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

|  |  |
| --- | --- |
| **N338** | 2021/08/25 |
| **Source** | Requirements (MCS) |
| **Title** | MPAI-MCS Use Cases, Components and Data formats  |
| **Target** | MPAI Community |

[1 Use Cases 1](#_Toc80648914)

[1.1 Use Case #1 – Multipoint videoconference 1](#_Toc80648915)

[1.2 Use Case #2 – Virtual e-learning 2](#_Toc80648916)

[1.3 Use Case #3 – Teleconsulting 3](#_Toc80648917)

[1.4 Use Case #4 – Avatar videoconference in a local 3D audio-visual space 4](#_Toc80648918)

[1.4.1 Description 4](#_Toc80648919)

[1.4.2 Steps 4](#_Toc80648920)

[1.4.3 Participant TX Reference Model 5](#_Toc80648921)

[1.4.4 MCS Reference Model 6](#_Toc80648922)

[1.4.5 Participant RX Reference Model 7](#_Toc80648923)

[1.4.6 Comments 7](#_Toc80648924)

[2 MCS description 7](#_Toc80648925)

[2.1 Context metadata 7](#_Toc80648926)

[2.2 Avatar metadata 7](#_Toc80648927)

[2.3 Object description 8](#_Toc80648928)

[2.4 General 8](#_Toc80648929)

[3 AIMs/Workflows required 8](#_Toc80648930)

[4 Data formats 9](#_Toc80648931)

[5 Terms and definitions 10](#_Toc80648932)

[6 References 10](#_Toc80648933)

# Use Cases

## Use Case #1 – Multipoint videoconference

The N participants in the conference reside at their locations, in their cultural environment. Their avatars sit around a virtual conference table located in a virtual room in an agreed cultural environment. A relevant quote is Marshall McLuhan’s “the medium is the message”.

This is how such a virtual shared-cultural conference could be managed:

1. The participants agree on and describe a shared cultural and/or context environment which can be real (representative of a physical space) or imagined (the components in the environment do not have a correspondence with the physical world):
	1. Conference style (board meeting, conference meeting, MPAI meeting etc.)
	2. Language that will be used in the shared space
	3. Room setting, furnishing, table and chairs, a CAV, outdoor
	4. Etc.
2. The organiser selects the multiconference service provider implementing the agreed setting
3. Participants provide/select and communicate to the multiconference service provider their own “personae”
	1. Avatar model
	2. Position in the meeting space
	3. Voice colour and style or own/synthetic
	4. Spoken language preference (e.g., EN-US, IT-CH) of the persona
4. Participant ensures that their own personae are authenticated
5. During the conference
	1. The camera of each participant
		1. Detects the participant’s body movements and extracts facial features and hand gestures
		2. Sends body movements and facial features to the multiconfer­ence unit
	2. The microphone set of a participant
		1. Captures the 3D field of the participant’s environment
		2. Separates the voice from the rest of the sound field
		3. Extracts and sends the sound field with descriptors of the speech
		4. Displays a choice of which sound field components should be preserved
	3. The multiconference unit
		1. Animates avatars at their assigned position using their body motions, facial features, hand gestures and speech descriptors
		2. Translates the cultural/context setting (speech etc.) of a participant to the agreed common setting
		3. Merges and sends to participants all sound fields as specified by each participant
		4. Sends participants an attendance table with metadata
	4. Participants
		1. Use the attendance table to, e.g., mute or reduce the influence of a particular source
		2. Place objects on their desks which are shown in front of them at the meeting or placed in the space for individual participants to engage, e.g., rotate etc.

## Use Case #2 – Virtual e-learning

A teacher holds a lecture to N students (in the following called participants to signify that the lecture is highly interactive and “non-frontal”). The teacher and the participants in the lecture res­ide at their own locations, in their cultural environments. Their avatars can sit classroom-style but the school or the cultural institution (hosting organisation), under whose aegis the lecture is held, could offer different arrangem­ents.

This is how such a virtual e-learning environment could be managed:

1. The hosting organisation makes available:
	1. Virtual spaces equipped with appropriate furnishings
	2. Populated by speaking and moving avatars
	3. The ability to convert
		1. Input speech from a language selected by the teacher and the speech in the agreed language to the languages of other participants
		2. Ditto for text
		3. Ditto for sign language
	4. Other objects.
2. The teacher selects
	1. A shared virtual space: real (e.g., representative of a physical space) or imagined (the virtual space does not correspond to an existing physical space) arranged as:
		1. Classroom style
		2. An evocative place, e.g., the Stoa of Athens
		3. With an orderly or scattered arrangement
	2. The language that will be used in the shared space
	3. Etc.
3. Participants provide/select and communicate their own “personae” to the hosting organisation:
	1. Avatar models or models with their affordances (i.e., the attributes of the model)
	2. Initial position in the meeting space
	3. Voice colour and style or own/synthetic
	4. Spoken language preference (e.g., EN-US, IT-CH) of their personae
4. During the conference
	1. The camera of each participant
		1. Detects the participant’s body movements and extracts facial features and hand gestures
		2. Sends body movements and facial features to the hosting organisation
	2. The microphone set of each participant
		1. Captures the 3D field of the participant’s environment
		2. Separates the voice from the rest of the sound field
		3. Extracts and sends the sound field with the descriptors of the speech
		4. Displays a list from which a participant can select the sound field components they wish to be preserved.
	3. Each participant has an acoustic echo cancellation
	4. The hosting organisation
		1. Animates avatars at their assigned positions moving their avatars and using their body motions, facial features, hand gestures and speech descriptors
		2. Translates the cultural/context setting (speech etc.) of a participant to the agreed common setting
		3. Merges and sends to participants all sound fields as specified by each participant
		4. Sends participants an attendance table with metadata.
	5. The teacher
		1. Uses the attendance table to, e.g., mute or reduce the influence of a particular participant
		2. Calls a synthetic 3D object from a DB and use it in support of the lecture
		3. Starts an experiment using a physical machine
		4. Place objects on his/her desk which are reproduced as (moving) 3D objects at participants’ locations so that they can engage interactively, e.g., rotate objects etc.

## Use Case #3 – Teleconsulting

An entrepreneur (E) offers teleconsulting services on a class of objects of particularly difficult use. A Customers (C) contacts E for advice on how to use a particular machine.

This is how the envisaged MCS teleconsulting service can take place:

1. C contacts E
2. E requests C to provide a 3D scan of the object
3. C provides the requested scan
4. E starts its MCS composed by
	1. the virtual representation of the object placed, e.g., on a table, or movable
	2. the avatar of E sitting in front of the object
	3. the avatar of C sitting next to the avatar of E
5. While speaking, the avatar of E manipulates the object
	1. e.g., rotates it
	2. touches a particular point of the object
	3. uses a virtual tool to indicate a type of operation
6. C and E see their own and the other avatars’ actions as if they were sitting in the virtual position of the avatar
7. While speaking, C acts on the physical object and the actions are reflected on the avatar and the virtual object
8. Avatars can move around the object (e.g., in the case of a large object)

## Use Case #4 – Avatar videoconference in a local 3D audio-visual space

### Description

Today’s videoconference falls short from being a satisfactory supplement to a physical meeting. Participants are able to see the full face of a speaker but are unable to have similar details for other speakers at the same time.

This use case is characterised by:

1. Each participants in a videoconference is represented by an avatar sitting at synthetic table of an MCS.
2. The body of each avatar is static.
3. The face/head of each avatar is animated by
	1. Movement of face/head,
	2. Emotion and meaning detected on the head and face of the avatar’s physical twin.
	3. Emotion and meeting of speech
4. Speech is transmitted in a compressed form.
5. The MCS:
	1. Creates a full description of the 3D visual space using the table, the avatars’ bodies and the heads and faces of the avatars’ bodies.
	2. Collects the speeches from the different participants.
	3. Assigns the spatial coordinates of the avatars in the MCS to the speeches.
	4. Sends the description of the 3D audio-visual space to each participant.
6. Each participant:
	1. Creates the 3D audio-visual space according to their preferences.
	2. Navigates the 3D audio-visual space without moving their avatar.

In other words, the MCS does not create and send the 3D visual space because that would be very demanding on bandwidth, but only the description. The 3D AV space is created locally by each participant.

### Steps

1. Each participant (sending side)
	1. Has an acoustic echo canceller.
	2. Sends before the meeting:
		1. Model of the body of the avatar.
		2. Model of the head and face of the avatar.
		3. Files containing any 2D- or 3D audio-visual presentation.
	3. Has a video camera that:
		1. Is pointed at the participant.
		2. Detects/sends head and face movements, emotion and meaning.
		3. Recognises the speaker.
		4. Transmits ii. and iii.
	4. Has a microphone that:
		1. Captures the environment audio.
		2. Separates speech from environment sound.
		3. Sends compressed speech.
		4. Detects and sends emotion-meaning.
		5. Recognises the speaker.
		6. Transmits iii., iv and v.
	5. Sends visual messages, e.g., raising a hand, calling for silence, in a coded form.
	6. Transmits
		1. Appropriate pointers to previously sent presentation(s).
		2. Information about the portion of the presentation that is being shown.
		3. Authorisation to other participants to control some aspects of the presentation.
2. MCS
	1. Receives speech signals with their identities.
	2. Describes a 3D visual scene with table and chairs.
	3. Describes avatars’s animations using 1.c.ii, 1.c.iii and 2.d.iv. from each participants.
	4. Describes the animation of a limited part of the body by using 1.e.
	5. Sends each participants
		1. b., c. and d.
		2. The speech signals with the corresponding chaor coordinates.
3. Each participant (receiving side)
	1. Creates the visual 3D space using:
		1. The environment with the table.
		2. The chairs in a number equal to the number of avatars.
		3. The presentation.
		4. The avatars whose bodies are static and their heads and faces are animated as received from MCS.
		5. Avatars displaying the visual equivalents of coded messages (e.g., may I speak).
	2. Synthesises the 3D audio space with sound sources at:
		1. Each chair
		2. Location of presentation.
	3. May move in the room to get the best audio-visual experience keeping their avatar at its place.

### Participant TX Reference Model



*Figure 1 – Reference model of a transmitting client*

### MCS Reference Model



*Figure 2 – Reference model of a MCS*

### Participant RX Reference Model

The participant spatially navigates the 3D Audio field “follows” the spatial navigation.



*Figure 3 – Reference model of a Receiving client*

### Comments

This use case is described with a particular partitioning of roles. Other partitionings are possible where some functions that are executed by a participant are instead delegated to MCS.

# MCS description

## Context metadata

1. General features
	1. Real or imagined MCS.
	2. Indoor/outdoor.
	3. Room: setting, furnishing, table and chairs, inside a CAV.
	4. Language: Shared language.
2. Event type
	1. Meeting: board meeting, conference meeting, MPAI meeting etc.
	2. Education: classroom style, evocative place, orderly/scattered arrangement.
	3. Entertainment.
	4. One-to-one consulting.
	5. Work.
3. Interaction attributes:
	1. Among participants
	2. Between participants and objects.

## Avatar metadata

1. Static
	1. Language preference (e.g., EN-US, IT-CH)
	2. Culture (nationality,…), e.g., English spoken with accent
	3. Real/synthetic avatar
	4. Real/synthetic voice
2. Dynamic
	1. Visual
		1. Motion description and animation: initial and subsequent positions
		2. Body parts description and animation
		3. Gesture description and animation
		4. Face description and animation
			1. Eye motion description and animation
			2. Description and animation of lips
		5. Description and animation of a point/object of interest (e.g.. laser from the fingertip)
	2. Speech
		1. Real speech
			1. Description: colour, style, language
			2. Modification: with specified emotions
		2. Synthetic speech from
			1. Text
			2. Text with emotion
			3. Concept with emotion
	3. Visual and speech: detection and animation

## Object description

1. Visual
	1. Real/synthetic
	2. Object position/motion
	3. Object shape, affordance (physical properties)
2. Audio
	1. Real/synthetic
	2. Object position/motion
	3. Object description (ambisonic audio)
3. Visual and audio: association of audio object with a visual object

## General

1. Objects
	1. Authentication (guarantee that an object is what it looks or says it is)
	2. Access (ability to make a specific action on an object)
2. Text
	1. Text analysis
	2. Media annotations to objects
	3. Visually represent mathematical formulae

# AIMs/Workflows required

NB: Context metadata are not included.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Real** | **Virtual** |
| Avatar | Visual | Body motion recognition | Body animation |
|  |  | Gesture recognition | Gesture animation |
|  |  | Face Emotion recognition | Face emotion animation |
|  |  | Face meaning recognition | Face meaning animation |
|  |  | Head motion recognition | Head animation |
|  |  | Eye motion recognition | Eye animation |
|  |  | Face recognition | Face reproduction, Authentication |
|  | Speech | Speaker recognition | Speech synthesis, Authentication |
|  |  | Speech recognition | Speech synthesis, Face animation |
|  |  | Language understanding | Speech synthesis, Face animation |
|  |  | Emotion recognition | Speech synthesis, Face animation |
|  |  | Language translation | Language translation |
|  | Text | Language understanding | Speech synthesis, Face animation, Authentication |
|  |  | Emotion recognition | Speech synthesis, face animation |
|  |  |  | Language translation (same as for speech) |
|  | Vis/Spe | Emotion fusion (T-S-F-G-B) | Speech synthesis, face animation, body animation |
|  |  | Meaning fusion (T-S-F-G-B) | Speech synthesis, face animation, body animation |
| Object | Visual | Object recognition | Visual object synthesis/reconstruction |
|  |  | Object position/motion | Visual object synthesis/reconstruction |
|  |  | Object metadata extraction (e.g., affordance, semantics) | Visual object synthesis/reconstruction |
|  |  |  | Avatar position selection |
|  | Audio | Sound separation | Audio object synthesis/reconstruction |
|  |  | Sound source recognition | Audio object synthesis/reconstruction |
|  |  | Sound classification | Audio object synthesis/reconstruction |
|  |  | Sound selection | Audio object synthesis/reconstruction |
|  |  | Sound metadata extraction | Audio object synthesis/reconstruction |
|  |  | Acoustic echo cancellation |  |
|  |  |  | Audio object position selection |
|  |  | Audio scene personalisation |  |
| Security | Vis/Aud |  | Object authentication |
|  |  |  | Object access |
| Scene |  |  | Audio/Speech/Visual Scene creation and interaction |
|  |  | Audio/Speech/Visual Scene personalisation & interaction |  |
|  |  |  | Cultural translation |

# Data formats

|  |  |  |  |
| --- | --- | --- | --- |
| **Data format** | **L1** | **L2** | **Initial requirements** |
| Context |  |  | Define format elements  |
| Persona |  |  | Define format elements  |
| Avatar description | Face | Identification | For security |
|  |  | Emotion |  |
|  |  | Meaning |  |
|  |  | Text |  |
|  | Speech | Identification | For security |
|  |  | Emotion |  |
|  |  | Meaning |  |
|  |  | Text |  |
|  | Gesture | Emotion |  |
|  |  | Meaning |  |
|  |  | Text |  |
|  | Motion |  |  |
| Visual object description | Real/synthetic |  |  |
|  | Coordinates  |  |  |
|  | Shape |  |  |
|  | Affordance |  |  |
|  | Metadata |  |  |
| Audio object description | Real/synthetic |  |  |
|  | Coordinates  |  |  |
|  | Metadata |  |  |

# Terms and definitions

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Access | The credential allowing a participant to act on an MCS object |
| Affordance | The properties of an object that define its possible uses |
| Authentication | The ability to associate an object to its physical twin |
| Colour | Set of characteristics defining the speech uttered by an individual |
| Context | Set of characteristics defining the nature of an MCS |
| Persona | Set of characteristics defining an individual in an MCS |

# References

1. Martin Scheffler, Jan P. Springer, Bernd Froehlich: Object-Capability Security in Virtual Environments; Proceedings of IEEE VR 2008; https://www.uni-weimar.de/fileadmin/user/fak/medien/professuren/Virtual\_Reality/documents/publications/capsec\_vr2008\_preprint.pdf
2. Ceenu George, Daniel Buschek, Mohamed Khamis, Heinrich Hussmann; Investigating the Third Dimension for Authentication in Immersive Virtual Reality and in the Real World; IEEE VR; http://daniel-buschek.de/assets/pubs/george2019ieeevr/george2019ieeevr.pdf