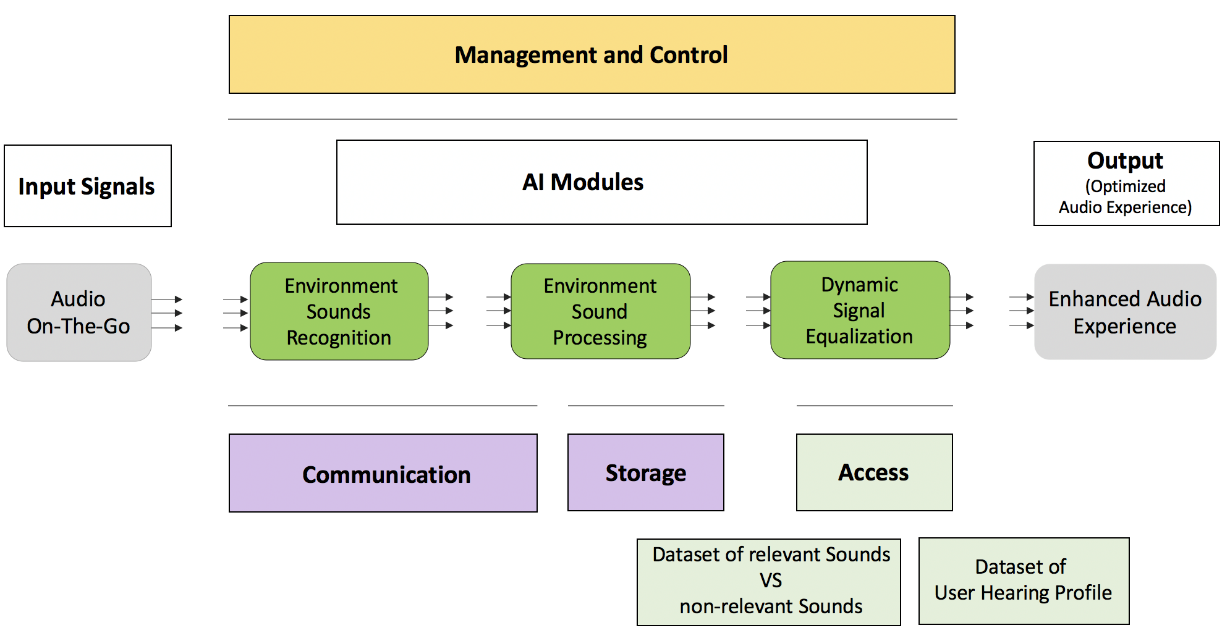
The input signal (Music Content) is treated using a combination of processing by three different processing modules:

* Environment Sounds Recognition, which is able to regognize and categorize the surrounding environment sounds
* Environment Sound Processing, which is able to determine which sounds are relevant for the user (sounds which the user needs to aware of e.g. Car Noise, Car Horn ) VS sounds which are not and that can be therefore removed
* Dynamic Signal Equalization, which based on the current envinronment noise lever and user hearing profile, can dynamically equalize the sound to produce the best possible quality output

The output signal, resulting from the combined process above, will then be delivered using the most suitable delivery protocol for the current usage scenario e.g. Bluetooth low latency if suitable hardware is used.

The AI Framework for this usage example is given by the following *Figure 3*

.



*Figure 3 –Audio-on-the-go*

### Environment sounds recognition

|  |  |
| --- | --- |
| Function | Recognize and categorize surrounding environment sounds |
| Inputs | 1. Single microphone with its physical characteristics 2. GPS Position 3. Accelerometer\Gyroscope 4. Datasets of Sounds and relative categorizations |
| Outputs | 1. Array of recognized and categorized sounds |

### Environment sound processing

|  |  |
| --- | --- |
| Function | Determine which sounds are relevant for the user VS sounds which are not and that can be therefore removed |
| Inputs | 1. Array of recognized and categorized sounds (from Environment Sound Recognition AIM) |
| Outputs | 1. Relevant Sounds 2. Non-Relevant Sounds |
| Access | 1. Dataset of relevant Sounds VS non-relevant Sounds |

### Dynamic Signal Equalization

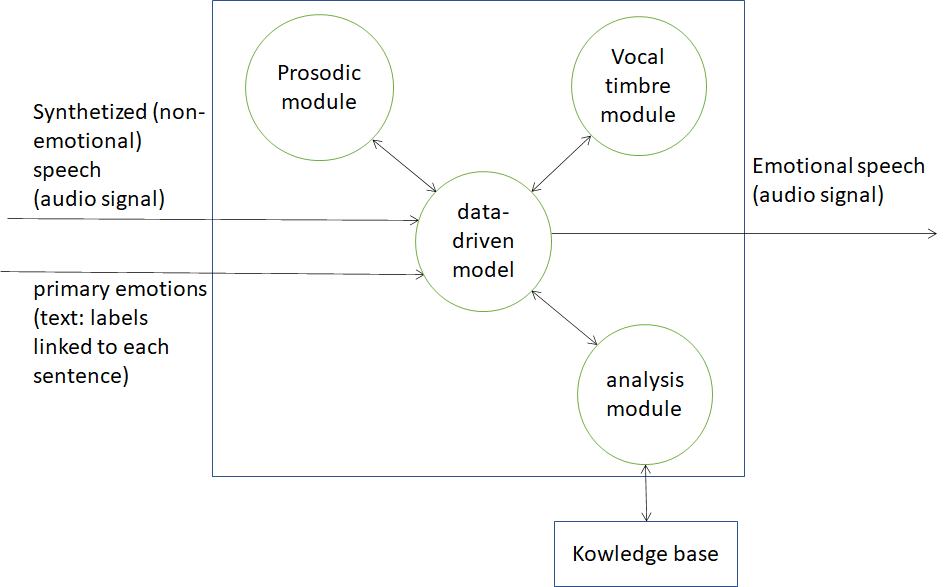
|  |  |
| --- | --- |
| Function | Dynamically equalize the sound to produce the best possible quality output |
| Inputs | 1. Relevant Sounds (from Environment Sound Processing AIM) |
| Outputs | 1. Dynamically equalized Sound |
| Access | 1. Dataset of User Hearing Profile. |

## Emotion enhanced synthesized voice

Voice quality is recognized to play an important role for the rendering of emotions in verbal communication. This application field is related to the analysis and synthesis of **emotional speech**. A set of acoustic cues have to be selected to compare the voice quality characteristics of the speech signals on a voice corpus in which different emotions are reproduced. The psychoacoustic parameters of emotions in speech can be separated into two groups: prosodic (rhythm, speed of speech, intonation and intensity) and vocal timbre-related parameters (position of the formants and distribution of the spectral energy).

**Data driven voice transformation algorithm** can be profitably used to alter the timbre of a neutral (non-emotional) synthesized voice in order to reproduce a particular emotional (fear, hap­piness, sadness, or anger) vocal timbre, based on a (**data-driven**) model obtained with training on real data for both **prosodic** and **vocal timbre** modules.

The Environment Component of the AI Framework for this usage example is represented by *Figure 4*



*Figure 4 – Emotion enhanced synthesized voice*

### Knowledge base

|  |  |
| --- | --- |
| Function | To allows the analysis module, based on training on real data, to learn the spectral characteristics of the voice. |
| Inputs | Query by similarity |
| Outputs | Audio (speech) signal. |
| Access | Dataset populated with recordings of different speakers reading/reciting a corpus of texts with different emotional styles: fear, happiness, sadness, or anger and a *neutral* style of reference. |

### Analysis module

|  |  |
| --- | --- |
| Function | To analyse the audio of the dataset and the spectral characteristics of the voice to make emotional. |
| Inputs | Audio (speech) signal. |
| Outputs | Audio (speech) signal with emotional descriptors (metadata). |

### Data-driven model

|  |  |
| --- | --- |
| Function | To process a neutral (non-emotional) synthesized voice in order to reproduce a particular emotional (fear, happiness, sadness, or anger) vocal timbre |
| Inputs | 1. audio (speech) signal with emotional descriptors (metadata); 2. synthesized (non emotional) speech; 3. verbal description (text) to make speech emotional. |
| Outputs | 1. Speech features extracted from audio. 2. Emotional speech (audio signal). |

### Prosodic module

|  |  |
| --- | --- |
| Function | To process the prosodic parameters of emotions in speech (rhythm, speed of speech, intonation and intensity). |
| Inputs | Speech features extracted by data-driven model. |
| Outputs | Emotional prosodic parameters. |

### Vocal timbre module

|  |  |
| --- | --- |
| Function | To process the vocal timbre-related parameters of emotions in speech (position of the formants and distribution of the spectral energy). |
| Inputs | Speech features extracted by data-driven model. |
| Outputs | Emotional prosodic parameters. |

## Audio documents cultural heritage

Computer science offers multiple possibilities to study the fields of humanities: a major topic that has been rapidly growing along the past decades is the implementation of **AI algorithms** in musical cultural heritage, with a particular relation to the audio documents preservation.

Recordings contain information on their artistic and cultural existence that goes beyond the audio signal itself. In this sense, a faithful and satisfying access to the audio document cannot be achieved without its associated contextual information, that is, to all the content-independent information represented by the container, the signs on the carrier, the accompanying material, and so on.

In particular, music on analog magnetic tape is characterized by several carrier-related specificities that must be considered when creating a copy for digital preservation. The magnetic tape could have some intentional or unintentional alterations. During both the creation and the musicological analysis of a digital preservation copy, the quality of the work may be affected by human inattention.

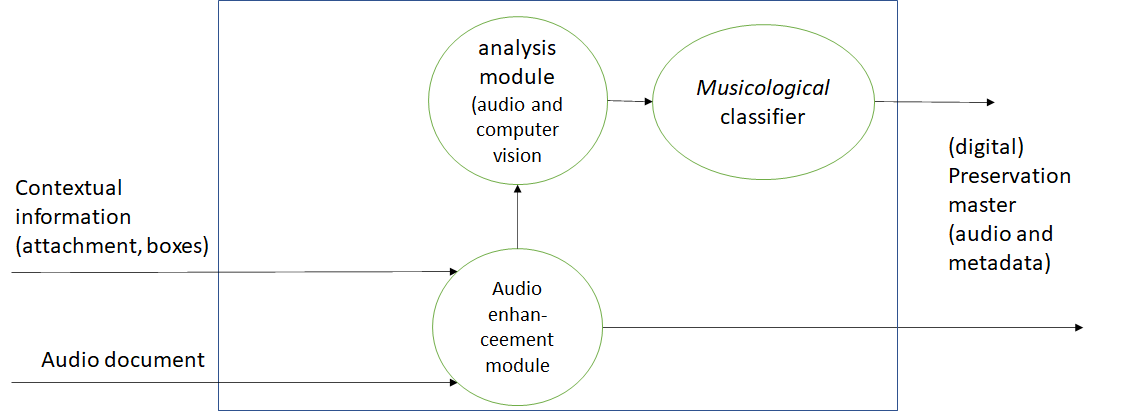
There are many aspects that need to be considered during the digitization of a tape. There is the primary information (i.e., the audio signal recorded). Then there is the secondary information, such as alterations of the carrier (corruptions, splices, signs, etc.). All of these metadata need to be stored with the preservation copies alongside the digital audio. In this sense, an important feature of the preservation process is the video recording of the tape as it passes the head of the tape recorder, which is important to preserve important ancillary information. The video recording offers infor­mation on the operations of the magnetic tape assembly, such as the splices used to join different pieces of tape and possible corruptions of the carrier; instructions for the performance of the piece (markings on the tape, representing points to be synchronized with a musical score, or indicating particular sound events); description of the irregularities in the playback speed of analog recor­dings, such as wow and flutter.

This application field emphasize the “textual” aspects of a sound document, considering the A/D transfer as a philological operation of *restitutio textus*.

**Automatic techniques** to extract information from audio and video of the tapes are useful to relieve technicians and musicologists of repetitive, tiresome, or otherwise error-prone tasks that are better performed by a machine.

During **pre-processing**, the first step, the video is examined frame by frame, and each image showing a potentially significant discontinuity is recognized (by means **computer vision techniques**) and saved. The exact content of the images is not determined. That task is the aim of the second step, classification, in which a **classifier** is used to determine the content of each image saved during pre-processing. In this way, the **video (not only the audio signal) is compressed**, considering only the “interesting” few frames.

The Environment Component of the AI Framework for this usage example is represented by *Figure 5*



*Figure 5 – Audio documents cultural heritage*

### Audio enhancement module

|  |  |
| --- | --- |
| Function | To digitize audio signal and contextual information (attachment, boxes). |
| Inputs | Original sound document (audio magnetic tape) |
| Outputs | 1. High quality digital audio; 2. video recording of the tape as it passes the head of the tape recorder. |

### Analysis module

|  |  |
| --- | --- |
| Function | To carry out feature extraction from audio and video. |
| Inputs | 1. High quality digital audio; 2. video recording of the tape as it passes the head of the tape recorder. |
| Outputs | Audio and video frames. |

### Musicological classifier

|  |  |
| --- | --- |
| Function | To classify the features extracted by analysis module. |
| Inputs | Audio signal excerpts and video frames. |
| Outputs | Audio excerpts (signal) and video frames (images) with verbal description (text). |

## (Serious) gaming

## Efficient 3D sound

## Normalization of TV volume

## Automotive

## Audio mastering

## Voice communication

## Audio (post-)production

# Conclusions

The document in its current form is work in progress. MPAI intends to add more details to the existing to enable MPAI to issue a Call for Technologies. MPAI may also add more usage exam­ples.

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

1. *Data formats* suitable as inputs and outputs of the identified AIMs
2. 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
3. New *usage* *examples* 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-CAE when licencing their technologies if included in the MPAI-CAE 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* willgive the MPAI-CAE standard a *clear IPR licensing* framework.

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