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

|  |  |
| --- | --- |
| **Public Document** | |
| N39 | 2020/10/21 |
| Source | MPAI-1 |
| Title | MPAI Application Note #6 - MPAI-MMC – Multi-Modal Conversation |
| Target | MPAI Members |

### MPAI Application Note #6

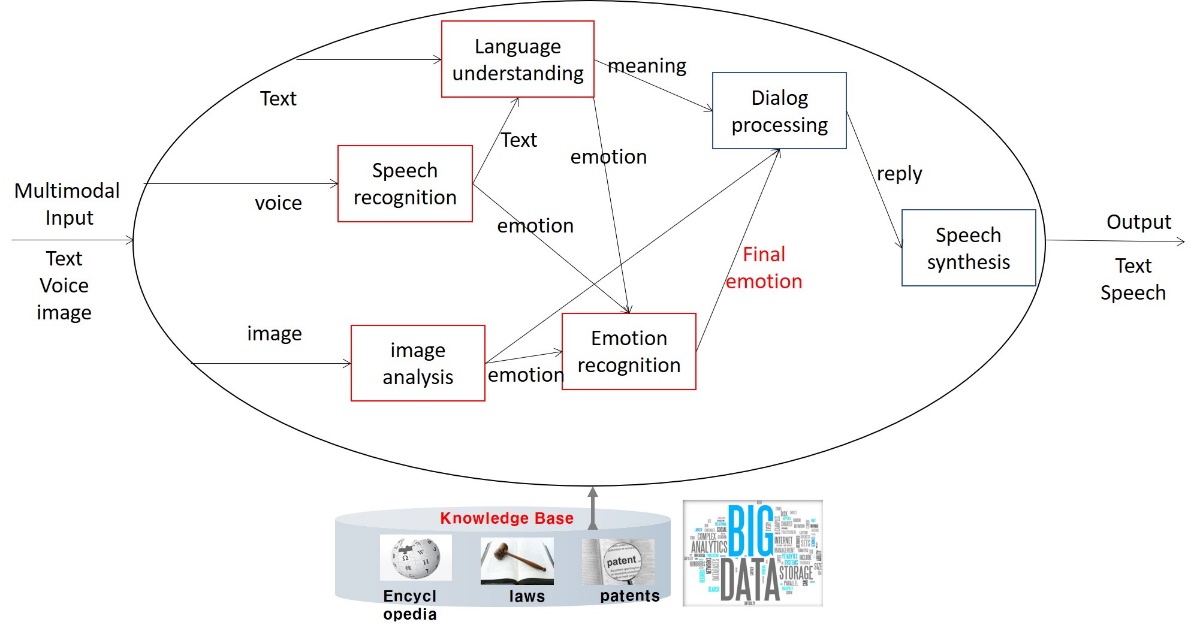
### Multi-Modal Conversation (MPAI-MMC)

**Proponent**: Miran Choi (ETRI)

**Description**: Owing to recent advances of AI technologies, natural language processing started to be widely used in various applications. One of the useful applications is the conversational partner which provides the user with information, entertains, chats and answers questions through the speech interface. However, an application should include more than just a speech interface to provide a better service to the user. For example, emotion recognizer and gesture interpreter are needed for better multi-modal interfaces.

Multi-modal conversation (MPAI-MMC) aims to enable human-machine conversation that emulates human-human conversation in completeness and intensity by using AI.

The interaction of AI processing modules implied by a multi-modal conversation system would look approximately as presented in *Figure 1*, where one can see a language understanding module, a speech recognition module, image analysis module, a dialog processing module, and a speech synthesis module.



*Figure 1 – Multi-Modal Conversation*

**Comments**: The processing modules of the MPAI-MMC instance of *Figure 1* would be operated in the MPAI-AIF framework.

**Examples**

The example of MMC is the conversation between a human user and a computer/robot as in the following list. The input from the user can be voice, text or image or combination of different inputs. Considering emotion of the human user, MMC will output responses in a text, speech, music depending on the user’s needs.

* Chats: “I am bored. What should I do now?” - “You look tired. Why don’t you take a walk?”
* Question Answering: “Who is the famous artist in Barcelona?” – “Do you mean Gaudi?”
* Information Request: “What’s the weather today?” – “It is a little cloudy and cold.”
* Action Request: “Play some classical music, please” - “OK. Do you like Brahms?”

**Processing modules involved in MMC:**

A preliminary list of processing modules is given below:

1. Fusion of multi-modal input information
2. Natural language understanding
3. Natural language generation
4. Speech recognition
5. Speech synthesis
6. Emotion recognition
7. Intention understanding
8. Image analysis
9. Knowledge fusion from different sources such as speech, facial expression, gestures, etc
10. Dialog processing
11. Question Answering
12. Machine Reading Comprehension (MRC)
13. Speech Synthesis

**Requirements:**

This is the initial functional requirements that will be developed in the full set in the Functional Requirements (FR) phase..

1. The standard shall specify the following natural input signals

* Sound signals from microphone
* Text from keyboard or keypad
* Images from the camera

1. The standard shall specify a user profile format (e.g. gender, age, specific needs, etc.)
2. The standard shall support emotion-based dialog processing that uses emotion result from the emotion recognition as input and decides the replies based on the user’s intention as output.
3. The standard should provide means to carry emotion and user preferences in the speech synthesis processing module.
4. Processing modules should be agnostic to AI, ML or DP technology: it should be general enough to avoid limitations in terms of algorithmic structure, storage and communication and allow full interoperability with other processing modules.
5. The standard should provide support for the storage of, and access to:

* Unprocessed data in speech, text or image form
* Processed data in the form of annotations (semantic labelling). Such annotations can be produced as the result of primary analysis of the unprocessed data or come from external sources such as knowledge base.
* meta-data (such as collection date and place; classification data)
* Support for the structured data produced from the raw data.

1. The standard should also provide support for:

* The combination into a general analysis workflow of a number of computational blocks that access processed, and possibly unprocessed, data such as input channels, and produce output as a sequence of vectors in a space of arbitrary dimension.
* The possibility of defining and implementing a novel processing block from scratch in terms of either some source code or a proprietary binary codec
* A number of pre-defined blocks that implement well-known analysis methods (such as NN-based methods).
* The parallel and sequential combination of processing modules that comprise different services.
* The real time processing for the conversation between the user and the robot/computer.

**Object of standard**: Interfaces of processing components utilized in multimodal communication.

* **Input interfaces**: how to deal with inputs in different formats
* **Processing component interfaces:** interfaces between a set of updatable and extensible processing modules
* **Delivery protocol interfaces**: Interfaces of the processed data signal to a variety of delivery protocols
* **Framework**: the glue keeping the pieces together => mapping to MPAI-AIF

**Benefits:**

1. Decisively improve communication between humans and machines and the user experience
2. Reuse of processing components for different applications
3. Create a horizontal market of multimodal conversational components
4. Makes the market more competitive

**Bottlenecks:**

Some processing units should be improved because end-to-end processing has lower performances compared to modular approaches. Therefore, the standard should be able to cover the traditional method as well as hybrid approaches.

**Social aspects**:

Enhanced user interfaces will provide accessibility for people with disabilities. MMC can also be used in care giving services for elderly and patients.

**Success criteria**:

* How MMC can be extended to different services by combining several processing modules easily and easily.
* The performance of multi-modality compared to uni-modality in the user interface.
* Interconnection and integration among different processing modules