VOXReality

VOICE DRIVEN INTERACTION IN XR SPACES

Definition and Analysis of VOXReality Use Cases V1

D2.1

04-04-2023

Funded by
the European Union

Version	1.0
WP	WP2
Dissemination level	Public
Deliverable lead	NWO-I
Authors	Sueyoon Lee, Moonisa Ahsan, Irene Viola, Pablo Cesar, Drakoulis Petros, Konstantoudakis Konstantinos, Mpiliousis Stefanos, Papadopoulos Georgios, Zarpalas Dimitrios, Yusuf Can Semerci, Manuel Toledo, Elena Oikonomou, Clayton Gordy
Reviewers	Ana Luísa Alves (F6S); Stavroula Bourou (SYN); Spiros Borotis; Nikos Achilleopoulos (MAG)
Abstract	This deliverable provides the reader an overview of the activities of WP2, focusing on the description of the use cases and a first version of the user and technical requirements. The list of requirements will be further defined, when considering the integration and validation requirements (D2.3).
Keywords	*Use Cases, User Requirements, Technical Requirements, User-Centred Methodology, Focus-Groups.*
License	 This work is licensed under a Creative Commons Attribution-No Derivatives 4.0 International License (CC BY-ND 4.0). See: https://creativecommons.org/licenses/by-nd/4.0/

Dissemination Level	
PU	**Public**
PP	Restricted to other programme participants (Including the Commission Services)
RE	Restricted to a group specified by the consortium (Including the Commission Services)
CO	Confidential, only for members of the consortium (Including the Commission Services)

Nature	
PR	Prototype
RE	**Report**
SP	Specification
TO	Tool
OT	Other

Version History

Version	Date	Owner	Author(s)	Changes to previous version
0.1	2022-12-22	NWO-I	Pablo Cesar	ToC and initial contents
0.2	2023-01-10	NWO-I	Sueyoon Lee	Section 4.2 - Initial list of user requirements before the design sessions
0.3	2023-03-03	CERTH	Drakoulis Petros, Konstantoudakis Konstantinos, Mpiliousis Stefanos, Papadopoulos Georgios, Zarpalas Dimitrios	Section 6.2
0.4	2023-03-03	UM	Yusuf Can Semerci	Added Chapter 3
0.5	2023-03-09	NWO-I	Sueyoon Lee	Section 2, Section 4.1
0.6	2023-03-09	VRDays	Manuel Toledo	Section 5.2
0.7	2023-03-10	Athens Festival	Elena Oikonomou	Section 5.1
0.8	2023-03-10	HOLO	Clayton Gordy	Section 5.3
0.9	2023-03-11	NWO-I	Moonisa Ahsan	Section 4.2
0.10	2023-03-12	NWO-I	Pablo Cesar and Irene Viola	Sections 1 and 6; consolidation of the text
0.11	2023-03-12	NOW-I	Moonisa Ahsan	Section 5.2
0.12	2023-03-16	NWO-I	Sueyoon Lee	Section 7
0.13	2023-03-20	NWO-I	Moonisa Ahsan, Sueyoon Lee	Section 7 (Table 9, 10, 11)
0.14	2023-03-22	NWO-I	Moonisa Ahsan	Section 7.1, Overall Revisions
0.15	2023-03-22	NWO-I	Pablo Cesar, Irene Viola, Sueyoon Lee, Moonisa Ahsan	Full Draft
0.16	2023-03-28	SYN	Stavroula Bourou	Peer revision
0.17	2023-03-28	F6S	Ana Luisa Alves	Peer revision
0.18	2023-03-30	MAG	Spiros Borotis	Peer revision
1.0	2023-04-04	NWO-I	Moonisa Ahsan Irene Viola Pablo Cesar Sueyoon Lee	Implementation of feedback from partners and overall revisions

Table of Contents

List of Abbreviations & Acronyms

3D	:	Three-dimensional
AI	:	Artificial Intelligence
AR	:	Augmented Reality
AR3S	:	Augmented Reality Engineering Space
ASR	:	Automatic Speech Recognition
CAD	:	Computer-Aided Design
CV	:	Computer Vision
DA	:	Digital Agent
FAQ	:	Frequently Asked Questions
HMD	:	Head-Mounted Display
ISAR	:	Interactive Streaming for Augmented Reality
ML	:	Machine Learning
NLG	:	Natural Language Generation
NLP	:	Natural Language Processing
NLU	:	Natural Language Understanding
NMT	:	Neural Machine Translation
PC	:	Personal Computer
PDF	:	Portable Document Format
Q&A	:	Questions and Answers
SDK	:	Software Development Kit
UI	:	User Interface
VCE	:	Virtual Conference Environment
VFX	:	Visual Effects
VL	:	Vision-Language
VR	:	Virtual Reality
XR	:	eXtended Reality

List of Figures

List of Tables

Executive Summary

This deliverable provides the reader an overview of the activities of WP2, focusing on the description of the use cases and a first version of the user and technical requirements. The list of requirements will be further defined, when considering the integration and validation requirements (D2.3).

Section 2 overviews the timeline and adopted user-centred methodology (M06). Initially, a number (N=7) online meetings took place from October to December 2022 with at least one representative of each partner. The calls (Section 3) helped in better defining and understanding the use cases. As a result, an initial set of requirements were identified (Section 4). Based on the initial set of requirements, three (N=3) in-situ visits to the premises of the use case owners were scheduled (Section 5). In these focus group workshops, the initial set of requirements were refined and further developed together with the use case owners.

Thanks to the user-centred methodology, this deliverable reports a detailed definition of the use cases (section 6) and the first version of the user and technical requirements (section 7). These three will be used as an input for the next iteration of the work, with a focus on the architecture, integration and validation requirements.

1 Introduction

1.1 VOXReality concept and approach

VOXReality is an ambitious project with the goal of facilitating and exploiting the convergence of two important technologies, NLP and CV. Both technologies are experiencing a huge performance increase due to the emergence of data-driven methods, specifically ML and AI. CV/ML are driving the XR revolution beyond what was possible up to now, and speech-based interfaces and text-based content understanding are revolutionizing human-machine and human-human interaction. VOXReality will employ an economical approach to combine these two. The Project will pursue the integration of language-based and vision-based AI models with either unidirectional or bidirectional exchanges between the two modalities. Vision systems drive both AR and VR, while language understanding adds a natural way for humans to interact with the back-ends of XR systems or create multimodal XR experiences combining vision and sound.

The results of the Project will be twofold: [a] a set of pretrained next-generation XR models combining, in various levels and ways, language and vision AI and enabling richer, more natural immersive experiences that are expected to boost XR adoption, and [b] a set of applications using these models to demonstrate innovations in various sectors. The above technologies will be validated through three use cases: VR Conferences, Augmented Theatres and Training Assistants.

1.2 Purpose of the deliverable

This deliverable provides a detailed analysis of the VOXReality's use cases and a list of technical and user requirements. It concludes with an initial information flow diagram towards the architecture of the system and the development/deployment plan, which will be provided in the subsequent deliverable D2.2 Definition and Analysis of VOXReality Use Cases V2.

1.3 Intended Audience

This deliverable is addressed, first of all, to the consortium, with an emphasis on the technical research team and the end-users. It is also addressed to the interested public and potential participants of the open calls, with an interest on the system and technologies to be developed in VOXReality.

1.4 Structure

The document is divided into eight sections, providing a list of requirements (user and technical) and a detailed description of the use cases. First, Section 2 details the overall user-centered methodology used for gathering the requirements and capturing the essence of the use cases (weekly calls and focus groups). Section 3 details the initial weekly calls between the different partners (technical, user-centered and use cases) which allowed gathering of an initial set of requirements and created a shared understanding. Section 4 provides the initial set of requirements, resulting from this first phase. Later, Section 4 describes the in-situ visits to the use case owners, as interactive focus group workshops, for further detailing the requirements and use cases; and providing the user-centered partners a better understanding of the context of the use cases. Based on the methodology, Section 5 provides a concrete and specific definition of the three use cases and Section 6 lists the identified users and technical requirements. Section 7 concludes the deliverable.

2 Methodology

In this Project, the consortium has agreed to adopt a user-centric approach in the requirement-gathering procedure which comprises two phases: 1) Weekly calls and 2) In-Situ visits as shown in Figure 1 below.

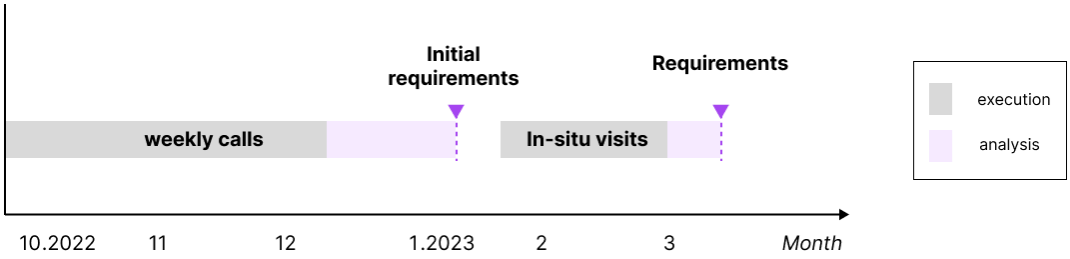


Figure 1. Methodology timeline

2.1 Weekly calls

Weekly calls had taken place between the use case partners and the consortium (technical and user-centered team) to scope the use cases and align viewpoints. Seven online meetings took place from October to December 2022 via Microsoft Teams platform, each of which lasted around 2 hours. At least one representative from each partner participated in the calls, and minutes were recorded, documented, and distributed to the consortium after each call to align understanding between partners.

The outcome of the meetings includes three use case documents filled by the use case partners (VRDays, AEF, HOLO), which provided an initial version of the use case with details such as scenario, required technology, needed equipment, potential assessment protocol, and its target population. The minutes and three use case documents were analysed and turned into an initial list of user requirements, which then were used to scope and prepare the in-situ visits (see section 2.2).

2.2 In-Situ Visits

Based on the initial user requirements gathered during the weekly calls, extended user research was undertaken via in-situ visits to each use case partner. Each meeting was prepared as an 'interactive focus group workshop' where the goal was to understand the latent user needs and gather deeper insights through playful and participatory activities.

The workshops were conducted on January 27th (Amsterdam, VRDays), January 30th (Athens, AEF), and February 28th (Ismaning, HOLO) 2023. Each session lasted up to three hours, depending on the number and characteristics of participants (Figure 2).

(A) Amsterdam, VRDays **(B)** Athens, AEF **(C)** Ismaning, HOLO

Figure 2. Participants discussing their ideas during the VRDays, AEF and HOLO workshop

The workshops were conducted by a moderator accompanied by an assistant. Tailored presentation slides and activities sheet materials were prepared to form a creative and structured workshop (Figure 3). These materials were based on the initial list of requirements and description of the use cases. The detailed experiment setup can be found in Sections 4.1 VR Conference, 4.2 Augmented Theatre and 4.3 Training Assistant.

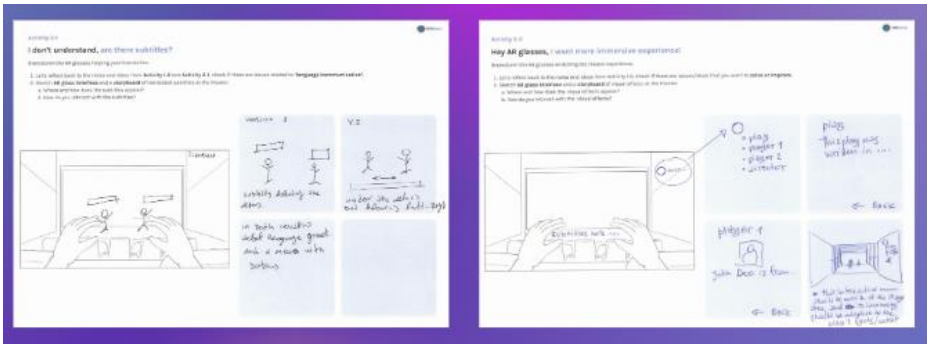


Figure 3. Results of the activity sheets from the AEF workshop

After the workshop, all the activity sheets were collected, documented, and the voice recordings of the participants were transcribed. The different forms of workshop outcomes were analysed by the user-centred research team and turned into a final list of user requirements, which can be found in Sections 6.1 VR Conference , 6.2 Augmented Theatre and 6.3 Training Assistant. These visits helped the use case partners to further reflect over their use cases, resulting in a more detailed description of them, which can be found in Section 5.

3 Details of Weekly calls

The VOXReality consortium comprises teams specialised in user-centred research, technical research and end-users. As expected, each team possesses unique terminologies, perspectives on use cases, and levels of familiarity with the expertise of other teams. In the beginning of the Project the consortium scheduled weekly online calls led by the Scientific Coordinator to synchronize the teams' viewpoints, enhance their comprehension of each other's terminologies, facilitate familiarity with the diverse disciplines represented within the consortium and drive the discussions towards user requirements, technical requirements and formal use case definitions.

3.1 Weekly calls *modus operandi*

The consortium conducted the 2-hour alignment meetings via Microsoft Teams platform, utilizing the project-specific channel. Partners were instructed to ensure at least one representative from each partner attended and participated in the discussions. A total of seven online meetings took place on October 21st, October 31st, November 7th, November 14th, November 21st, December 5th, and December 12th. Simultaneously with each meeting, the minutes were recorded and collected and the minutes from the previous meeting were distributed to the consortium before the next meeting.

The first two meetings focused on comprehending the partners' perspectives regarding the use cases. During the initial meeting held on October 21st, the discussions revolved around the questions of the use case designers (VRDAYS, AEF and HOLO) related to the utilization of technological components in each use case. On October 31st, the second meeting was devoted to the technical partners' point of view, where the questions related to the execution of each use case experiment were addressed by the use case designers.

The structure of the remaining meetings consisted of three equally separated sessions dedicated to each use case individually and a final shorter session to discuss technical or use case related topics that would affect all use cases and technical components. Furthermore, these meetings were used to take initial decisions on certain technical and use case related aspects by taking advantage of the participation of all members of VOXReality consortium.

The outcome of these meetings was intended to drive the discussions towards user, functional and technical requirements. To facilitate these discussions in the next steps (described in chapter 4), a use case definition template has been prepared to produce complete documents where all details could be collaboratively recorded regarding each use case. Table 1 presents the sections of the mentioned template document with the description of each section. The meetings held in December were used to complete the contribution of each partner to these use case documents and the meetings contributed to the clarification of any misunderstandings and missing details.

Table 1. Description of the use case definition template

Section	Explanation
Title	Title of the use case
Description	High-level description and goals of the overall use case
Scenario	Description of the specific scenario. (How long is it? Description of a user's experience in the scenario. Maybe a step-by-step explanation of one walkthrough.)
Technologies	The software components that will be used. (How and where they should be utilized? What should they provide considering their and the proposal's limitations?)

Equipment	Hardware configurations of the scenario such as laptop, headset, mobile, controller, microphone, speaker etc.
External Context	The surrounding sentences are internal context. What could be the external ones in the scenario? For example, slides, a dictionary of words, description of the scene, agents' knowledge base e.g., the venue, the manual of a training.
Assessment Protocol	The criteria (what will be assessed) and tools (questionnaires etc.) that will be used in the scenario. Where will the scenario take place? The number of participants in each session. Total number of participants in the scenario to be reached.
Target Population	The target users/audience of the scenario. Demographic background to be considered e.g., age, gender, language, technology acceptance, technology experience. Languages to be used. If possible/needed, exclusion criteria (maybe eyeglasses for headsets?).

3.2 Results

The alignment meetings were held until the start of the focus group meetings, resulting in four significant outcomes:

- The partners involved in the Project were able to familiarize themselves with the various disciplines represented within the consortium.
- All members were informed about the terminologies that will be utilized throughout the Project.
- There was a consolidation of viewpoints regarding the design, implementation, and execution of each use case.
- The use cases were defined and documented using a common structure.

The outcomes of these meetings enabled the consortium to easily transition from internal meetings to focus groups involving other stakeholders, such as end-users and application designers of each use case; including conference organizers for VR conference use case, trainers for Training-Assistant use case and directors and actors for Augmented-Theatre use case. Furthermore, the common language, mutually agreed terminologies and the formal representation of use case definitions used among the VOXReality researchers provided a common voice to reach these stakeholders easily. Finally, the use case definitions provided the foundation of the functional and technical requirements as well as enabled the researchers to initiate the implementation of independent technological components of VOXReality.

3.3 Coordination with technical partners

Following the initial definition of the use cases, the technical partners met in person on November 30th, in a general meeting held during Immersive Tech Week in Rotterdam, The Netherlands, to refine and consolidate the technology requirements based on the input from the use-case owners. During the meeting, the partners identified a common architecture to be used in all use cases, and decided on its modularity and on the timeline of its implementation. For each use case, they identified which modules should be implemented and/or activated, what input would be provided (both as direct input and as context), and what is the envisioned output. The outcome of the meeting was a roadmap for technical partners to proceed with their development, which informed the subsequent weekly meetings.

4 Use cases initial requirements

This section lists the initial requirements for each VOXReality's use cases (VR Conference, Augmented Theatre and Training Assistant), based on the analysis of the weekly calls MoM and on the initial description of the use cases provided by the use case owners.

4.1 VR Conference

The goal of the VR conference use case is to provide user navigation and language translation at a virtual conference. The conference will only be held in VR, not hybrid, but a stream of video can happen in the virtual environment. The virtual venue consists of entrances, lobby areas, trade shows, meeting rooms, social areas and a conference/plenary room. The language transition will mainly support the 1 to many conferences setting.

User interaction

Users will be represented by virtual avatars in the conference venue. User avatars will be programmed to deliver visual feedback complementing available real-time multilingual translation and captioning. Users can ask questions to the virtual agents during the navigation, and also allowed to engage with speakers and guests during Q&A sessions.

Virtual agents

Virtual agents are virtual only setting, and help navigate the user by informing about the context and contents of the rooms. Everything happening on the way to the rooms could potentially be considered a part of the virtual agents (digital navigation assistants).

Virtual agents can answer questions from the users and deliver personalized info related to the conference programme-related activities. A dedicated virtual agent will engage with users when they log into the virtual conference venue.

Table 2. Initial requirements for VR Conference use case

Type	Requirements	*Notes*	Phase gathered
Environment	The experience is in VR.		conference use case scenario document
Objective	User navigation and language translation at the virtual conference.		
Interaction	VR controllers will provide navigation, menu selections and haptic interactions.		conference use case scenario document
Setup	The conference is only in virtual, not hybrid.		technical-use case meeting 1
Setup	A stream of video can happen in the virtual environment.		technical-use case meeting 1
Input	Contextual visual information will be used for path generation (navigation)		technical-use case meeting 6
Input	Voice communication will be used for path generation (navigation) will use.		technical-use case meeting 6
Input	Instruction generation (instruction) will be based on the dialog (text) and visual information.		technical-use case meeting 6

Source contents	The length of the speech should not be longer than 30'.		conference use case scenario document
Source contents	The speech is based on a predefined conference format.	*ideally, ex: TED, Pecha Kucha	conference use case scenario document
Output	Language output is in English, but can be translated.		technical-use case meeting 6
Output	All outputs from the models will be text.		technical-use case meeting 3
Venue design	The virtual venue consists of entrances, lobby areas, trade shows, meeting rooms, social areas and a conference/plenary room.		conference use case scenario document
User interaction	Users will be represented by virtual avatars in the venue.		conference use case scenario document
User interaction	User avatars will be programmed to deliver visual feedback complementing available real-time multilingual translation and captioning.		conference use case scenario document
User interaction	Users are allowed to engage with speakers and guests during Q&A session.		conference use case scenario document
User interaction	In the conference room, users can choose between two in-room view options.	*The options include a full screen and conference room setting.	conference use case scenario document
User interaction	Both in-room views (full screen and conference room) include on-demand virtual assistance and real-time multilingual translation and captioning.		conference use case scenario document
User interaction	At the end of every conference session, users will be presented with all available options.	*The options include vacating the conference room for the conference venue, re-visit the conference and/or exiting the event completely.	conference use case scenario document
User interaction	Users can ask the agent questions during the navigation.	*ex) "Please describe the scene for me.", "What is in this room?", "Is there an empty chair in the room?" etc.*	technical-use case meeting 3
Translation	Language translation assists the 1 to many conferences setting.	*1: many is the main focus use case scenario	1:1 meeting - CWI
Virtual agents	Virtual agents help navigate the user.		technical-use case meeting 2
Virtual agents	Virtual agents are virtual only settings.	*no physical agents	technical-use case meeting 1
Virtual agents	Virtual agents inform the users about the context/content of the room.		technical-use case meeting 2
Virtual agents	Everything happening on the way to the rooms could potentially be considered a part of the digital navigation assistant.		technical-use case meeting 3
Virtual agents	Virtual agent can integrate scene description from the Visual Language models.		technical-use case meeting 3

Virtual agents	Virtual agents answer questions from the users.		conference use case scenario document
Virtual agents	A dedicated virtual agent will engage with users when they log into the virtual conference venue.		conference use case scenario document
Virtual agents	Virtual agents deliver personalized info feeds related to the conference programme-related activities.		conference use case scenario document
Virtual agents	The agent will act as the dialogue generator and the component communicating with the user.		technical-use case meeting 6

4.2 Augmented Theatre

The goal of the AR Theatre use-case is to provide language translation and VFX experience with AR glasses during the play. The target audience is Greek or Greek play international audiences with different age, gender, technology acceptance and using different language. The experiment will take place at the avant-garde theatre, modern, indoor, instead of the outdoor ancient theatre.

Play design

The play, which will be used for the AR experience, is an ancient tragedy with the length of 10-12 minutes. The play is in English and not dialogue heavy. Maximum 2-3 actors are on the stage at the same time to avoid the technical and design complexity. Additionally, one narrator, who is not on the stage, could be included.

VFX

VFX will be triggered by certain words or phrases from the play. A narrator could be the person who triggers the VFX. The style of the visual effects should be artistically relevant to the opera, and delivered in a high quality rather than having ambitious challenges. The location of the visual effects should not exceed the range of the stage.

Evaluation

Two users will watch the play at the same time due to the availability of the AR devices, and up to 50 people will participate in the experiment. The evaluation may include combined demonstration – e.g., 5-minute narration + 5-minute dialogue, 5-minute VFX + 5-minute dialogue/narration with VFX. While user experience is the main evaluation of the Project, the creator's experience of how their work is presented through the technological medium (VFX) could also be evaluated.

Table 3. Initial requirements for Augmented Theatre use case

Type	Requirements	*Notes*	Phase gathered
Environment	The experience is in AR.	Environment	
Objective	Language translation and VFX experience with AR glasses at the theatre.	Objective	
Setup	The experiment takes place at the avant-garde theatre	*Modern, indoor theatre*	technical-use case meeting 6
Source contents	Contextual information (summary, description etc) will be provided instead of the entire script.		technical-use case meeting 1

Device	The simulator sickness limits for the selected AR glasses should be checked.		technical-use case meeting 2
Target user	The audience is Greek or Greek play international audience		technical-use case meeting 1
Target user	The audience has different technology acceptance or experience		technical-use case meeting 1
Target user	The audience is different age, gender and use different language	*People with disabilities will not be considered; we can put emphasis on the potential of VOX technologies in providing access to people with disabilities (maybe auditory disabilities)	technical-use case meeting 1
Play design	The play is in English.		technical-use case meeting 1
Play design	The play is not dialogue heavy		technical-use case meeting 2
Play design	Maximum 2-3 actors are on stage at the same time.		technical-use case meeting 2
Play design	A narrator, who is not on stage, could be included.		technical-use case meeting 2
Play design	The play includes a short dialogue on stage.		technical-use case meeting 2
Play design	The length of the play will be 10-12 minutes.		technical-use case meeting 4
Play design	Play is an ancient tragedy	*planned to be	technical-use case meeting 6
VFX	Certain words or phrases will trigger VFX.		technical-use case meeting 1
VFX	A narrator could be the person who triggers the VFX		technical-use case meeting 2
VFX	Style of the visual effects should be artistically relevant to the opera	Not too cartoonish	technical-use case meeting 6
VFX	Location of the visual effects should not exceed the range of the stage		technical-use case meeting 6
VFX	The visual effect should be delivered in a high quality rather than having ambitious challenges.		technical-use case meeting 6
Evaluation	Up to 50 people will participate in the experiment		theatre use case scenario document
Evaluation	At least two users watch the play at the same time with AR devices	*based on the availability of AR devices)	technical-use case meeting 4
Evaluation	The evaluation should focus on the experience of users with the technological features of VOX.		technical-use case meeting 2
Evaluation	Users should evaluate VFX.		technical-use case meeting 2
Evaluation	The evaluation may include combined demonstration.	For example, 5-minute narration + 5-minute dialogue + 5-minute VFX + 5-minute	technical-use case meeting 2

		dialogue/narration with VFX	
Evaluation	Creator experience could also be evaluated based on rehearsal performance including VFX.		technical-use case meeting 2
Evaluation	Creators could evaluate how their work is presented through the technological medium regarding VFX.		technical-use case meeting 2
Evaluation	The testing of the Project should not influence the normal production of the theatre play.		technical-use case meeting 6
Evaluation	People without glasses should not be neglected from the experience.		technical-use case meeting 6
Evaluation	VFX and the subtitles should not have delays on AR screen.		technical-use case meeting 6

4.3 Training Assistant

The goal of the training use case is to provide machine assembly training using AR glasses. Users will assemble machines with virtual tools, screws, and objects. User interaction is set with hands, without extra controllers. The target user is a beginner trainee with some knowledge about the scenario.

Training design

Users can select different difficulty modes during the assembly experience. In easy mode, 3D model will always fit correctly to its location when in the correct place; in hard mode, user needs to assemble the parts into the correct location and orientation.

User interaction

Users communicate with the agents through voice and hand input. Users can ask for instruction to assemble a machine/object.

Virtual agents

Virtual agents use the visual scenes as context. Virtual agents can provide verbal/textual explanation and display corresponding visual contents. They can show and demonstrate the movements and also direct the user towards objects when necessary. If the virtual agents are in 3D avatar format, they will be animated.

Evaluation

Evaluation may include the speed of the assembly with and without the virtual agent.

Table 4. Initial requirements for Training Assistant use case

Type	Requirements	*Notes*	Phase gathered
Environment	The experience is in AR.		technical-use case meeting 4
Objective	Machine assembly training using AR glasses.		technical-use case meeting 4
Interaction	User interaction is set with hands, without controllers.	*Available at HOLO & hardware SDKs	technical-use case meeting 3
Setup	The place to assemble machine is a physical grey base.		technical-use case meeting 6

Setup	Tools, screws, objects are virtual.		technical-use case meeting 6
Setup	The visual instructions, step-by-step guide can be pre-set.		technical-use case meeting 3
Input	Hand motions and audio input from the user will be used.	*HOLO collect the hand movement videos for additional contents	technical-use case meeting 4
Target user	The target user is a beginner trainee with some knowledge about the scenario.		technical-use case meeting 3
Training design	User can select different difficulty mode.		technical-use case meeting 6
Training design	In easy mode, 3D model always fits correctly to its location when in the correct place		technical-use case meeting 6
Training design	In hard mode, user needs to assemble the parts into the correct location and alight the part's orientation.		technical-use case meeting 6
User interaction	Users communicate with the agents through voice and controller input.		technical-use case meeting 1
User interaction	User can ask for instruction to assemble a machine/object.	*e.g., "Teach me how to this", "Am I doing this right?"	technical-use case meeting 6
Virtual agents	Virtual agents will be animated if in 3D avatar format.		technical-use case meeting 3
Virtual agents	Virtual agents direct the user towards objects that are necessary for the application.	*e.g., certain stance for assembly of a machine part	technical-use case meeting 3
Virtual agents	Virtual agents can show and demonstrate movements that are necessary for the application.		technical-use case meeting 3
Virtual agents	Virtual agents can provide verbal/textual explanation and display corresponding visual contents.		technical-use case meeting 3
Virtual agents	Virtual agents use the visual scene as context.		technical-use case meeting 4
Evaluation	Evaluation may include speed of the assembly with and without the agent.	*Planned to be	technical-use case meeting 3

5 In-Situ Visits

The focus group workshops were conducted with use case owners to understand the latent user needs and gather deeper insights and requirements. Each workshop was conducted at the lead partner location, specifically on January 27th in Amsterdam (VRDays), January 30th in Athens (AEF), and February 28th in Ismaning (HOLO). The following sections describe the detailed experiment setup and procedure (5.1 Methodology) and provide the results of the sessions (5.2 Data Analysis and Results) respectively.

5.1 Methodology

The followed methodology allowed interactive focus group workshops [3][4] of 2-3 hours, where the partners and users were invited to share and brainstorm ideas for a new experience with VR/AR experience. Every workshop was conducted in English by one moderator and one assistant. Activity sheets and presentation slides were accompanied during the workshop as supplementary materials. These materials were developed based on the initial list of requirements. The meetings were recorded for later analysis. No preparatory tasks were required.

5.1.1 VR Conference

The goal of the focus group workshop was to:

1) Understand the users' needs and organizers' wants on virtual agents and language translation at the VR conference; and
2) brainstorm ideas for the role and design of virtual agents and language translation at the VR conference.

The workshop was conducted on January 27th 2023 from 14:00 to 17:00 at Spaces Herengracht[1], in Amsterdam. In total 6 participants attended the workshop, including VRDays conference organizers and end-uses who had previous experience attending the conference.

The structure of the workshop was as follows:

- Introduction;
- Part A: My current conference experience;
- Part B: Bringing virtual agents and language translation;
- Part C: Designing a future VR conference;
- Conclusion.

Figure 4. Participants working on activities during the VR Conference (focus-group) workshop

[1] https://www.spacesworks.com/nl/amsterdam-nl/herengracht/

Introduction

In the beginning of the workshop, the moderator shortly introduced the study and provided an activity workbook sheet, a pen, and stickers. Participants, moderator and assistant introduced themselves with a warm up activity, information including name, position, expertise in VR, etc.

Part A: My current VR conference experience

Part A activities were designed to understand the current VR conference experiences and to find a design space for adopting language translation and virtual agent solutions.

Activity 1-1: My experience with a VR conference

The first activity was to ask participants to recall the time they attended a VR conference. Using the activity sheet, participants reflected and wrote down their experience of participating the VR conference on:

- When and where was it?
- What was the purpose or goal of attending the conference?
- What did you enjoy the most?
- What is the one thing that could have been improved?

Activity 1-2: My conference experience

Following activity 1-1, participants were asked to write down or visualize the activities they did on the timeline when they attended the VR conference, from entering the VR venue to leaving it. After the drawing, they marked positive and negative moments with green and red stickers, marking at least 3 moments that they thought could be improved/assisted/richer in some way, assuming they have a superpower, with blue stickers. After individuals completed the timeline drawing, participants shared and discussed the similarities and differences of the timelines as a pair.

Activity 2-1: My experience with virtual agents

The moderator first introduced the concept of virtual agents to help understand the possibilities with visual and audio examples about existing virtual agents (see Figure 5).

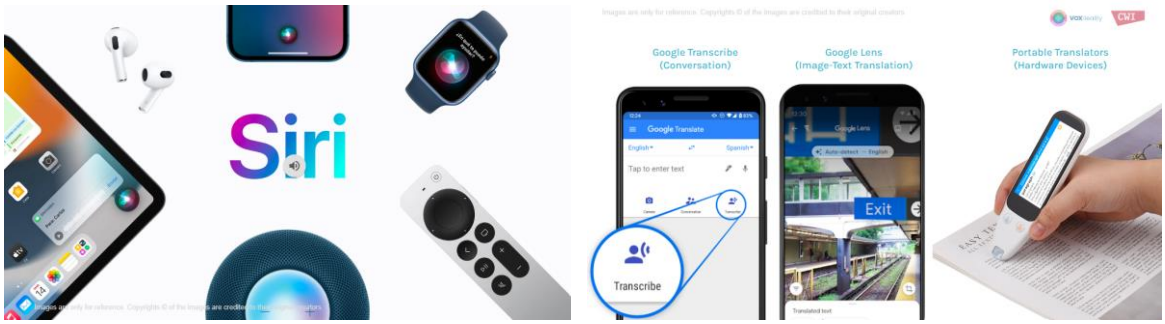


Figure 5. Virtual agent (left) and language (right) translation examples

The participants were later asked to reflect on their experience with virtual agents. They selected one situation and briefly sketched and explained the situation, with supplementary questions (Figure 6):

- Why did you use/had to interact with the virtual agent at that time?
- How did it help solve your problem?
- What was the pain point?

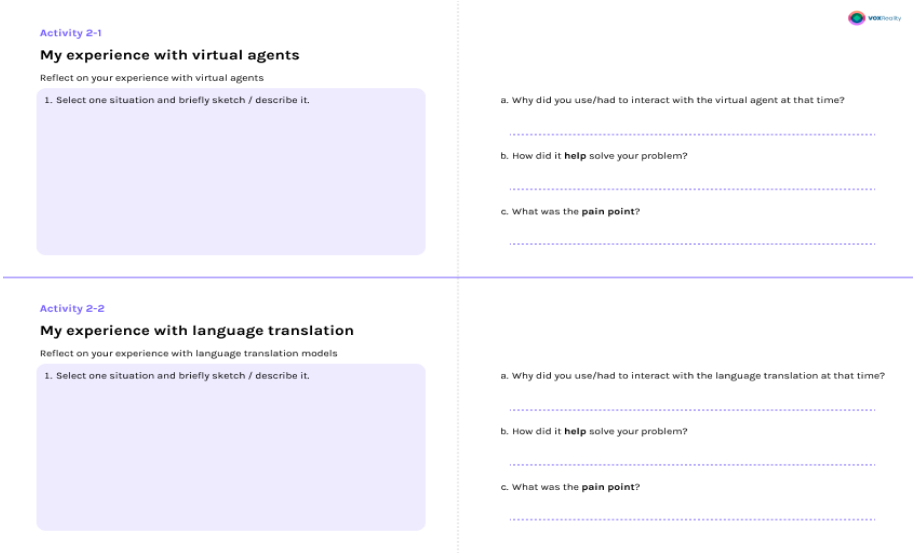


Figure 6. Activity 2-1, 2-2 sheet for the VR Conference use case

Activity 2-2: My experience with language translation

The moderator introduced the concept of language translation models to help understand the possibilities with visual examples.

Participants were asked to reflect on their experience with language translation models. They selected one situation and briefly sketched and explained the situation, with supplementary questions:
Why did you use/had to interact with the language translation model at that time?
How did it help solve your problem?
What was the pain point?

Part B: Bringing superpowers: virtual agents and language translation

Part B activities were designed to help the participants brainstorm ideas on how to apply language translation and virtual agents to solve issues raised in Part A.

Activity 3-1: [Virtual assistant], please help me navigate!

For activity 3, the moderator provided a specific scenario of attending a VR conference and asked the participants to imagine they were faced with the situation. In the first scenario, one had just entered the VR conference venue and wants to attend the 'XR haptics' session in Room A-2, without knowing the room location (Figure 7).

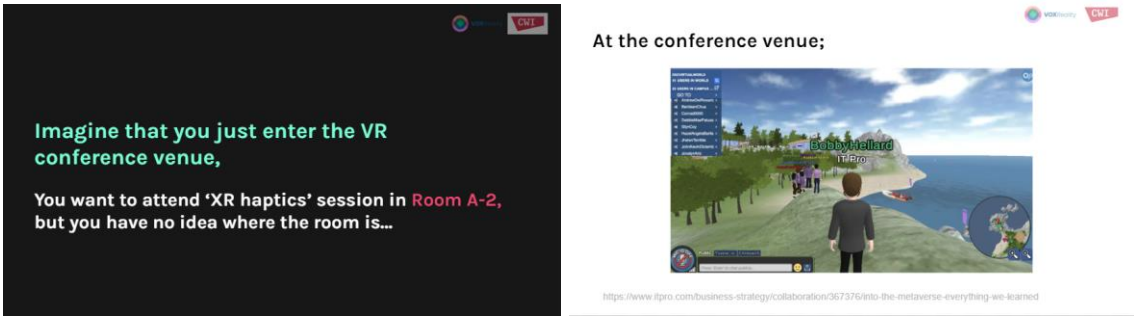


Figure 7. Scenario case and supplementary image on navigating in VR conference

The participants were asked to brainstorm on virtual agents helping their navigation, considering:

- How does the virtual agent look like?
- How do you interact with the virtual agent?

The brainstorming sheet included eight blank boxes – where to sketch the imaginary scenario or screen interface. After finishing the drawing, each participant shared and explained their brainstormed idea.

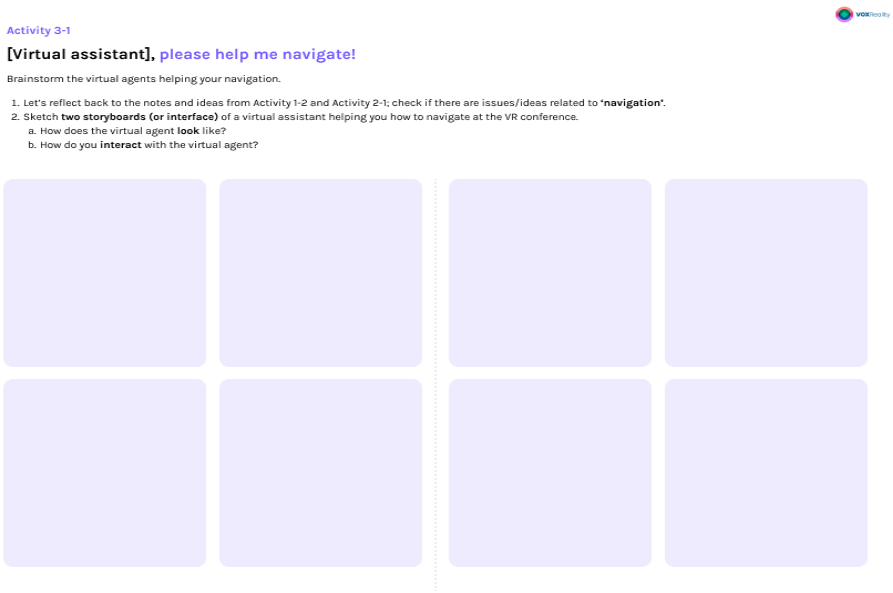


Figure 8. Activity 3-1: Brainstorming sheet for drawing ideal VR conference navigation scenario

Activity 3-2: I do not understand, are there subtitles?

The second scenario was about the subtitles. Participants arrived at Room A-2 and the speaker started to give a presentation. However, the speaker started to talk in French, a language they do not understand, and they seek the language translation. The participants were asked to brainstorm on virtual agents helping the translation, considering:

- Where and how does the subtitles appear?
- How do you interact with the subtitles?

The brainstorming sheet with the same format as activity 3-1 was provided. After finishing the drawing, each participant shared and explained their idea (Figure 9, left). The moderator and assistant placed the activity sheets with the brainstormed ideas on the whiteboard. Participants were provided with four stickers and had to vote for the two best ideas for each scenario, namely navigation and language translation (Figure 9, right).

Figure 9. (Left) participants sharing their brainstormed ideas, (Right) participants voting for the two best ideas for each scenario

Part C: Designing a future VR conference

Part C activities were designed to come together and design the ideal future VR conference as a group based on the accumulated ideas from the previous parts.

Activity 4-1: Four scenarios

The moderator introduced four different forms of virtual assistants that could possibly be integrated into VR conference platforms (Figure 10). Participants were asked to write the order that they would like to use to start exploring the VR space, by selecting a minimum of 1 and a maximum of 5, including their own idea as the fifth option.

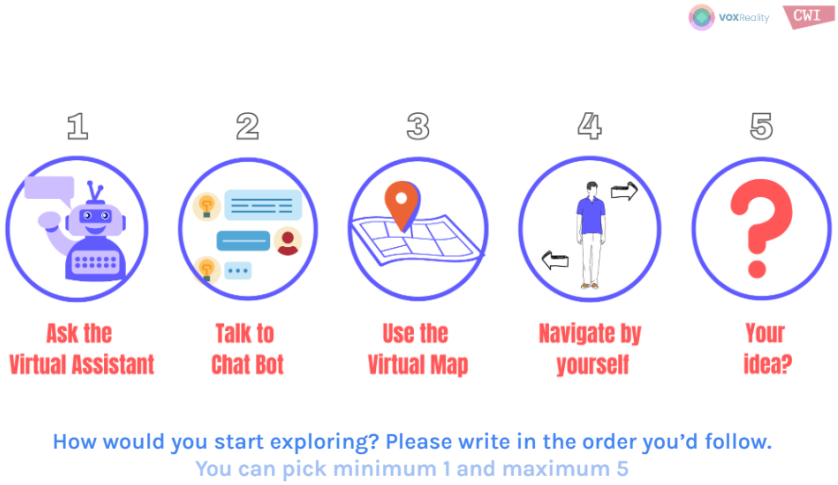


Figure 10. Different forms of virtual assistant in VR conference

Activity 4-2: Brain-drawing – designing an ideal VR conference

Participants were divided into two groups to work on two scenarios: navigation and language translation. They were provided a blank sheet to start the brain-drawing activity, with the steps:

- As a group, sketch an ideal interaction on one topic:
  - How does it look like?
  - How to interact?
- Write the list of components it should have and not have;
- Swap the sketch between two groups and continue on drawing. Repeat once more.
- Each group pitch the generated idea and together reflect on the list of requirements.

Conclusion

The moderator concluded the workshop by appreciating the participants for their active participation and input. The moderator and assistant collected the activity sheets filled with responses and drawing from participants. After the workshop, the researchers analysed activity sheets, along with the voice recordings, and presented the process (5.2 Data Analysis and Results).

5.1.2 Augmented Theatre

The goal of the focus group workshop with AEF was to:

1) Understand the needs and desires of both users' and organizers for subtitles (Language translation) and VFX while watching a theatre play;
2) And brainstorm ideas for the UI and interaction method between the audience and AR glasses.

The workshop (see Figure 11) was conducted on January 30th 2023 from 10:00 to 13:00 at the Theatre Peiraios 260[2], in Athens, Greece. In total 6 people attended the workshop, with both AEF organizers and theatre enthusiasts with diverse profile.

Figure 11. The moderator introducing the workshop to the participants in focus group workshop

The structure of the workshop included the following parts:

- Introduction;
- Part A: My current theatre experience;
- Part B: Bringing superpowers: subtitles and visual effects with AR;
- Conclusion, wrap-up.

Introduction

At the start of the experiment, the moderator gave the participants a short introduction to the study and provided an activity workbook sheet, pen, and stickers. Participants, moderator and assistant shortly introduced themselves with a warm up activity, sharing their name, position, expertise in theatre, expertise in AR, favourite play, etc.

Part A: My current theatre experience

Part A activities were designed to understand the current theatre experiences of the audiences and to find a design space for integrating AR solutions. There were three activities.

Activity 1-1: My experience with a theatre

The first activity was to ask participants to recall the time they watched a theatre play. Using the activity sheet, participants reflected and wrote down their experience of visiting a theatre on:

- When and where was it

[2] https://aefestival.gr/venues/peiraios-260/?lang=en

- Which language that you watched?
- What did you enjoy the most?
- What is the one thing that could have been improved?

Activity 1-2: My visit to the theatre

On an extension to the first activity, participants were asked to write down or visualize the activities they did on the timeline when they visited a theatre, from entering the theatre to leaving it (see Figure 12). After the drawing, they marked positive and negative moments with green and red stickers. They marked at least 3 moments that they thought could be improved/assisted/richer in some way, assuming they have a superpower, with brown stickers. After individuals completed the timeline drawing, participants shared and discussed the similarities and differences of the timelines as a pair.

Figure 12. Activity sheet 1-2: My visit to the theatre (drawing timeline of a visit to the theatre)

Activity 2-1: My experience with language/communication

The next activity was to reflect on **language/understanding/communication issues** during the theatre play. Participants were asked to select one situation and briefly sketch and explain the situation, with supplementary questions:

- Why was it a problem to you?
- Did you take any action to solve the problem? Why? Or why not?
- Will there be a better way to solve the problem? How?

Part B: Bringing virtual agents and language translation

Part B activities were designed to help participants learn about AR and brainstorm ideas on how to apply AR technology to solve issues raised in Part 1.

The moderator first introduced the concept of AR and played the video on "The Future of Augmented Reality: 10 Awesome Use cases[3]" to easy the understanding of AR possibilities with visual examples.

[3] https://youtu.be/WxzcD04rwc8

Activity 3-1: I do not understand, are there subtitles?

For activity 3, the moderator provided a specific scenario of watching a theatre and asked the participants to imagine they were faced with the situation. In the first scenario they were seated in the theatre wearing AR glasses, and two actors on the stage started to speak Korean – a language that the participant could not understand (Figure 13).

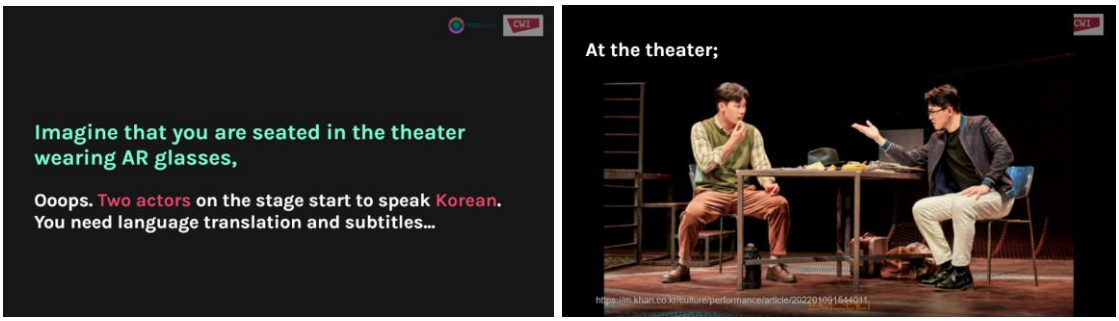


Figure 13. Scenario case and supplementary image on watching a play in a foreign language

The participants were asked to brainstorm on use of AR glasses to assist in the translation, considering:

- Where and how do the subtitles appear?
- How do you interact with the subtitles?

The brainstorming sheet consisted of an illustration of theatre – where to draw the UI of AR glasses – and four boxes – where to sketch the scenario (Figure 14). After finishing the drawing, the moderator took a photo of each idea and shared it on the screen. Each participant shared and explained their idea.

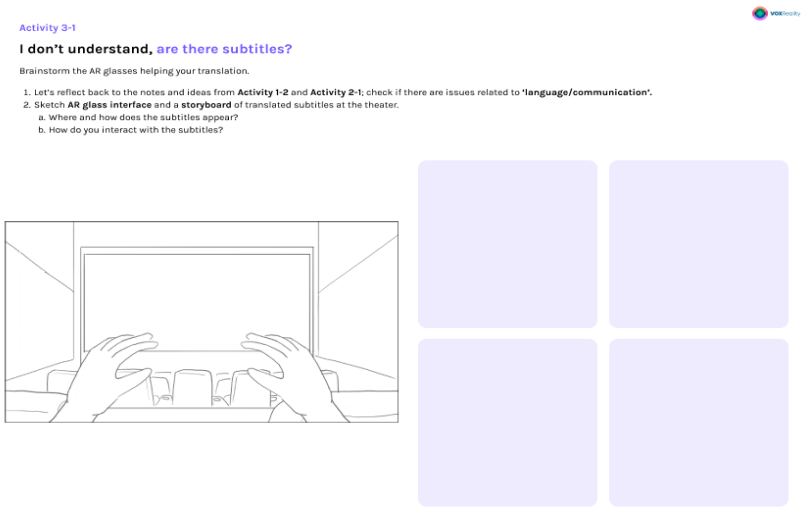


Figure 14. Activity sheet 3-1: Brainstorming sheet for drawing AR glass UI and scenarios for the augmented theatre use case

Activity 3-2: Hey AR glasses, I want more immersive experience!

The second scenario was that they were at the theatre wearing AR glasses, and any text or visual effects could assist them or enrich their experiences. The participants were asked to brainstorm the AR glasses enriching the theatre experience, considering:

- Where and how does the visual effects appear?

- How do you interact with the visual effects?

The brainstorming sheet with the same format as activity 3-1 was provided. After finishing the drawing, the moderator took a photo of each idea and shared it on the screen. Each participant shared and explained their idea.

Conclusion

The moderator concluded the workshop by appreciating the participants for their active participation and input. The activity sheets filled in by the participants were collected for researchers' analysis, together with the meeting recordings. The process can be found later in this document, in section 5.2 Data Analysis and Results.

5.1.3 Training Assistant

The goal of the focus group workshop with HOLO was to:

1) Understand the requirements for providing machine assembly training via virtual agents using AR glasses;
2) And brainstorm ideas for the role and design of virtual agents using AR glasses.

The workshop (Figure 15) was conducted on February 28th 2023 from 15:00 to 18:00 at Holo-Light offices[4] in Ismaning, Germany. Two people – a chief project manager and a scientific researcher – participated in the workshop as company representatives.

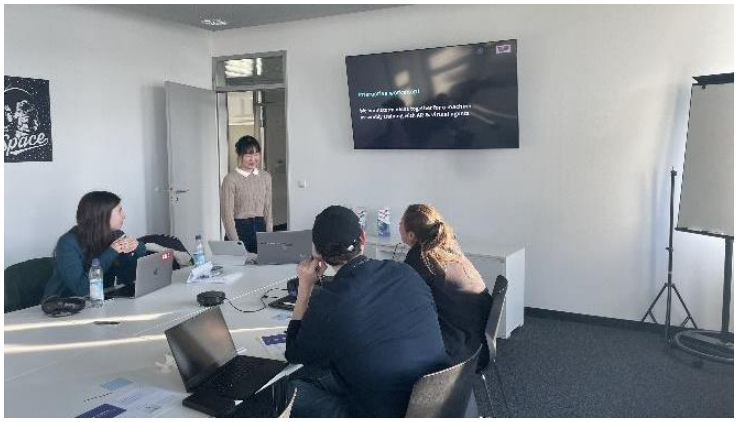


Figure 15. The moderator introducing the workshop to the participants

The structure of the workshop is as follows:

- Introduction;
- Part A: Current machine assembly experience;
- Part B: Brining superpower: voice enabled virtual agent;
- Conclusion.

[4] https://holo-light.com/contact/

Introduction

At the start of the experiment, the moderator gave the participants a short introduction to the study and provided an activity workbook sheet, a pen, and stickers. Participants, moderator and assistant introduced shortly introduced themselves with a warm up activity, information including name, position, expertise in AR, etc.

Part A: Current machine assembly experience

Part A activities were designed to understand the current machine assembly experiences of the users and to find a design space for integrating AR solutions. However, as the participant themselves were not the target user of the product and service, the moderator provided an imaginary persona to help the ideation process (Figure 16). Participants were asked to imagine him/herself as James when working on the activities.

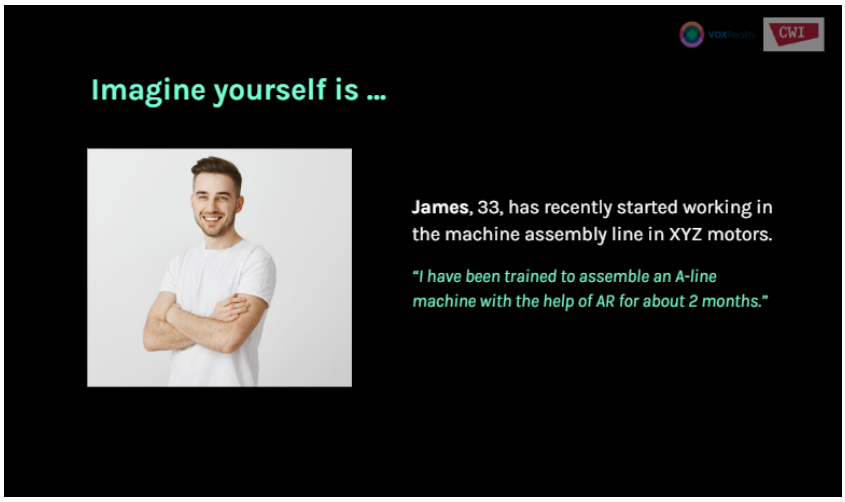


Figure 16. A user persona, James, with a beginner level machine assembly experience

Activity 1-1: My machine assembly experience

For the first activity, the participants were asked to recall the time they assembled machines with AR glasses. Then they had to write down or visualize the activities they did on the timeline, from starting to completing the machine assembly. After the drawing, they marked at least 3 moments that they thought could be improved/assisted/richer in some way, assuming they have a superpower, with brown stickers. After individuals completed the timeline drawing, participants shared and discussed the similarities and differences of the timelines as a pair.

Activity 2-1: I think there should be a better way…?!

The next activity was to select one situation from marked moments from activity 1-1; and briefly sketch and explain the situation, with supplementary questions:

- Why was it a problem to you?
- Did you take any action to solve the problem? Why? Or why not?
- Will there be a better way to solve the problem? How?

Part B: Brining superpower: voice enabled virtual agent

Part B activities were designed to help the participants learn about voice enabled virtual agent and brainstorm ideas on how to apply the technology to solve issues raised in Part A.

The moderator first introduced the concept of voiced enabled virtual agents to help understand the possibilities with visual and audio examples. After, the moderator provided a specific scenario and asked participants to imagine that they were faced with the situation, again, assuming they were James (Figure 17).

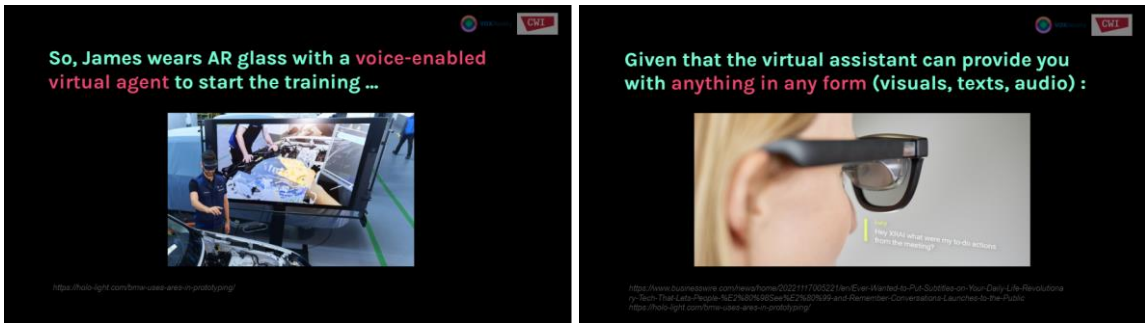


Figure 17. Scenario case and supplementary image on machine assembly experience with AR glasses

Activity 3-1: [Before] So, how do I start the assembly training?

The first scenario was that they wore the AR glasses to start the training but were unclear on how to initiate or operate the training with the AR glasses. The participants were asked to brainstorm starting experience with a virtual agent, considering:

- How would you start the virtual assistant helping your experience?
- What feedback do you expect to see/hear/get from the virtual assistant?

The brainstorming sheet consisted of an illustration of a car with the bonnet opened – where to draw the UI of the AR glasses – and four boxes – where to sketch the scenario (Figure 18). After finishing the drawing, each participant shared and explained their brainstormed idea.

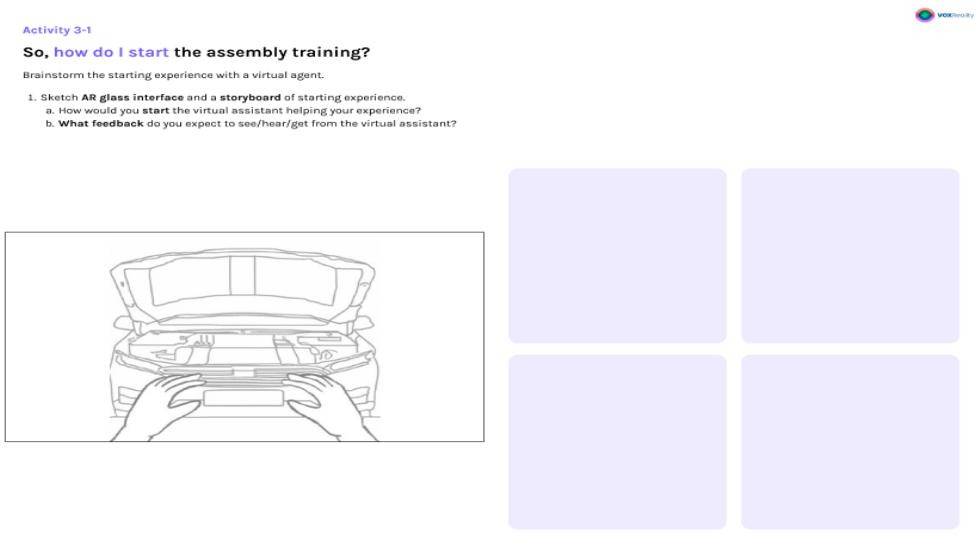


Figure 18. Activity sheet 3-1: Braining storming sheet for ideal AR UI and scenario during machine assembly training

Activity 3-2: [During] Please guide me to the next step!

The second scenario was that they had followed until step 5 of the machine assembly procedure, but now they were stuck and unsure what is the next step. The participants were asked to brainstorm the guidance experience with a virtual agent, considering:

- What action will you take?
- What feedback do you expect to see/hear/get from the virtual assistant?

The brainstorming sheet with the same format as activity 3-1 was provided. After finishing the drawing, each participant shared and explained their brainstormed idea.

Activity 3-3: [Feedback] Hey agent, how was my performance today?

The third scenario was that they had assembled a complete machine, however, they wanted feedback on the entire process so that they can reflect on today's performance. The participants were asked to brainstorm the ending experience with feedback, considering:

- What kind of feedback do you want to receive?
- What form should the feedback be?

The brainstorming sheet with the same format as activity 3-1 was provided. After finishing the drawing, each participant shared and explained their brainstormed idea.

Conclusion

The moderator concluded the workshop by appreciating the participants for their active participation and input. The moderator and assistant collected the activity sheets filled with responses and drawing from participants. After the workshop, the researchers analysed activity sheets, along with the voice recordings, and presented the process (5.2 Data Analysis and Results).

5.2 Data Analysis and Results

This section presents a comprehensive analysis of the qualitative data collected from the user-focused group workshops, highlighting the key themes and insights that emerged from the discussions. The focused groups successfully provided deeper insights from various workshop activities including some in (See Figure 19) about the responses of participants on user-needs and wants on AR/VR agents, translation mechanisms and subtitles in the AR/VR use-cases. The aim for each of the focused group's workshop were:

VR Conference (VRDays):

1) Understand the users' needs and organizers' wants on virtual agents and language translation at the VR conference;
2) And brainstorm ideas for the role and design of virtual agents and language translation at the VR conference.

Augmented Theatre (AEF):

1) Understand the users' needs and organizers' wants on subtitles (Language translation) and VFX while watching a theatre play;
2) And brainstorm ideas for the UI and interaction method between the audience and AR glasses.

Training Assistant (HOLO):

1) Understand the requirements for providing machine assembly training via virtual agents using AR glasses;
2) And brainstorm ideas for the role and design of virtual agents using AR glasses.

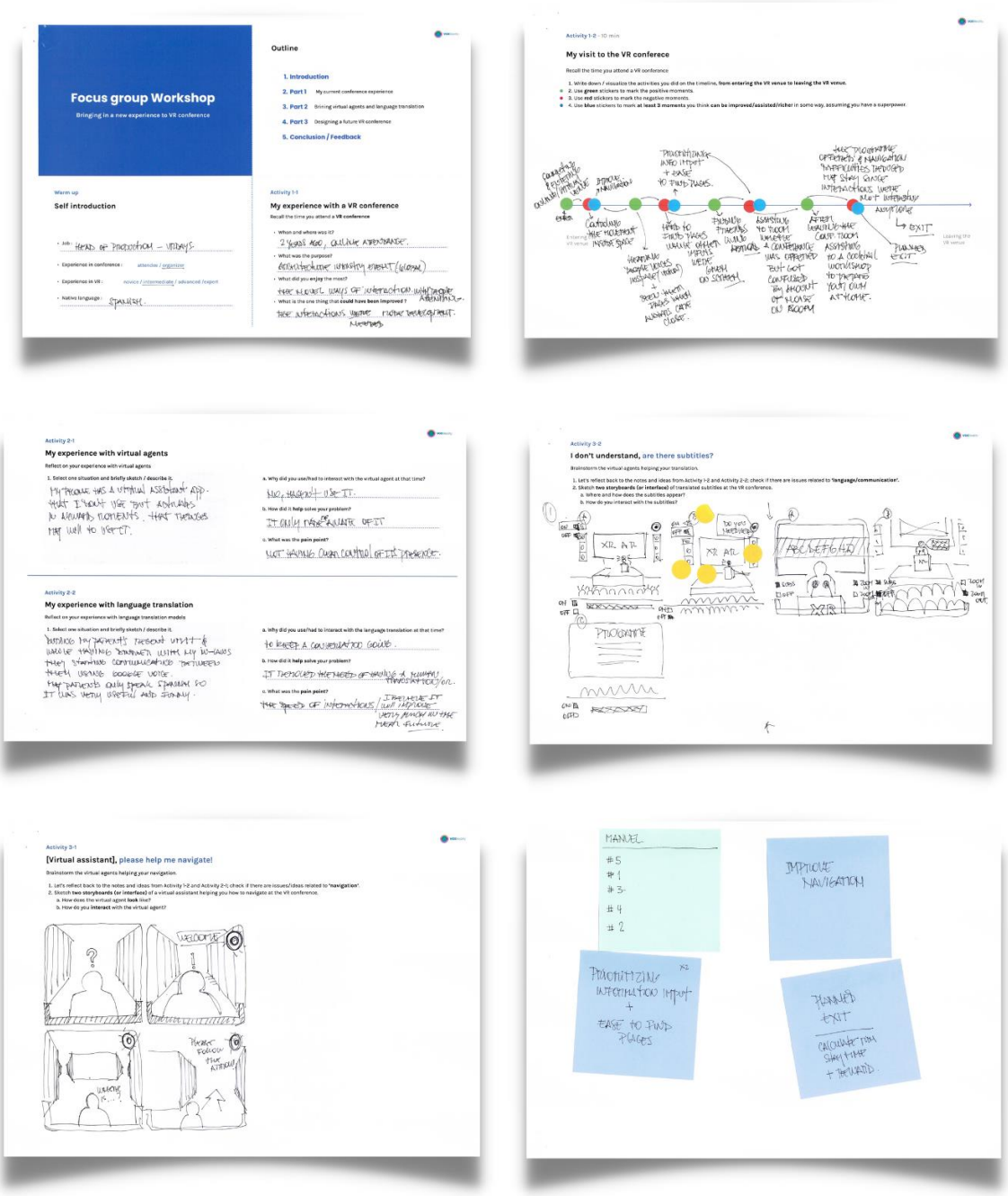


Figure 19. Selected user-activities scans from the data worksheets filled by participants

5.2.1 Participants

The success of virtual conferences largely depends on the needs and preferences of the users who attend them. In order to gather the maximum number of inputs, our focus group brought together a diverse group of participants who had different backgrounds, interests, and experiences with AR/VR technology. By bringing these participants together, we aimed to gather insights and perspectives that could improve the design and implementation of our use-case within the VOXReality project. This section describes the participants in the three focused group workshops, including their demographics, job type, experience in conference/theatre, and prior experience with AR/VR technology. Details of the participants can be found in following Table 5, Table 6 and Table 7.

Table 5. Participants in VRDays focus group

User #	Job	Experience in conference	Experience in VR	Native Language
1	Marketing Manager	Attendee, Organizer	Intermediate	Dutch
2	Director	Organizer	Expert	Dutch
3	Head of Production	Organizer	Intermediate	Spanish
4	Digital Marketing Specialist	Organizer	Intermediate	Hindi
5	Head of Partnerships and Business Development	Organizer	Advanced	English / French
6	Marketing Coordinator	Organizer	Advanced	Hindi

Table 6. Participants in AEF focus group

User #	Job Title	Experience in theatre	Experience in AR	Native Language
1	Retired School Teacher	Intermediate	Novice	Greek
2	Project Manager	Intermediate	Novice	Greek
3	Creative Labourer	Expert	Novice	Italian
4	Actress, Director	Expert	Novice	Greek
5	Composer, Soundtrack Artist, Curator of Mixed Media Projects	Expert	Intermediate	Greek
6	Musician, Guitar Teacher	Intermediate	Novice	Greek

Table 7. Participants in HOLO focus group

User #	Job	Experience in VR	Native Language
1	Chief Project Manager	Expert	German
2	Researcher, Writer	Novice	English

5.2.2 Type of Data

From each of the sessions, a significant amount of data was collected, including:

Text:

- Worksheets with questions and answers referring to the user's introduction, and level of expertise on the VR/AR and related domain(s);
- Their past experiences and anecdotes regarding virtual conferences and events;
- Descriptions of their preferences and interaction details.

Images/drawings:

- Timeline drawings visualizing their experience from start to finish as participants in the given scenarios / activity roles;
- Visual markers (stickers) and sticky notes indicating various user-choices during the activities.

Audio

- Recordings of the interactive discussion sessions, which were transcribed for easier analysis.

5.2.3 Requirements Analysis Methodology

We used Thematic Analysis [6] to analyse the collected qualitative data from the focus groups. It is a method of analysing qualitative data that involves mining of patterns or themes within the data and organizing those themes into categories. It involves reviewing the data multiple times to get a sense of its content, generating labels or tags called codes [7] that reflect relevant features within the data, grouping those codes into potential themes, and then refining and naming those themes in a way that accurately reflects their meaning and relevance to the scientific goal. Due to the wide applications [8] of thematic analysis in contexts of social sciences research and content analysis of media or literature, we used this method to process the acquired data. We identified and assessed the patterns and themes within our data from the workshops and their results are compiled into tabular categories in Chapter 7. We conclude that the accumulated results strongly support the aims and objectives of the focused groups of VR Conference, Augmented Theatre and Training Assistant.

5.2.4 Data Analysis and Classification

This section enlists the comprehensive classification of the qualitative data collected from VOXReality's user-focused workshops, highlighting the major categories and themes that emerged from the discussions. After collecting the data from the participants, it was critical to organize and interpret the data into meaningful knowledge. For this purpose, we classified each dataset into themes and categories defined below, which allowed to identify patterns and relationships in the data, leading to insights and recommendations for the design and development partners in the VOXReality team.

User Journey (Tags)

- **Positive Moments:** The useful or memorable instances in the AR/VR experiences for the user.
- **Negative Moments:** The disturbing or poor instances in the AR/VR experiences for the user.
- **Could be improved (moments):** The instances that were not significantly bad, but could be improved or fixed for a better user-experience.

User Requirements (Tags)

- **New function(s) / feature(s**): Suggestions for new functionalities in the respective application.
- **Behaviour / action / pattern:** A set of actionable practices agreed to be useful by the users.
- **Requirement(s):** Technical or functional requirements for the system(s).
- **Added value:** Features or experiences that provided additional value to the overall experience.
- **Raised issue / concern:** Considerations for the issues or problems to be addressed.
- **Interface (suggestions):** UI ideas for the applications / platforms.
- **Interaction:** User interaction practices for intuitive immersive exploration.

Figure 20. Insights Canvas for VR Conference (VRDays) Use-case

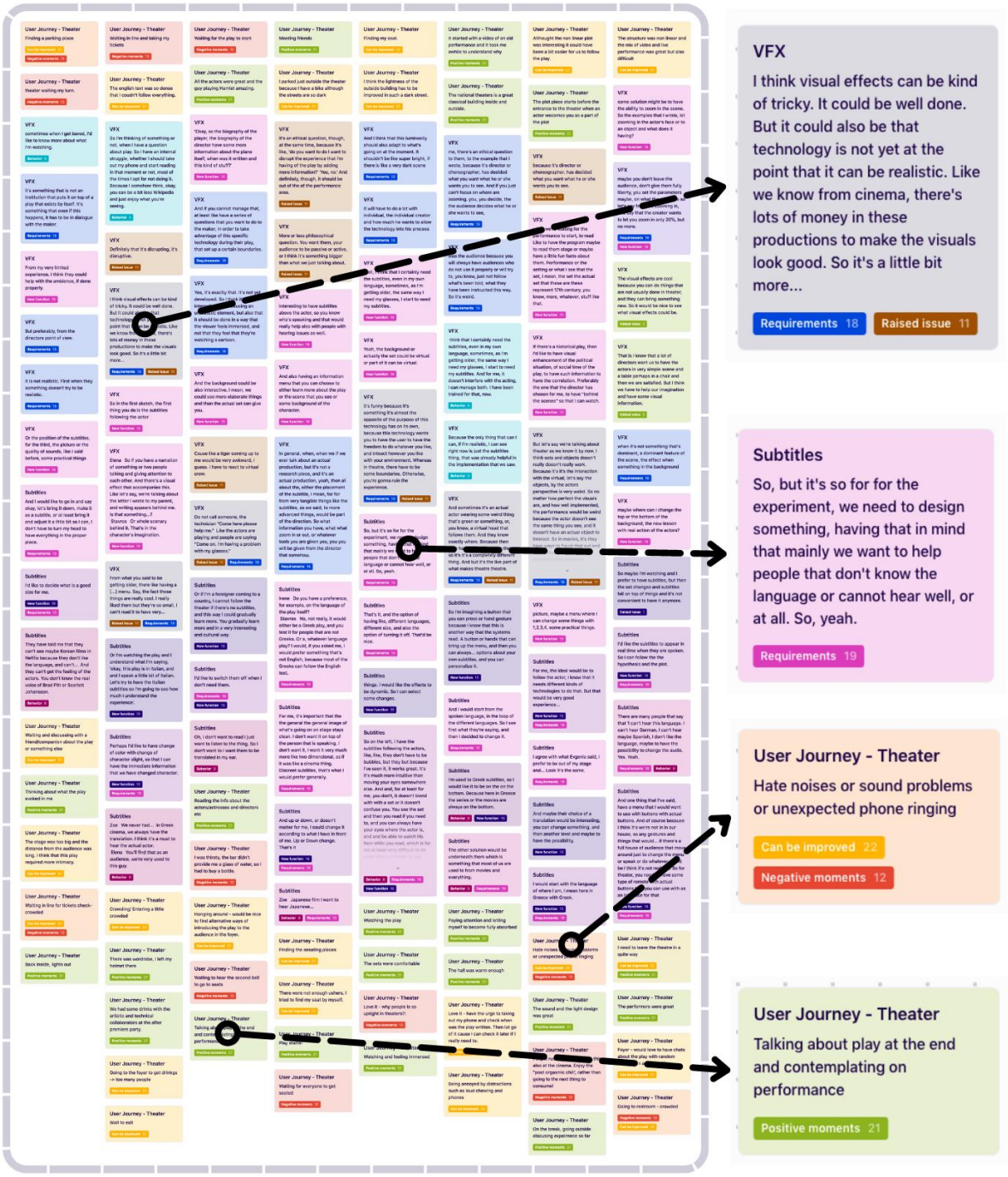


Figure 21. Insights Canvas for Augmented Theatre (AEF - Athens Festival) Use-case

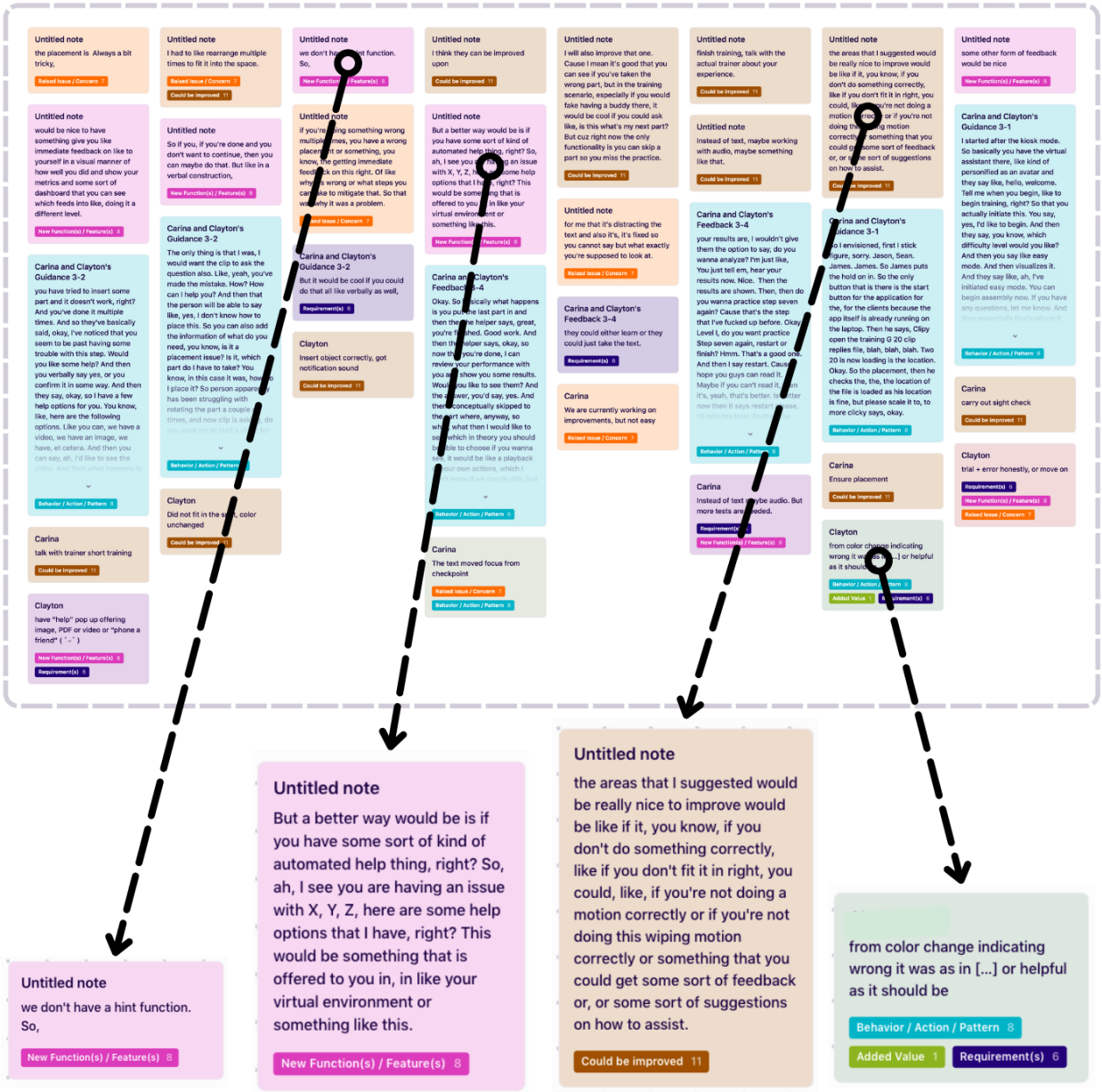


Figure 22. Insights Canvas for Training Assistant (HOLO) Use-case

Table 8 summarizes the number of insights for each category and theme that emerged from the analysis of the data. Each category represents the key considerations discussed by the participants in response to dedicated tasks, questions and group activities.

Table 8. Total number of insights for each of the category/sub-categories

#	Categories	VR Conference (VRDays)		Augmented Theatre (Athens Festival)		Training Assistant (HOLO)
1	Positive Moments	20		21		N/A
2	Negative Moments	25		12		N/A
3	Could be improved	25		22		11
Sub-themes		Virtual Assistant	Subtitles	VFX	Subtitles	Training (AR)
4	New function(s) / feature(s)	40	27	16	15	08
5	Behaviour / action / pattern	17	5	4	8	08
6	Requirement(s)	27	28	18	19	06
7	Added value	21	-	2	-	01
8	Raised issue / concern	10	15	11	1	07
9	Interface (suggestions)	08	29	-	-	-
10	Interaction	05	-	-	-	-

5.2.5 Results

By analysing and synthesising the data, we revealed important themes and patterns, as well as gained a better understanding of the attitudes and behaviours of the participants in augmented and VR applications of all the project use-cases.

The next paragraphs provide insights about the user-experience:

- **Intuitive and easy-to-use interface:** The AR/VR application should have an intuitive and user-friendly interface that allows attendees to navigate and interact with the environment and other attendees easily. It should be easy to access the platforms for virtual conference program, theatre experience or virtual training session, while supporting a quick learning curve for all three use-cases.

- **Comfortable and immersive experience:** VR and AR experiences can be overwhelming and cause motion sickness for some users. Therefore, the application(s) should be designed to create a comfortable and immersive experience, with well-balanced audio and visuals, and smooth movements. Also, timings of the content presented should suit the considerations of wearing headsets for long durations.

- **Interactive and engaging content:** The application should offer a range of interactive and engaging content to keep attendees/viewers/trainees interested and involved throughout the conference/theatre/training sessions.

- **Personalization and customization:** The virtual agent should be able to personalise its responses and recommendations based on attendee preferences and behaviour. This may include recommending sessions or events based on an attendee's past interests (if applicable), or offering personalised assistance based on an attendee's needs.

- **Clear and concise communication:** In all use-cases, (virtual conference room, theatre and training session), very clear and concise communication is crucial. The virtual environment should support high-quality voice and video communication, with minimal latency and interruptions. Also, ability to turn on/off main audio with translation audio and noises from surroundings should be included to make experience more user friendly.

- **Intuitive navigation and controls:** The UI should be intuitive and easy to use, with simple navigation controls. Users should be able to move around, adjust their position (if needed), and interact with other virtual objects and controls with ease.

- **Customizable avatars:** Attendees should be able to customize their avatars to represent their personalities and preferences. This can help foster a sense of community and create a more engaging experience. Many users reported lack of engagement in their experience due to fixed or stereotypical representations of avatars.

- **Realistic environment:** The conference room should be designed to create a realistic environment that immerses attendees in the virtual experience. With precise care on size of the modalities, including screen size, height and placement that gives naturalistic dimensions to users in virtual and augmented environments. Since some users reported problems with too high or low viewing angles of virtual objects and scenes was quite uncomfortable.

- **Multimodal communication**: The virtual agents should be able to communicate with attendees using multiple modes of communication, including voice, text, and visual cues. This will help ensure that attendees can interact with the agent in a way that is most comfortable and convenient for them. Many users suggested that virtual agents should be helpful but not intrusive.

- **Assistance and guidance:** The virtual agent should be able to provide attendees with intuitive assistance and guidance throughout the conference/theatre/training without being asked several times. This may include proactive answering to questions about the conference/theatre/training schedule, providing directions to specific locations or sessions, or helping users connect with other attendees/viewers/trainers (if required).

The next paragraphs provide insights about the expected interfaces for all use cases:

- **Platform Interface:** The AR/VR application(s) interface should display the most frequently used features e.g., conference program and schedule, controls for the theatrical play, or training demos etc to be easily visible so that users can easily view and interact with the systems.

- **Navigation and movement:** The interface should include intuitive navigation controls that allow attendees to move around the environment with ease, e.g., walking, fast run, flying etc.

- **Avatar customization:** The interface should allow attendees to customize their avatars, such as changing their appearance, selecting different outfits, or adding personal touches. Some users suggested various avatar representations such as a wrist/hand band (like smartwatch), sun or cloud character, blobby (cartoony character for virtual assistance) etc.

- **Non-intrusive Communication:** The interface should include tools for communication, such as voice chat, text chat, or virtual hand gestures, that allow attendees to interact with each other in real-time but without being forced or intrusive in the overall experience.

- **Virtual objects:** The interface should include virtual objects that attendees can interact with, such as whiteboards, presentation screens, or interactive models. These objects can be used for presentations, display screen, discussions, or interactive activities.

- **Personalization options:** The interface should allow attendees to personalize their experience, such as adjusting the audio and visual settings, selecting their preferred language or time-zone, or customizing their profile.

- **Help and support:** The interface should include tools for help and support, such as a quick demo, short tutorials, knowledge base, FAQs, etc.

The above recommendations are derived after analysis of the available data and suggestions by the participants of the workshops; however, their feasibility or conformance with the project's goals has not been validated. Therefore, their implementation is subjective to the feasibility in accordance with technical requirements. We report them here for completeness, and to provide guidance for third party developers in furthering the development of the VOXReality software.

6 Use Cases descriptions

In this section, we describe the scenario, technologies, equipment, external contexts, assessment protocol and target population for each of the use case of VR Conference, Augmented Theatre and Training Assistant respectively.

6.1 VR Conference

This scenario objective is to provide virtual conferencing environments with real-time multilingual translation and captioning, and to develop a virtual assistant providing users with relevant information to navigate virtual spaces, interact with other visitors, and receive relevant information.

6.1.1 Scenario

The ideal scenario should emulate the most recognizable features of a professional conference setting: Conference rooms, social spaces (cafes and lounges), business dedicated areas (trade show stands and business meeting areas). In addition, all information delivered to users within the virtual venue should be structured, prioritizing venue navigation, programme information and business-related interactions.

The length of the scenario should be 30 minutes max, ideally based on a predefined conference format, ex: TED, Pecha Kucha, etc. Minimum scenario times are defined by the full range of actions needed to validate the scenario successfully.

The virtual conferencing user case is a virtual door-to-door, assisted experience, including real-time multilingual translation and captioning assisting in conferences, virtual trade shows and social areas. Once users log into the virtual conference venue, a dedicated virtual assistant will engage with them. This virtual assistant is programmed to facilitate navigation, trained to answer questions and conditioned to deliver info feeds related to the conference programme-related activities.

All users of the VCE scenario will benefit from real-time translation and captioning service during conferences (1:X – one-to-many interactions), social interactions and business exchanges (1:1 one-to-one interactions). Once inside the virtual venue, users will be represented by virtual avatars. These avatars will be programmed to deliver visual feedback complementing available real-time multilingual translation and captioning.

Entering the venue

Once users log in to the virtual conference space/environment, they will face recognizable spatial features common to conference/event spaces: Entrances, lobby areas, trade shows, meeting rooms, social areas and ultimately, a conference/plenary room. From the user perspective, the virtual conference space will provide a door-to-door experience. Once at the virtual conference spaces, users can find their way around it thanks to a virtual assistant, who will answer questions, provide wayfinding guidance and inform visitors about relevant conference programme information.

Social interactions

Social interactions within virtual conference spaces are defined as 1:1 interaction. While an informal setting may produce unstructured conversation patterns, business interactions may follow a more structured communication and predefined vocabulary.

Figure 23. Social Area - Immersive Tech Week 2022 (©SnapBoys.nl)

In addition to real-time translation and captioning, virtual conference users will use avatars to provide a recognizable feature in the virtual space.

Trade shows

As part of the virtual conference space, a section dedicated to business interactions will be included: a Trade show where exhibitors and their spatial representations (stand/booth) will be allowed to engage with users/visitors via direct marketing on a visual level, or messages mediated by the personal assistant and the user's preferences.

Figure 24. Tradeshow - Immersive Tech Week 2022 (© JaynoBrk)

Into the conference rooms

Once users access the conference room area, they may choose between two in-room view options: Full screen and conference room setting. Both in-room views include on-demand virtual assistance and real-time multilingual translation and captioning.

Since this is a 1:X (one to many) kinds of interaction, users will be at the receiving end most of the time. However, exceptions may be allowed when Q&A opportunities are available, allowing the audience to engage with speakers/guests. Third parties/agents will enable these interactions.

At the end of every conference session, users will be presented with all available options, including vacating the conference room for the conference venue, re-visit the conference and/or exiting the event completely.

Figure 25. Main conference - Immersive Tech Week 2022 (© JaynoBrk)

A step-by-step walkthrough of the use case includes:

1) The user logs into the VCE with the assistance of VR headsets.
2) Once connected to VCE, users will be assigned a virtual assistant. This virtual assistant will help users to navigate the space, answer programme-related questions, and deliver relevant information feeds from social and business activities.
3) Each virtual assistant will present users with information on available spaces and interactions, such as social and business-related options. Among the options available in the VCE social, exhibitor stands and business meeting areas.
4) Users may choose between Full screen and Conference room settings once they have entered any conference rooms. Interactions with speakers and/or presenters may be available.
5) Once conferences finish, users can opt to vacate the conference room for the conference venue, where further interactions are possible, re-visit the conference (pre-recorded) and/or exit the event completely.

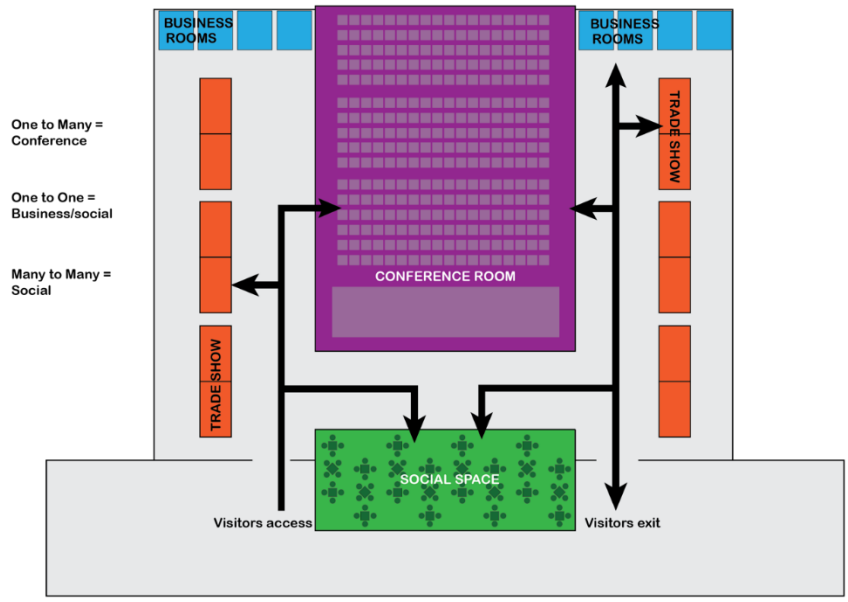


Figure 26. Virtual Conference space – Conceptual floorplan (©VRDays)

6.1.2 Technologies

The Virtual Conference case will utilize the ASR component developed by VOXReality to generate textual counterparts of the speech of users interacting with each other, users interacting with a digital assistant and speakers in the keynote speech.

The textual information acquired from the ASR will be utilized in the NMT component for automatic translations in the interaction cases (keynote speech and 1-1 interaction). NMT component will also make use of contextual information such as terminology used in the conference and abstract of the keynote speech.

Furthermore, the contextual information will also be included from the outputs of the visual language component that provides descriptions of the scene the user sees and the description of the events in a room that the user is in. Apart from the scene description, the VL component will be also capable of answering simple questions regarding various properties or spatial relations between the visible entities via dialog, using the DA. The textual information acquired from ASR will also be utilized in the DA component where a dialogue system helps the users to navigate through the virtual venue. DA component will facilitate a dialogue between the user and the corresponding navigation instructions and descriptions generated by VL component using English as a mediation language, hence, if the user speaks in a different language than English, the NMT model will provide the English translation for DA to generate its responses. The user will be able to ask questions and keep a continues dialogue going with the agent while navigating through the venue. The response generated from DA will be translated to the language of the user from English and off-the-shelf text to speech tools are envisioned to be utilized to generate an audio response in the desired language.

For navigation purposes the DA will make use of predefined map of venue.

6.1.3 Equipment

This user case supports its functioning on VR technology (VR Headsets), allowing users interactions via an in-headset microphone. In addition, VR controllers will provide navigation, menu selections and haptic interactions. The VR application will be accessible via a web-based application using a VR-ready computer (windows-based PC), enabling a live conferencing component.

6.1.4 External Context

External contexts considered for the VR Conferences use case are conference presentation slides, exhibitors' promotional content, TED data sets, conference presentation recordings, social interactions, technical word dictionaries, industry-related papers, industry-related press releases, among others that may be considered relevant.

6.1.5 Assessment Protocol

A single session will invite less than 20 conference goers with different backgrounds. The questionnaires will be used to assess the quality of the experience, followed by a semi-structured interview to get qualitative insights. Dutch and English (Dutch translated to English) will be used.

6.1.6 Target Population

The target population includes conference partners and visitors with different backgrounds.

6.2 Augmented Theatre

This use case aims to introduce AR technologies to theatre audiences. A special performance will be produced for the purpose of VOXReality. A small audience will be invited to experience AR applications in theatre. Specifically, this use case will create an AR experience for the audience via providing AI-generated and by incorporating visual effects in the performance.

6.2.1 Scenario

Up to 50 theatre goers, split in small groups, will arrive at the venue in assigned timeslots. Each group, upon arrival, will be informed about the Project, the technologies and the performance by AEF personnel and informational materials. They will sign the relevant consent forms before being led to the theatre where they will be given a short presentation of the use of the AR equipment prior to watching the performance. Technical staff will assist the audience in wearing and familiarising with the equipment.

A scene of the play “Hippolytus” by Euripides will be performed. The duration of the performance will not exceed 15 minutes. Afterwards the audience will be asked to fill out a questionnaire regarding the performance and the use of AR technologies. The duration of each groups experience, including pre- and post-performance activities, should be no more than 30 minutes.

It is yet to decide if all members of the audience will experience both technologies used, the automatic translation and the visual effects, or if there will be members of the audience without AR equipment. The creators of the performance may also be asked to evaluate the experience, through semi structured interviews [5].

6.2.2 Technologies

The use case will utilize the ASR component developed by the VOXReality consortium to generate textual counterparts of the speech of actors on stage and if the play includes a narrator, the speech of the narrator. The textual information acquired from the ASR will be utilized in the NMT component for automatic translations. NMT component will also make use of contextual information such as director notes, summary of the play, scenography and description of the characters and the theme of the play. Furthermore, the contextual information will also be included from the outputs of the visual language component. The VL component will use as input video from the viewpoint of the audience and periodically output the description of the scene in textual form.

6.2.3 Equipment

The AR equipment to be used as well as other specialised equipment is yet to be determined. The purchase costs of the equipment will be analysed against budget allowance and project needs, without compromising the expected results.

6.2.4 External Context

The external context that will be provided is the following:

- The summary of the ancient Greek play “Hippolytus” by Euripides;
- Mythological context regarding the characters of the tragedy;
- The summary of the scene, and of the events preceding and following;
- Description of the play’s themes, historical context, character relationships;
- Information regarding scenography and kinesiology: information on the set, costumes and movement of actors on the stage;
- Director notes on the scene;
- Music used, if any.

6.2.5 Assessment Protocol

The use case will take place at the Athens Epidaurus Festival’s Peireos 260 venue located at Peireos Str. Athens, Greece. The total participants, 50 theatre goers, will be split into smaller groups for each performance. The size of each session’s group will be determined based on the AR equipment that is to be used. The audience will be selected so as to represent demographic diversity, considering age, gender, language, technological knowledge, opinion/acceptance and experience. Specific demographic targets will be set later on during the project implementation.

Figure 27. A potential stage at the Athens Epidaurus Festival's, (Left) Distance between the stage and the audience, (Right) Chairs for the audience

The language of the performance will be either Greek or English and the translation will be available in all the languages under the scope of the VOXReality project. Questionnaires will be developed to assess the audience experience. Similarly, semi structure interviews with selected personnel from the performance's creator team will be constructed.

6.2.6 Target Population

Theatre goers, native speakers of any of the VOXReality languages or English. The selected group will fulfil diversity criteria regarding age, gender, knowledge of and exposure to AR technologies and opinion on AR and AI technology.

6.3 Training Assistant

This use case aims to provide a guided AR industrial training scenario. In this use case, a trainee will work on an assembly task through interaction with 3D virtual content superimposed on the real world. During this AR assembly task, users will be guided and supported by a virtual assistant. Users will verbally engage with the virtual assistant who will be able to provide support when prompted and/or offer support when elicited based on trainee performance in the task. The AR assembly task here will involve the visualization of a holographic computer-aided-design (CAD) file. This holographic CAD file object will be interactable and composed of multiple parts which need to be assembled together by the trainee. Physical manipulation (e.g., picking up, moving, and inserting) of specific object components will be used to assemble these constituent parts into the object's frame. Different levels of training will be offered, ranging from easy to difficult, and assembly levels can be performed in repetition. The virtual assistant will provide a unique language-centric interface with the AR training environment and aid end-user trainees in industrial assembly instruction.

6.3.1 Scenario

This training use case scenario will focus on the experience of a single user participating in the AR training environment. In this scenario, a user will be provided with a HoloLens 2 (and/or an android) AR smart glass device. The user will engage with the training experience as detailed below in the user journey walkthrough, but in general are predicted to consist of running through the individual difficulty levels at least one time each. The use case is currently envisioned to take approximately around 3 hours (for one person), which includes time for set-up, introduction to the hardware device, training application scenarios, clean-up, and feedback assessment periods.

The tentative predicted user journey walkthrough will in general occur with the following assumptions: 1) Only one user will be in the AR training session at a time. 2) Trainees will be inexperienced in using the HoloLens 2, and 3) Trainees will be relatively inexperienced in the industrial assembly task (i.e., unfamiliar with the CAD file to be assembled).

A user will first be introduced to the HoloLens 2 and provided an explanation of what will be happening during their AR training experience. They will then be provided with relevant consent forms to sign, and after signing they will be fitted with the HoloLens 2 and the training scenario will begin. After launching the application, either by a VOXReality researcher or the participant verbally with the virtual assistant, the user will be greeted by the virtual assistant and prompted to confirm they would like to begin training. After confirming, the user will then be asked by the virtual assistant which difficulty level they would like to be trained on. Following a verbal decision by the user, the training will begin. Users will then be able to see a holographic 3D industrial CAD file object which is to be assembled. This object will not be superimposed over a physical twin object in real life, but rather existing over the local surrounding environment. This object is composed of a certain number of parts which are loaded initially external to a core frame. The constituent parts are then to be picked up and inserted in the correct spatial position into the object's frame. Depending on the difficulty level, the parts may be visually ordered in a sequential fashion or have a corresponding color-coded deposit location on the object frame. A user would then begin moving the holographic parts in space and attempting to position them into the frame. Throughout this process, the virtual assistant will be readily available to respond to help inquiries and offer support. Support is predicted to come in the form of e.g., documentation such as images or PDFs. As users pick and move the individual assembly parts into the frame, relevant UI elements are predicted to provide information which may aid in the assembly process or provide feedback on how they are performing, which could scale depending on the difficulty level selected. The virtual assistant will also monitor a users' process and intervene when the user appears to need help. For example, if a user is attempting to insert a part in a wrong location or takes a long period of time to move forward with the assembly, the virtual assistant could engage verbally with the user and offer support.

This process is intended to continue iteratively until all constituent parts of the CAD object are assembled. The user will have learned how to assemble this object and have an understanding of the assembly process. The virtual assistant would then be able to offer the user the opportunity to restart the training again along with the option to choose a different, likely more challenging, difficulty level. The user would then be expected to try the different difficulty levels in sequence. This serial assembly practice and continued help/assistance with the virtual assistant will, in the end, have provided a unique training environment for the user. Following the completion of each difficulty level, it would be beneficial if the user could be provided with metrics associated with their performance (e.g., time spent doing the entire task, per each step, etc.) in some format. After the final difficulty level is reached, the user will end the training session by instruction to the virtual assistant. The user would then remove the HoloLens 2 and the session would be closed. Following, users will be provided with paper assessment forms to get feedback on their experience as well as discuss with VOXReality experimenters on site.

6.3.2 Technologies

The specific technologies intended to be utilized in this use case consist of hardware devices, which include the HoloLens 2, and likely a PC laptop where the AR assembly application will run. This application will be based in Unity and will incorporate Holo-Light's Interactive Streaming for Augmented Reality (ISAR) SDK. ISAR-enabled applications provide a client-server environment

which offloads the computationally heavy rendering process from the AR device to a dedicated server environment (e.g., laptop) with a high-quality graphics processing unit and streams the application. ISAR allows the streaming of this entire application to the HoloLens 2, which itself has an ISAR-client application installed, which permits high resolution AR visualizations with low latency. This application will further make use of Holo-Light's Augmented Reality Engineering Space (AR3S), which will provide the basis for visualization and interaction of 3D CAD files for the assembly training tasks. The application developed for this use case, which combines both the ISAR streaming functionality and the AR training and assembly features of AR 3S will be bolstered by the virtual assistant which will be described below.

The use case will utilize the ASR component developed by the consortium to generate textual counterparts of the speech of users interacting with a digital assistant. The textual information acquired from ASR will be utilized in the DA component where a dialogue system helps the users to assemble a machine. DA component will facilitate a dialogue between the user and the corresponding assembly instructions and descriptions generated by VL component using English as a mediation language. The response generated from DA will be in English and off-the-shelf text to speech tools are envisioned to be utilized to generate an audio response.

Moreover, the agent will be composed of a NLU model that will understand what the user is talking about as well as a NLG model, that generates meaningful responses. The agent will make use of manual as an unstructured textual document that guides the agent in interacting with users. The agent will be able to display pdf, corresponding videos, or specific files, if it is asked by the user.

6.3.3 Equipment

As mentioned above, this use case will make use of the HoloLens 2 as the AR device of choice that the end user (trainee) will utilize to interact in the AR training environment. In addition to allowing holographic visualizations of the 3D CAD data, the HoloLens 2 possesses a microphone and speaker, enabling a full audio experience and thus leverages the ASR component, NLU, and NLG models in order to have a dynamic language experience with the virtual assistant during the training sessions. As mentioned above, a PC laptop will likely be used to host the server-side AR application which will be streamed directly to the HoloLens 2. No additional equipment is envisioned to be necessary for this use case as all interactions within the training environment occur with virtual holographic content and audio information.

6.3.4 External Context

Beyond the holographic CAD file objects that users will interact with in this use case, external context will include support materials which the virtual assistant will provide/offer to the user when engaging with the training session. Such content will likely consist of different images of the CAD object and its constituent parts that the user could view. Additionally, instruction manuals may also be offered in the form of PDFs which the user could view and read at their discretion. While not guaranteed to be present, a further external factor which would be beneficial to have would be videos which can show an individual part being assembled in AR space. Such factors are envisioned to be incorporated into the AR application stream which would allow the user to visualize them during the training session. Given that the entirety of the session will be occurring in AR, the external environment having a defined location is not essential for this use case. However, an ideal location would be an industrial shop floor or similar location which would approximate a canonical environment where a typical user would engage in such training.

6.3.5 Assessment Protocol

Following the training session, users would be asked to participate in a feedback session designed to collect information on their experiences with assembly training with emphasis on interaction with the virtual assistant. Participants will be asked to complete a questionnaire/survey and/or provide direct verbal feedback to VOXReality experimenters about the quality of their experience either verbally or by questionnaire. The results from user profiling will provide insights into what went well, what could be better, and overall impressions of the use case and its technology provided. In addition, assessment protocols are envisioned to be application metrics which are profiled during the training sessions such as e.g., time spent doing a virtual task, number of incorrect steps taken, number of times the virtual assistant had to intercede and offer help due to the latter, etc. This data can be used to scientifically approximate the efficacy of the training experience over sessions with increasing difficulty. This data will be collected from the application itself.

This assessment protocol is intended to be acquired from 3-5 users; more would be ideal if possible. Despite such a presumptive low sample size however, a robust coverage of all possible end user demographics is a goal of this use case. In particular, equal numbers of all genders will be aimed for. Language of operation will likely be in English. Users are envisioned to be readily open to technology, with little experience in XR technologies and some experience in industrial tasks.

6.3.6 Target Population

Target populations in this use case will be users who have a desire to work with virtual assistants for task performance during industrial assembly training. Ideal participants would be those who have no prior knowledge of the task to be performed in this use case, and those who have not had much experience with AR, so they can be instructed without prior influence biasing the training scenario. Users will likely be factory workers with some assembly experience and knowledge of assembly training in general.

7 Requirements

This section lists all the requirements gathered in the first six months of VOXReality implementation, based on the weekly calls and in-situ visits. They are divided into user (Section 7.1) and technical (Section 7.2) requirements.

7.1 User Requirements

This section provides the assessed requirements of each use case – VR Conference, Augmented Theatre and Training Assistant. The list is the result of the cumulated user requirement gathering process, from the initial weekly calls (Section 3), in-situ visits (Section 5), and the final use case description (Section 6).

The initial user requirement (Section 4) was derived from the weekly calls where all the consortium partners worked on aligning the scenarios with a mutual understanding of its technologies and possibilities. Based on the initial user requirements, the user-centred partner designed activity materials and conducted interactive focus group workshops with each use case partner through In-situ visits (Section 5). The focus of the workshops was to elicit the latent needs of each use case owners and the actual users so that can help develop the initial user requirements to in details.

The workshops benefited the user-centred partner and use case owners bidirectionally. On the one hand, the analysis of the activity results and discussions during the workshop led to a list of original insights addition to the initial understanding. On the other, the workshop participation of each use case owners helped them to clarify the project setting and define the scenarios into a deeper level. Consequently, the user requirements (Section 7.1) resulted to have modifications from the initial list (Section 4), by combining the outcomes of In-situ visits (Section 5) and use case description (Section 6).

At this stage it is too early to conclude final requirements from the project, thus these requirements in section 7.1 are not binding to development, instead it provided an overarching perspective collected from users. The implementation of features is subjective to technical feasibility mutually agreed among requirements and technical partners. The following sub-sections present updated user requirement lists with tables and summaries per use case scenario.

7.1.1 VR Conference

Following section describes the overview, scenario, assessment, virtual-agent(s), navigation, subtitles and translation, interface and user-interaction themes derived after carefully analysing the data by thematic analysis, whereas the individual requirements within each theme is enlisted in Table 9.

Overview

- The VR experience should primarily focus on user navigation and language translation features along with support for all the standard functionality required for a VR conference platform. A virtual agent should also be developed to provide users with relevant information to navigate virtual spaces, interact with other visitors, and exchange information. The operational output language will be English but can be translated into other languages.

Scenario

- The virtual conference venue should simulate a professional conference setting, including an entrance, lobby, trade shows, meeting rooms, social areas, and a conference room. The trade show should consist of spatial booths for business interaction and engagement. Participants

will be represented as virtual avatars, complete with predefined gestures and features such as hand-shakes, high-fives, nods, and thumbs up gestures. Presentations should follow a specific conference format, such as TED talks or Pecha Kucha, with a maximum duration of 30 minutes, in order to keep time limit feasible for users, and to prevent discomfort caused by prolonged headset use.

Assessment

- The VR conference targets a diverse group of conference partners and visitors, with a recommended limit of 20 participants per session (room-space). The experience will be assessed using a questionnaire and a semi-structured interview to ensure high quality.

Virtual agents

- A virtual agent will be assigned to each user for the duration of the conference, providing help and navigation assistance. The agent should offer getting-started help, a welcome greeting, FAQs, and personalized info feeds related to the conference program. The agent should also suggest call-to-action depending on the current location and activity status of the user, and present all available options at the end of each session. Communication with the agent should be enabled by voice typing and/or text, and users should have the option to skip the virtual assistant help. The agent should be likeable, friendly, pleasant, customizable, non-intrusive, and realistic. Users should be able to access the agent via a smartwatch/wristband, and the agent should auto-save relevant information for export in a downloadable or email-able PDF.

Navigation

- The navigation experience should be clear with consistent navigation cues and visual landmarks that enabled attendees to easily find their way around the virtual venue. To enhance the user experience, several features related to navigation should be available during the virtual conference. Users should have the option to take a quick tour of the venue, which should last between 30-120 seconds and can be replayed or skipped. A virtual map should be accessible to help users navigate the conference venue, allowing them to identify their current location. To facilitate quick and easy navigation between different locations, users should be provided with an option of a virtual taxi or cab. In case the user takes the wrong direction, they should receive a notification. Arrows should guide users towards their target location. Additionally, a flyover view of the venue should be available with zoom in/out options.

Subtitles and Translation

- The users should be able to communicate and translate both text and voice. They should also have the option to mute the entire room or individual speakers and control the volume. Users should be able to turn on/off the program, subtitles, auto-translation, and voice using interactive buttons. For multi-speakers, the subtitles should highlight the active speaker among the group, with priority given to the speaker that the user is currently looking at.
- The subtitles should be displayed relative to the context using various suggestions like speech bubbles or text strips, with the default placement being at the bottom of the screen. To avoid eye-strain, subtitles should be placed in level to viewing angle and also ensure to not mask/block important parts of the display or screen. For HMDs, subtitles should be located over or in the middle of the screen. The subtitles and translation experience should promote inclusivity and accessibility, enabling attendees from around the world to fully participate in the conference and engage with conference content in a meaningful way.

Interface

- There should be a customizable dashboard that is always visible, allowing easy access to frequently used features. A help button should also be present on the corner of the screen for quick assistance. Additionally, two in-room view options should be available: full screen and conference room. To facilitate audience engagement, questions and feedback should appear near the presenter's screen, ensuring that they are noticed and answered promptly. When an attendee asks a question, it should be sent visually or textually to the speaker to avoid it going unanswered.

User Interaction

- The system should provide options for users to interact with speakers and guests during Q&A sessions without interrupting the presentation. Users should be able to ask questions to an agent during navigation, using voice commands such as "Please describe the scene for me," "What is in this room?" or "Is there an empty chair in the room?" Additionally, users should be able to use hand gestures or interactive buttons on the screen for navigation.

Table 9. Final Requirements for VR Conference use case

Type	Sub category	Requirements	*Priority*	*Notes*	Phase gathered
General		The experience is in VR.	High		Conference use case scenario document
General	Objective	Provide user navigation and language translation at the virtual conference.	High		Use Case Description
General	Objective	Develop a virtual assistant providing users with relevant information to navigate virtual spaces, interact with other visitors, and exchange relevant information.	High		Use Case Description
General		The operational output language is English, and it can be translated into other languages.	High		Technical-Use case meeting 6
Scenario	Space	The conference venue should imitate the environment of a professional conference setting.	High	*e.g., entrance, lobby, trade shows, meeting rooms, social areas, and a conference room, etc. *The trade show includes spatial booths for business interaction and engagement.	Use Case Description
Scenario	User Representation	The users will be represented as virtual avatars.	High		Use Case Description
Scenario	Social Interaction	The virtual avatars of the participants will offer predefined set of gestures and features.	Low	*e.g., hand-shake, high-five, nodding, thumbs up gesture etc.	Use Case Description

Scenario	Contents	The presentation should be in a predefined conference format.	High	*e.g., TED, Pecha Kucha etc.	Use Case Description
Scenario	Contents	The suggested duration for a single talk is 30 minutes at maximum.	High	*wearing headsets for too long is uncomfortable and might cause headaches / visual discomfort.	Use Case Description
Assessment	Target User	The target user includes conference partners and visitors with diverse backgrounds.	High		Use Case Description
Assessment	Target User	The suggested number of participants per suggestion is 20 people.	Low		Use Case Description
Assessment		The quality of experience will be assessed by a questionnaire followed by a semi-structured interview.	High		Use Case Description
Virtual Agent	Function	A dedicated virtual agent will be assigned to each user and will stay with him/her during the complete duration of the conference.	High	*Virtual agent should offer getting-started help without the user asking for it.	Conference use case scenario document
Virtual Agent	Function	The virtual agent should provide welcome greeting.	High		In-situ visit - VRDays
Virtual Agent	Function	The virtual agent should provide help and FAQs.	High		In-situ visit - VRDays
Virtual Agent	Function	The virtual agent will help user to navigate the space, answer programme-related questions, and deliver relevant information.	High		Use Case Description
Virtual Agent	Function	Virtual agents inform the users about the context/content of the venue and potential to-do items in advance or relative to the current activity.	Low	*The virtual agent should offer call-to-action depending on the current location and activity-status of the participant.	Technical-Use case meeting 2
Virtual Agent		The users should have a skip option for the virtual assistant help.	High	*In case they have already done the training or they are aware of similar experiences	In-situ visit - VRDays
Virtual Agent		The communication with virtual agent should be enabled by voice typing and/or text.	High	*Text typing is very time-consuming in VR settings.	In-situ visit - VRDays
Virtual Agent	Function	Virtual agents deliver personalized info feeds related to the conference programme-related activities.	Low	*The virtual assistant should keep track of the activities of the user and suggest personalized suggestions to each participant.	Conference use case scenario document, In-situ visit - VRDays

Virtual Agent		The virtual assistant can be accessible via a smart-watch / wrist-band.	Low	*to access available options at all times during the conference	In-situ visit - VRDays
Virtual Agent	Function	The virtual-assistant should auto-save relevant information and provide them in an exportable file.	Medium	*e.g., the day log, rooms visited, events attended, people met with their contact information, etc in form of a pdf file that could be downloaded or emailed to the user.	In-situ visit - VRDays
Virtual Agent	Function	The virtual agent should present users with all available options at the end of every session.	High	*The options may include vacating the conference room for the conference venue, re-visit the conference and/or exiting the event completely.	Conference use case scenario document
Virtual Agent	Behavior	The virtual agent should interact with the users on-demand, and should not be intrusive.	High		In-situ visit - VRDays
Virtual Agent	Behavior	Virtual agent should be likeable, friendly, pleasant, customizable, realistic and complete.	Low	*e.g., for behavior, the interaction of agent with user should be intrinsic. For appearance, the agent should be complete and not be missing graphical details like missing limbs, parts or other natural features.	In-situ visit - VRDays
Virtual Agent	Appearance	Virtual agent could look like a cartoony avatar.	Medium	e.g., cloud, taxi, humanic personification, a robot, an egg (blobby), magic wand etc.	In-situ visit - VRDays
Navigation		Quick tour option should be available.	High	*Suggested duration is 30-120 seconds. *It should be skipped and/or replayed.	In-situ visit - VRDays
Navigation	Interface	Virtual map should be available to navigate the conference venue.	Medium	*Users should be able to identify their current location.	In-situ visit - VRDays
Navigation		Users should be assisted for quick and easy navigation between different places in the venue.	Low	*e.g., through a virtual taxi/cab option to go to the desired location in the conference. * Users should be notified in case s/he goes in wrong directions.	In-situ visit - VRDays
Navigation	Interface	Navigation to the target location should be guided by arrows.	High		In-situ visit - VRDays
Navigation	Interface	Flyover view of the venue should be available with zoom in/out options.	High		In-situ visit - VRDays

Subtitles	Function	Translation should be available in both text and voice formats.	Medium		In-situ visit - VRDays
Subtitles	Function	Option to mute the complete room or individual speaker(s) should be available.	High		In-situ visit - VRDays
Subtitles	Function	Users should be able to control the volume +/-	High		In-situ visit - VRDays
Subtitles	Function	Users should be able to turn on/off the program, subtitles, auto-translation, and voice using interactive buttons.	Medium		In-situ visit - VRDays
Subtitles	Interface	For multi-speakers, the subtitles should highlight and different the active speaker among group of speakers.	Medium	*e.g., semi-transparent speech bubble; highlighting the active speaker *When the users look at a particular speaker among many, that speaker's subtitles should have priority of view/voice over other presenters/speakers.	In-situ visit - VRDays
Subtitles	Interface	The subtitles should be displayed relative to the context.	High	*Subtitles display suggestions: - strip on the top, - moving text strip on top, - speaker-attached speech bubble, - strip of text on bottom, - text box on side of the speaker.	In-situ visit - VRDays
Subtitles	Interface	The default placement of subtitles should be on the bottom of the screen.	High	*To match the standard/classic subtitles style in videos and movies. *Also, to prevent eye-strain, the subtitles should not be too high or too low and relative to the screen.	In-situ visit - VRDays
Subtitles	Interface	The subtitles should be visible without masking other important parts in the display/screen.	High		In-situ visit - VRDays
Subtitles	Interface	For HMD, the subtitles should be over or in middle of the screen, not on the bottom of the screen.	High		In-situ visit - VRDays
Interface	Dashboard	There should be a customizable dashboard visible at all times.	Medium	* The dashboard should allow participants to access most frequently used features.	In-situ visit - VRDays
Interface	Help	A help button should be visible on the corner of the screen.	High		In-situ visit - VRDays

Interface		In the conference room, users can choose between two in-room view options i.e., full screen and conference room.	Medium		Conference use case scenario document
Interface	Feedback	Questions/feedback from the audience should appear textually/visually near the presenter's screen.	Medium	*When a user asks a question, his/her question should be sent visually / textually to the speaker, to avoid being unanswered.	In-situ visit - VRDays
User Interaction		Users should be able to engage with speakers and guests during Q&A session(s).	High	*The interaction between the speaker/presenter and the attendee/user should be available without interrupting the presentation/talk.	Conference use case scenario document
User Interaction		Users can interact with and ask the agent questions during the navigation.	High	*e.g., “Please describe the scene for me.”, “What is in this room?”, “Is there an empty chair in the room?” etc.	Technical-Use case meeting 3
User Interaction		Users should have options to use hand-gestures or interactive-buttons on the screen.	Low		In-situ visit - VRDays
Extra		Chatbot should be available on demand.	Low		In-situ visit - VRDays
Extra	Payment	Users should be able to make digital and crypto payments in the conference items shop.	Low	*This is low-priority / extra suggestion	In-situ visit - VRDays

7.1.2 Augmented Theatre

This section describes the most prominent themes of overview, setup, scenario (play-design), assessment, interface, subtitles, VFX, and additionally recommended requirements that have been acquired after a thorough evaluation of the qualitative data through thematic analysis, whereas the Table 10 enlists details of each individual requirement within relative theme.

Overview

- The main objectives are to understand the users' needs and the organizers' wants for subtitles and VFX while watching a theatre play, and to brainstorm ideas for the UI and interaction method between the audience and AR glasses. It's important to ensure that the gestures and setup required to operate the AR glasses do not interfere with other audience members' theatre experience, and that the possibility of distracting the audience with additional information from the glasses is taken into consideration.

Setup

- The AR glass setup should be user-friendly for those who are not familiar with AR technology. Also, it should provide enough control to the audience while limiting their ability to alter the default settings.

Scenario: Play design

- The play of "Hippolytus", story of a young man who has sworn a sacred oath of chastity and devotion to the goddess Artemis, written by Euripides will be specially produce for a performance of not more than 15 minutes.

Assessment

- The scenario involves two users watching the play simultaneously using AR devices, subject to availability. The target user base includes diverse theatregoers whose native language is any of the VOXReality languages. It may also include some participants with minor hearing or vision impairments. Up to 50 participants will be involved in the complete experiment, including non-Greek speaking audiences in the Greek play. Automatic translation and visual effects technology will be available for the audience, and the user evaluation may include a combined demonstration of narration, dialogue, and VFX. Creators of the performance may also be asked to provide feedback through semi-structured interviews.

Interface

- The AR glass will provide a menu on the screen to adjust practical settings, allowing the audience to learn more about the play or scene and zoom in or out. The elements on the AR screen should adapt to the luminosity of the stage and theatre, and the interface should not require extensive head movements from the audience. Additionally, the UI of the AR screen should consider accessibility, such as font size for those with minor visual impairments.

Subtitles

- The AR Theatre should provide various options for subtitles, such as real-time translations in different languages, adjustable sizes, and the ability to turn them on and off. The audience should be able to move the subtitles up or down on the screen and avoid overlapping with the stage setup, allowing them to watch the actors' faces while reading the subtitles text. Additionally, the glasses should offer different audio language options, with the default being the original spoken language of the play. However, the audience should have the option to switch to a different language or to listen to both the original and local languages.

VFX

- During the early production stage, the implementation of VFX should be discussed with the script and plot, and the director should specify the users' capability to change AR Glass settings and subtitles. VFX can accompany or reflect the narration of the scene, and certain words or phrases may trigger the VFX. The VFX could be more appropriate for substituting supplementary features. However, the location of the visual effects should not exceed the stage's range, and they should not affect the actor's actions. The VFX should help the audience immerse in the performance, and low-quality VFX should be avoided as they may decrease the user experience. Additionally, the style of the visual effects should be artistically relevant to the opera and not too cartoonish.

Additional

- The stage background can be interactive, allowing the audience to learn about the play or read fun facts while waiting for the performance. The audience can also learn more about the player's or director's biographies, as well as other information about the play. Additionally, the audience can watch behind-the-scenes footage or learn about the social/historical background of the play before and after watching it.

Table 10. Final Requirements for Augmented Theatre Use Case

Type	Sub category	Requirements	*Priority*	*notes*	Phase gathered
Objective		The objective is to introduce AR technologies to theatre audiences.	High		Use Case Description
Objective		Language translation and VFX experience with AR glasses at the theatre.	High		
General		The experience is in AR.	High		
General		The gestures and setup to operate AR glass should not intervene other audience's theater experience.	High	*The AR glass operation should involve physical buttons.	In-situ visit - AEF
General		The possibility of the audience getting distracted by the additional information from AR glasses has to be considered.	Medium		In-situ visit - AEF
Setup		The experiment takes place at the avant-garde theatre.	High	*modern, indoor theatre	Technical-Use case meeting 6
Setup		The AR glass setup should be easy and clear to audiences who are not familiar with the AR technology.	High		In-situ visit - AEF
Setup		AR glass should give the audience enough controls but also limit their ability to change default/preset settings.	Medium		In-situ visit - AEF
Scenario		A scene from the play "Hippolytus" by Euripides will be specially produced and played.	High		Use Case Description
Scenario		The length of the performance will not exceed 15 minutes.	Medium		Use Case Description
Assessment	Target User	Two users can watch the play at the same time with two AR devices in a single time.	High	*based on the availability of AR devices	Technical-Use case meeting 4
Assessment	Target User	The target user is a theatergoer with diverse background whose native language is any of the VOXReality languages.	High	*i.e., age, gender, language, technology knowledge, opinion/acceptance and experience, exposure to AR technology.	Use Case Description

				*People with disabilities will not be considered; we can put emphasis on the potential of VOX technologies in providing access to people with disabilities (maybe auditory disabilities)	
Assessment	Target user	The target user includes who don't know the language or cannot hear well.	Low		In-situ visit - AEF
Assessment	Target User	The target user base includes people with vision correction glasses.	Low		Technical-Use case meeting 6
Assessment	Target User	In Greek play, non-Greek audiences should participate in the experiment.	Medium		In-situ visit - AEF
Assessment	Target User	Up to 50 people will participate in the complete experiment.	Medium		Theatre use case scenario document
Assessment		The translation of the play will be available in all VOXReality languages.	High		Use Case Description
Assessment		The audiences may experience either one or both of the automatic translation and the VFX technology.	High		Use Case Description
Assessment		The user evaluation may include combined demonstration.	Medium	*For example, 5-minute narration + 5-minute dialogue + 5-minute VFX + 5-minute dialogue/narration with VFX	Technical-Use case meeting 2
Assessment		The creators of the performance may be asked to evaluate the experience through semi-structured interviews.	High		Use Case Description
Interface	Function	The AR glass could have menu on the screen to change practical settings.	High		In-situ visit - AEF
Interface	Function	Audience can learn about the play or the scene.	Low		In-situ visit - AEF
Interface	Function	Audience can zoom in or out the scenes.	Low		In-situ visit - AEF
Interface		The elements in the AR glass screen should adapt to the luminosity of the stage and theater.	Low		In-situ visit - AEF

Interface		The interface should not include extensive head movements.	Low	*The audience doesn't have to turn his/her head to see everything in the proper place.	In-situ visit - AEF
Interface		The UI of the AR screen should consider audience accessibility.	Low	e.g., small font size	In-situ visit - AEF
Subtitles	Default	The default subtitle starts with the local language.	High		In-situ visit - AEF
Subtitles	Setting	The subtitles should be real-time, provide different languages, sizes, and turn on and off functions.	Medium		In-situ visit - AEF
Subtitles	Location	The audience should be able to change the location of the subtitles, either up or down.	Medium		In-situ visit - AEF
Subtitles	Location	The subtitles should not overlap with the stage setup.	High	*The audience should be able to watch the actors' face while reading the subtitles.	In-situ visit - AEF
Subtitles	Location	The subtitles can be placed above or below the screen and should follow each actor.	Medium	*Greek audience is used to subtitles underneath the screen.	In-situ visit - AEF
Subtitles	Default	The default audio starts with the original spoken language of the play.	High		In-situ visit - AEF
Subtitles	Audio	The audience should have an option to change the language audio.	Low	*Some people don't like to hear the specific language.	In-situ visit - AEF
Subtitles	Audio	The audience should have option to listen to the original and/or local language.	Low		In-situ visit - AEF
VFX		The VFX implementation should be discussed earlier with script and the plot.	Low	*early production stage	In-situ visit - AEF
VFX		The director specifies the capability of the users to change AR Glass settings and subtitles.	Low		In-situ visit - AEF
VFX	Function	VFX can reflect or accompanies the narration of the scene.	Low		In-situ visit - AEF
VFX	Function	Certain words or phrases will trigger VFX.	Low		Technical-Use case meeting 1
VFX	Function	A narrator could be the person who triggers the VFX.	Low		Technical-Use case meeting 2

VFX	Scope	The VFX may be more suitable for substituting the supplementary features.	Low		In-situ visit - AEF
VFX	Scope	Location of the visual effects should not exceed the range of the stage.	Low		Technical-Use case meeting 6
VFX	Scope	The implemented VFX should not affect the actor's actions on the stage.	Low	*The added VFX should not confuse the actors in how they perform.	In-situ visit - AEF
VFX		The VFX should help the audience be immersed in the performance.	Low	*Low quality VFX should be avoided since they are reported to decrease the user experience.	In-situ visit - AEF
VFX		Style of the visual effects should be artistically relevant to the opera.	Low	*not too cartoonish	Technical-Use case meeting 6
Extra	Function	Background of the stage can be interactive.	Low		In-situ visit - AEF
Extra	Function	Audience can read about the stage or fun facts while waiting for the performance.	Medium		In-situ visit - AEF
Extra	Function	Audience can learn more about the biography of the player, director, or more information about the play.	Medium		In-situ visit - AEF
Extra	Function	Audience can watch "behind the scene" or social/historical background of the play before and after watching the play.	Medium		In-situ visit - AEF

7.1.3 Training Assistant

In this section, we define the most prominent themes of assessment, user-interaction, interface, virtual-agent, and feedback requirements derived after the careful thematic analysis of the data regarding personal assistant and its related features, while the individual requirements within each theme is enlisted in Table 11.

Assessment

- **Target User**: The *target user* is a factory worker with some assembly experience and limited AR knowledge. They have no prior knowledge of the task and are relatively inexperienced in industrial assembly tasks.
- **Training:** The *training* will be conducted in English, with 3-5 users assessed using metrics like time spent, incorrect steps, and virtual assistant intercession. User feedback will be collected via questionnaire/survey or verbal feedback on assembly training and interaction with the virtual assistant. Each use case scenario takes around 3 hours.

User interaction

- The AR experience involves physical manipulation of object components, where the user can ask for instructions to assemble a machine/object. Users can choose between textual or visual instructions, with options to control the narration speed. User interaction is hands-on without controllers, and verbal engagement with the virtual agent is also possible.

Interface

- The interface should have a dashboard for commonly used functions that users can customize and personalize to their needs. It should also have appropriate colour schemes and visual cues for each action. Access to help and FAQs should be available on the screen at all times.

Virtual agents

- Virtual Agents provide verbal/textual explanations with corresponding visual content. It would use voice commands, visual cues, and other interactive features to help trainees identify the correct tools and parts, navigate the assembly process, and troubleshoot any issues or errors that may arise. They direct users to necessary objects and offer on-demand help, including hints and quick tips in text and/or audio format. The intention is to understand user needs and enhance efficiency of the machine assembly training in real-time, providing trainees with the support and guidance they need to master complex assembly procedures. The virtual assistant monitors user actions, intervenes after a certain number of incorrect steps, highlights errors, and intuitively guides users towards solutions. Support from the virtual assistants may include documentation such as images or PDFs.

Feedback

- Users should receive immediate and self-explanatory feedback on their actions, with instructions clearly stating any placement errors and pointing towards areas that need to be fixed/reassembled. Visual and audible notifications should indicate the correctness of tasks, and suggestions and feedback should appear automatically in case of incorrect assembly. The training and assembly should end with results and feedback, including time taken, number of correctly and incorrectly assembled parts, and other related information collected during the interactive session.

Table 11. Final Requirements for Training Assistant use case

Type	Sub category	Requirements	*Priority*	*Notes*	Phase gathered
Objective		The goal is to provide a guided AR industrial training scenario.	High		Use Case Description
Setup		The external environment having a defined location is not essential.	High		Use Case Description
Setup		An ideal location is an industrial shop or similar which approximate a canonical environment.	High	**where a typical user would engage in such training.*	Use Case Description
Setup		Three levels - easy, medium, and difficult - will be offered.	High	**In easy mode, 3D model always fits correctly to its location when in the correct place.* **In hard mode, user needs to put the parts into the*	Use Case Description

				correct location and alight the part's orientation.	
Assessment	Target User	The target user is a factory worker with some assembly experience and knowledge of assembly training.	Low		Use Case Description
Assessment	Target User	The target user has no prior knowledge of the task performed in the use case.	High		Use Case Description
Assessment	Target User	The target user has not much experience with AR.	High	**specifically, HoloLens 2, to prevent influences of biasing the training scenario*	Use Case Description
Assessment	Target User	The target user is relatively inexperienced in the industrial assembly task.	High		Use Case Description
Assessment	Target User	The target user will cover all possible end-user demographics, including equal numbers of genders.	Low		Use Case Description
Assessment		The operational language is English.	High		Use Case Description
Assessment		One user at a time will be in the AR training session.	High		Use Case Description
Assessment		At least 3-5 users will be assessed.	High		Use Case Description
Assessment		The assessment metrics include - time spent doing a virtual task, number of incorrect steps, number of times the virtual assistant interceded to help - etc.	High	**data can be used to scientifically approximate the efficacy of the training experience over sessions with increasing difficulty.*	Use Case Description
Assessment		A round of use case scenarios takes approximately 3 hours.	Low	**including set-up, introduction to the hardware device, training application scenario, clean-up, feedback assessment.*	Use Case Description
Assessment		The user's assembly training experience with emphasis on interaction with the virtual assistant will be asked.	High		Use Case Description
Assessment		The user feedback will be collected via questionnaire/survey and/or direct verbal feedback.	High		Use Case Description
Interaction		The AR experience involves the user's physical manipulation of object components.	Low	** e.g., picking up, moving, and inserting*	Use Case Description
Interaction		User can ask for instruction to assemble a machine/object.	High	**e.g., "Teach me how to to this", "Am I doing this right?"*	Technical-Use case meeting 6
Interaction		Users should be able to choose between textual instructions or visual demo for the training.	High		In-situ visit - HOLO

Interaction		Users should have options to turn on/off, speed-up or slow-down the narration.	Low		In-situ visit - HOLO
Interaction		User interaction is with hands, without controllers and may also verbally engage with the virtual agent.	High	**Available at HOLO & hardware SDKs*	Technical-Use case meeting 3
User Interface	Dashboard	There should be a dashboard for most common functions, which users can customize and personalize for their needs.	Medium		In-situ visit - HOLO
User Interface		The interface should provide color schemes and visual cues appropriate for the action.	High	**If the step is correct, it should be highlighted as green, if it needs partial adjustments, it should be yellow and it needs to be redone completely, it should be red.*	In-situ visit - HOLO
User Interface	Help	Access to help and FAQs should be available on screen at all times.	Low		In-situ visit - HOLO
Virtual Agent	Appearance	Virtual agents will be animated if in 3D avatar format.	Medium		Technical-Use case meeting 3
Virtual Agent	Format	Virtual agents can provide verbal/textual explanation and display corresponding visual contents.	High		Technical-Use case meeting 3
Virtual Agent		Virtual agents direct the user towards objects that are necessary for the application.	High	**e.g., certain stance for assembly of a machine part*	Technical-Use case meeting 3
Virtual Agent		There should be on-demand help during the interaction.	High	**e.g., How-to-do suggestion' in a display/dialogue box/ The virtual assistant will provide support when prompted or/and elicited based on trainee performance in the task.*	
Virtual Agent		The help should include hints and quick tips to guide users.	High		In-situ visit - HOLO
Virtual Agent	Format	The tutorial should be in both text and/or audio.	Medium		In-situ visit - HOLO
Virtual Agent		The virtual assistant should monitor the user action and offer instructions to help demonstrate the user on how to assemble a part correctly.	High	**the form could be in format of video tutorials, demo, visual, textual and audible instructions.*	In-situ visit - HOLO
Virtual Agent		The virtual assistant should intervene and offer automatic help after a certain number of incorrect steps by users.	Medium		In-situ visit - HOLO
Virtual Agent		The virtual agent should highlight the errors and intuitively guide users towards the solution.	Medium		In-situ visit - HOLO

Virtual Agent	Format	Support from a virtual assistant may come in the form of documentation, such as images or PDFs.	High		Use Case Description
Feedback		Trainee should be provided with immediate and self-explanatory feedback on actions.	Low	*If there is any error in the placement, the instructions should clearly state the issue and point towards the area to be fixed/reassembled*	In-situ visit - HOLO
Feedback		Users should have both visual and audible notifications regarding the correctness of the tasks at hand.	High		In-situ visit - HOLO
Feedback		The suggestions and feedback should appear automatically in case of incorrect assembly.	High		In-situ visit - HOLO
Feedback		The training and assembly should end with results and feedback.	High	*feedback includes the time taken, number of correctly assembled part, number of incorrectly assembled parts and other related information collected during the interactive session.*	In-situ visit - HOLO

7.2 Technical Requirements

In the three following subsections, one for each use case, we categorize the various technical requirements into Functional and Non-Functional ones, in compliance with the common standard ISO/IEC 25010. The prioritization of the requirements is developed under the MoSCoW model.

Besides the use case-specific requirements, there are universal ones, applicable to all scenarios. These are:

Title	Components' connectivity
Description	All system components must be connected using the most convenient interface and topology, regarding both performance and usability.
Typology	Non-Functional
Priority	SHOULD
Requirement ID	UNI-NFR-01

Title	Context-aware reasoning
Description	The system must consider live audio-visual environmental information.
Typology	Non-Functional
Priority	MUST
Requirement ID	UNI-NFR-02

Title	Real-time performance
Description	The system must operate and allow interactions at near-real time.
Typology	Non-Functional
Priority	MUST
Requirement ID	UNI-NFR-03

7.2.1 VR Conference

Use case-specific technical requirements, tailored for the VR Conferences scenario are:

Title	VR HMD
Description	The system must support the use of VR glasses since the experience will be fully virtual.
Typology	Non-Functional
Priority	MUST
Requirement ID	VC-NFR-01

Title	VR HMD audio capture
Description	The HMD must support audio capturing to input what the user says into the system.
Typology	Non-Functional
Priority	MUST
Requirement ID	VC-NFR-02

Title	VR HMD sound reproduction
Description	The HMD should have built-in speakers to playback sounds to the user.
Typology	Non-Functional
Priority	SHOULD
Requirement ID	VC-NFR-03

Title	Communication with agent
Description	The system must allow users to communicate with virtual agents.
Typology	Non-Functional
Priority	MUST
Requirement ID	VC-NFR-04

Title	Communication with agent
Description	Users must be able to input their feedback through voice and hand-held controllers.
Typology	Functional
Priority	MUST
Requirement ID	VC-FR-01

Title	Virtual conference space
Description	The system must have a virtual conference 3D space.

Typology	Non-Functional
Priority	MUST
Requirement ID	VC-NFR-05

Title	Virtual conference space
Description	The 3D space must comprise entrances/exits, lobby areas, trade shows, meeting rooms, social areas and a grand conference/plenary room.
Typology	Functional
Priority	MUST
Requirement ID	VC-FR-02

Title	Virtual avatars
Description	The system must allow users to be represented by virtual avatars in the conference venue.
Typology	Non-Functional
Priority	MUST
Requirement ID	VC-NFR-06

Title	Virtual avatars
Description	These avatars could be programmed to deliver visual feedback complementing available real-time multilingual translation and captioning (sensible avatars).
Typology	Functional
Priority	COULD
Requirement ID	VC-NFR-03

Title	Virtual agents
Description	The users must be able to communicate with virtual agents that help users explore the virtual conference venue.
Typology	Non-Functional
Priority	MUST
Requirement ID	VC-NFR-07

Title	Virtual agents
Description	The agents should be able to provide information about the context and contents of the rooms and answer questions from the users.
Typology	Functional
Priority	SHOULD
Requirement ID	VC-FR-04

Title	Virtual agents
Description	The agents should be able to provide navigation information to the users.
Typology	Functional
Priority	SHOULD
Requirement ID	VC-FR-05

Title	Virtual agents
Description	The agents could have the form of a humanized avatar.
Typology	Functional
Priority	COULD
Requirement ID	VC-FR-06

Title	One-to-one communication
Description	The platform should allow user avatars to talk between them.
Typology	Non-Functional
Priority	SHOULD
Requirement ID	VC-NFR-08

Title	Talk session one-to-many
Description	The platform should allow a user to give a speech.
Typology	Non-Functional
Priority	SHOULD
Requirement ID	VC-NFR-09

Title	Talk session many-to-many
Description	The system could allow users to engage in Q&A many-to-many sessions.
Typology	Non-Functional
Priority	COULD
Requirement ID	VC-NFR-10

Title	Input video stream
Description	The system could allow users to up-stream video in order to make presentations or give a speech.
Typology	Non-Functional
Priority	COULD

Requirement ID	VC-NFR-11

Title	Language translation
Description	The system should be able to provide translation from and into the consortium-selected languages.
Typology	Non-Functional
Priority	SHOULD
Requirement ID	VC-NFR-12

Title	Language translation
Description	The system should be able to provide translation in the form of live captions.
Typology	Functional
Priority	SHOULD
Requirement ID	VC-FR-07

7.2.2 Augmented Theatre

Use case-specific technical requirements, tailored for the Augmented Theatre scenario are:

Title	AR HMD
Description	The system must support the use of AR glasses to provide transparent overlays with captions and effects.
Typology	Non-Functional
Priority	MUST
Requirement ID	TR-NFR-01

Title	AR HMD video capture
Description	The HMD must support video capturing to input what the user sees into the system.
Typology	Non-Functional
Priority	MUST
Requirement ID	TR-NFR-02

Title	Scene audio capture
Description	An array of microphones may be placed on-stage and record the action clearly.
Typology	Non-Functional
Priority	MUST
Requirement ID	TR-NFR-03

Title	Language translation
Description	The system should be able to provide translation from and into the consortium-selected languages.
Typology	Non-Functional
Priority	SHOULD
Requirement ID	TR-NFR-04

Title	Language Translation
Description	The system should be able to provide translation in the form of live captions.
Typology	Functional
Priority	SHOULD
Requirement ID	TR-FR-01

Title	Reactive VFX overlays
Description	The system must be able to provide unscripted visual effects in the form of aligned AR overlays.
Typology	Non-Functional
Priority	MUST
Requirement ID	TR-NFR-05

Title	Actor tracking
Description	The system must be able to track actors on-stage to ensure proper visual effects' alignment.
Typology	Non-Functional
Priority	MUST
Requirement ID	TR-NFR-06

7.2.3 Training Assistant

Use case-specific technical requirements, tailored for the Training Assistant(s) scenario are:

Title	AR HMD
Description	The system must support the use of AR glasses to provide transparent overlays with instructions.
Typology	Non-Functional
Priority	MUST
Requirement ID	PA-NFR-01

Title	AR HMD video capture
Description	The HMD must support video capturing to input what the user sees into the system.
Typology	Non-Functional
Priority	MUST
Requirement ID	PA-NFR-02

Title	AR HMD audio capture
Description	The HMD must support audio capturing to input what the user says into the system.
Typology	Non-Functional
Priority	MUST
Requirement ID	PA-NFR-03

Title	AR HMD sound reproduction
Description	The HMD must have built-in speakers to playback sounds to the user.
Typology	Non-Functional
Priority	MUST
Requirement ID	PA-NFR-04

Title	AR HMD hand-tracking support
Description	The HMD must support hand-tracking to enable virtual tools' manipulation.
Typology	Non-Functional
Priority	MUST
Requirement ID	PA-NFR-05

Title	Virtual components library
Description	The system must provide a library of virtual tools and components that can be manipulated by the user in the training scenario.

Typology	Non-Functional
Priority	MUST
Requirement ID	PA-NFR-06

Title	Difficulty-level system
Description	The system could provide users with the option to select a difficulty mode during the training scenario.
Typology	Non-Functional
Priority	COULD
Requirement ID	PA-NFR-07

Title	Difficulty-level system
Description	In easy mode, the 3D model always fits correctly to its location when in the correct place; in hard mode, the user needs to put the parts into the correct location and orientation.
Typology	Functional
Priority	COULD
Requirement ID	PA-FR-01

Title	Communication with agent
Description	The system must allow users to communicate with virtual agents.
Typology	Non-Functional
Priority	MUST
Requirement ID	PA-NFR-08

Title	Communication with agent
Description	Users must be able to input their feedback through voice and hand-tracking channels.
Typology	Functional
Priority	MUST
Requirement ID	PA-FR-02

8 Conclusion

This deliverable describes in detail the use cases and the requirements (user and technical) of the VOXReality project. It is a live document that will be updated and edited again in M22, producing D6.2.

The results in this deliverable are based in a two-phases user centric process, where end-users are placed in the centre of the methodology. Initial requirements were gathered in the first phase, consisting of weekly calls between the different stakeholders of the project and an initial description of the use cases. Based on the initial requirements, targeted interactive focus groups were organised in three in-situ visits (VRDays, AEF, HOLO). These focus groups provided sufficient results for the initial identification of the final version of the requirements and the final description of the use cases.

Overall, the participants of the focus groups provided valuable insights into the needs and preferences of users for the AR/VR use-cases. By bringing together a diverse group of participants, we were able to gain a comprehensive understanding of the different factors that contribute to a successful user experience. These insights are intended to guide the design and implementation of VOXReality applications; ultimately leading to more engaging, interactive, and successful AR/VR conferences that provide users with a truly immersive experience. The workshops were successful and provided a valuable opportunity for both VOXReality team and workshop participants to learn, explore, and collaborate in the exciting and rapidly evolving world of VR and AR. The results of this workshop demonstrate the importance of a user-focused approach and we look forward to seeing how these insights can be applied in scientific and industry practices.

Based on the final version of the requirements (user and technical), an initial workflow of information diagram has been created per use case. This enables further collaboration between technical and non-technical partners, providing a common understanding of the ultimate goal and ambitions. The follow-up of this deliverable is the work towards an architecture of the system and the adequate planning of the development and deployment of the system. This will be reported in the next deliverable of the work-package due in M8, D2.3 Development infrastructure and integration guidelines.

References

[1] M. Grieves and J. Vickers, “Digital twin: Mitigating unpredictable, undesirable emergent behaviour in complex systems,” *Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches*, pp. 85–113, Jan. 2016, Doi: 10.1007/978-3-319-38756-7_4/FIGURES/7

[2] K. M. Alam and A. el Saddik, “C2PS: A digital twin architecture reference model for the cloud-based cyber-physical systems,” *IEEE Access*, vol. 5, pp. 2050–2062, 2017, Doi: 10.1109/ACCESS.2017.2657006.

[3] van Boeijen, A., Daalhuizen, J., & Zijlstra, Delft Design Guide: Perspectives, models, approaches, methods. (2nd ed.) BIS Publishers, 2020.

[4] Kaner, Sam. Facilitator's guide to participatory decision-making. John Wiley & Sons, 2014.

[5] KALLIO, Hanna, et al. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. Journal of advanced nursing, 2016, 72.12: 2954-2965.

[6] G. Guest, K. M. MacQueen, and E. E. Namey, Applied thematic analysis. Sage Publications, 2011.

[7] L. S. Nowell, J. M. Norris, D. E. White, and N. J. Moules, “Thematic analysis: Striving to meet the trustworthiness criteria,” International journal of qualitative methods, vol. 16, no. 1, p. 1 609 406 917 733 847, 2017.

[8] J. Fereday and E. Muir-Cochrane, “Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development”, International journal of qualitative methods, vol. 5, no. 1, pp. 80–92, 2006.

Appendix 1: Weekly Call Minutes

1st Technical/Use case Meeting

The meeting was held on 21-10-2022 in Teams General Channel

Attendees: *(Names and Surnames hidden for privacy reasons)*

Name	Partner
	UM
	CERTH
	MAGGIOLI
	HOLO
	AF
	F6S
	ADAPT
	SYN
	CWI (NWO-I)
	VRDAYS

Minutes:

Use case owners:

- VR Conference – VRDAYS
- Augment Theatre – AF
- Training Assistant – HOLO

VR Conference:

- Multilingual Translation, Navigation Assistant, Avatars
- Not physical but virtual only settings
- Visitor can participate online/remote but we should not get public users. The users should participate in a controlled environment.
- Can we deliver the content that is happening physically in real time (stream)? Maybe a hybrid setting? Answer: We should not go for hybrid but a stream of video can happen in the virtual environment
- Digital agent:
    - Where to go?
    - What is happening currently in the venue?
    - Contextual information about where the user is or where they are headed.
- HOLO can integrate their work in any framework if that framework is compatible with Unity
- The context, how it will be integrated, how it will be provided should always be the focus of discussions regarding the usecase scenario

Augmented Theatre:

- A play in the festival: Play is English and the audience is Greek or Greek play international audience
- Real usage or controlled experiment? Answer: Controlled experiment, not public
- Audience demographic:
    - Different age, gender and language
    - Different technology acceptance or experience
- The duration of the play will be defined
- AF will start looking into collaborators who are familiar with similar technologies
- AF will start brain storming sessions with possible collaborators
- We will not go for disabilities as a case but we can put emphasis on the potential of VOX technologies in providing access to people with disabilities (maybe auditory disabilities)
- The context of the play, how it will be integrated, how it will be provided should always be the focus of discussions regarding the usecase scenario
- If we already have the script, why don't we translate the script offline?
    - We will not have the script but have contextual information (summary, description etc.)
    - For the VFX, we will not have the whole script again but define trigger words or phrases.

Training Assistant:

- We should start investigating which industry we will do the use case in
- Remote assistance to a user is aimed
- It could be repair, maintenance or training
- How will the user communicate with the agent? Answer: Voice + controller input
- HOLO plans to explore HoloLens, Oculus VR and Magic Leap
- This use case will follow a single example industry assistance case
- HOLO will investigate the headsets we can use in the usecase.
    - It is best if we use the same equipment for all the cases.
    - In other words, meaning if we are going to use HoloLens, all use cases that will use a headset should use HoloLens.
    - This would minimise development and deployment workload of technical partners and training/learning load of end-user partners.
- HOLO also confirms that with minimum modifications their framework can be integrated to different hardware configurations.
- The context, how it will be integrated, how it will be provided should always be the focus of discussions regarding the usecase scenario

2nd Technical/Use case Meeting

The meeting was held on 31-10-2022 in Teams General Channel

Attendees: *(Names and Surnames hidden for privacy reasons)*

Name	Partner
	UM
	CERTH
	MAGGIOLI
	HOLO
	AF
	ADAPT
	SYN
	CWI (NWO-I)
	VRDAYS

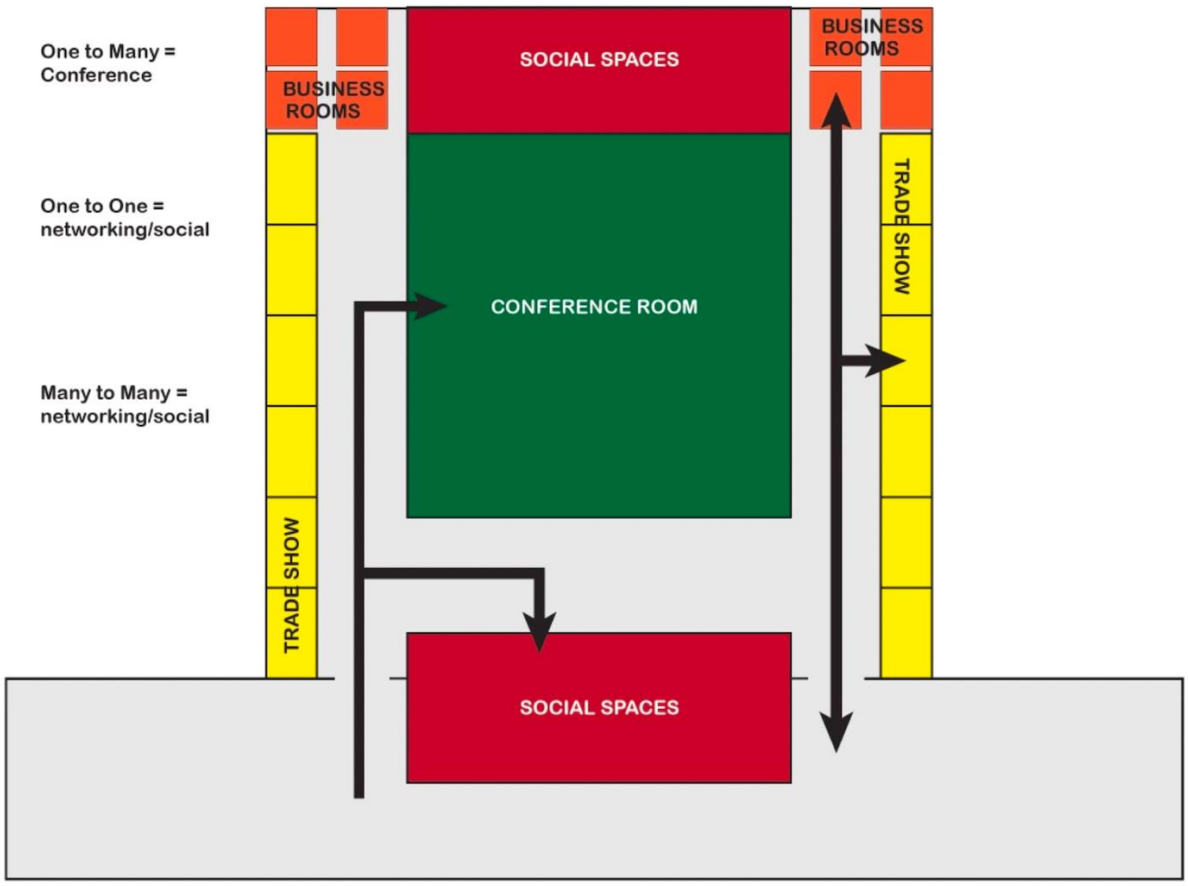
Minutes:

General:

- Next meeting: 7 November 2022 14:00 CET
- The use case scenario definition template has been uploaded to SharePoint.
- End user partners to start filling the document with their respective use cases
    - VRDAYS: VR Conference
    - AF: Augmented Theatre
    - HOLO: Training Assistant
    - ***Please download a copy and upload the filled document to the same folder to create a version of the document for each usecase***
- At least 1 Technical partner will support the usecase execution.
    - Which partner to support will depend on the expertise + location
- Do we need to have a use case demonstration (pilot) at the end of the project or the pilots of the third parties will act as the “second pilot”?

VR Conference:

- Manuel prepared an example venue sketch for a virtual conference:



- Possible cases that could be explored in the VR Conference case:
  - 1-to-1: Business meeting / Business rooms
  - 1-to-Many: Keynote speech / Conference room
  - Many-to-Many: Networking / Social spaces & Trade show
- Normally the time set for these cases are as follows:
  - Keynote: 30-60 minutes
  - Meeting: Up to 30 minutes
  - Networking: Around 10 minutes
- How are the emotions synthesized in the avatars?
  - Should be voice oriented.
  - Emotion (e.g., voice prosody) and Motion (e.g., Gestures)
- How to integrate or investigate sarcasm?
  - There are studies that detect sarcasm.
  - This should not be a priority.
  - Investigation of these kinds of models could be investigated in the future.
- Context:
  - What avatar is seeing + External visual data
  - 2 separate models for the above case could be developed and wrapped in a single module for deployment
- VR Conference Agent
  - Should help navigate the user
  - Should inform the users about the context/content of the room they are in using visual-language models.
- An example scenario could be the following:

- 1 VR User (V) + 1 PC User (P) + 2 experimenters (E)
- V meets E at the booths while P goes to the keynote (10 minutes)
- P meets E at the booths while V goes to the keynote (10 minutes)
- P and V have a 1-to-1 meeting at the meeting room (5 minutes)

Augmented Theatre:

- AF can start contacting the creators who will perform at the festival in Summer 2023 and build on top of their performance
- Equipment to be used should be finalized and an estimate budget should be decided for reaching multiple audience members at the performance.
  - This would determine the nature of the demonstration scenario
  - There is a use case equipment budget for each end user organization
  - Alberto to talk to Nikos for the equipment price estimate
- Theatre use case should be demonstrated with AR glasses.
- 15-20 minutes could be the duration of the performance
  - We should also check the simulator sickness limits for the selected AR glasses.
- Evaluation: subject-based or group-based?
- Depending on the power analysis (and the evaluation criteria, different groups, demographics) we can determine the minimum number of participants.
- We do not need to validate the performance of the translation with the end users.
- The use case evaluation should focus on the experience of users with the technological features of VOXReality.
- VFX should be evaluated by the users.
  - Creator experience could also be evaluated based on rehearsal performance including VFX.
  - They could evaluate how their work is presented through the technological medium regarding VFX
- The characteristics of an example scenario:
  - A play which is not dialogue heavy
  - Max 2-3 actors on stage at the same time
  - A narrator could be included who is not on stage
    - Could be the person who triggers the VFX
  - A short dialogue on stage
- The scenario must provide time for each technological component to be experienced by the audiences individually
  - A combined demonstration could also be included.
  - For example, 5-minute narration + 5-minute dialogue + 5-minute VFX + 5-minute dialogue/narration with VFX

Training Assistant:

- HOLO is analysing the different assistance paths: training, maintenance, repair etc.
- The usecase scenario should include a single path, a single industry and a single demonstration scenario.
- HOLO will update the consortium in the next meeting about this use case

3rd Technical/Use case Meeting

The meeting was held on 07-11-2022 in Teams General Channel

Attendees: *(Names and Surnames hidden for privacy reasons)*

Name	Partner
	UM
	CERTH
	MAGGIOLI
	HOLO
	AF
	ADAPT
	SYN
	CWI (NWO-I)
	VRDAYS
	F6S

Minutes:

General:

- Next meetings: Mondays in the next 5 weeks 14:00 CET will be reserved for these meetings
  - 14-11-2022, 21-11-2022, 28-11-2022, 5-12-2022, 12-12-2022
- **AP:** End user partners to start filling the document with their respective use cases
  - HOLO: Virtual Assistant
  - VRDAYS: VR Conference
  - AF: Virtual Theatre
  - ***Please download a copy and upload the filled document to the same folder to create a version of the document for each usecase***
  - *Template can be found here.*
- Do we need to have a use case demonstration (pilot) at the end of the project or the pilots of the third parties will act as the "second pilot"?
  - **AP:** F6S and MAG will clarify this.
- CWI will start the focus groups for the requirement analysis in the coming weeks.
  - They will organize meetings per use case for this.
  - The timeline for these meetings to be shared in one of the next meetings.

Training Assistant:

- Training vs Remote support
  - Training: A machine assembly training scenario
  - Remote support: Help replace a part of a machine or troubleshoot a software

- In both cases a step-by-step instruction set will be included
- Visual instructions are needed:
    - Showcasing small actions that are essential in the assembly procedure
- What are the training data for the digital agent?
    - Answer: The visual instructions, step-by-step guide can be pre-set but the assistance from the agent should be determined by a language model
    - Furthermore, the audio inputs from the users should be analysed to obtain the right instructions and right position in the guidelines
    - We need to find the appropriate down-stream task datasets based on the selected application area.
- We can use off-the-shelf text-to-speech tools after the fine-tuning phase to give audio feedback to the users if it is necessary for the use case.
    - We should not develop a new tool for this since it is out of scope
- User profile: Beginner trainee with some knowledge about the scenario
- If the instructor/agent/assistant will have a 3D avatar (character like the Microsoft Clippy), it should have movements:
    - Show and direct the user towards objects and demonstrate movements that are necessary for the application for example certain stance for assembly of a machine part
    - If the agent is a stable character next to the user, it would not have the desired effect.
- A model like this could be purchased or motion capture could be included for the use case development
    - We could collect motion/movement data from human instructors only once for the demonstration purposes.
- If we want to include only the hands of the user for the assembly of a machine, HOLO already has this technology but if we want body positioning for the assembly new tools should be developed/purchased.
- Controller or hand input? Answer: Hand input is more natural/realistic. It is also available at HOLO and hardware SDKs.
- Scenario example:
    - User asks: “How can I do this?”
    - Agent finds the correct guide and acquires its visual content directs the user step-by-step through the guide with verbal/textual explanations, also displays the corresponding visual content
- Speed of the assembly with and without agent could be an evaluation measurement
- English can be used/We do not need to utilize translation for this usecase
- Location:
    - The end user facilities / HOLO facilities / A rental venue
    - HOLO could reach out to FILL Austria for the demonstration

VR Conference:

- We are going to use VR headset
- Everything happening on the way to the rooms could potentially be considered a part of the digital navigation assistant
- MAG has experience with the creation of a venue in a virtual setting.
- ISAR SDK (HOLO) has compatibility with Oculus Quest, WebXR, MagicLeap 2, Qualcomm Lenovo, HoloLens

- Proximity audio integration will be investigated
    - This is the audio from the nearest location to enable an immersive experience (just like being in a venue)
- All outputs will be text from the models
- Agent can integrate scene description from the Visual Language models
    - This is the output of the VL model with the scene the user is looking at
    - This output will need to be generated already for the context integration
    - User can ask the agent: “Please describe the scene for me.”, “What is in this room?”, “Is there an empty chair in the room?” etc. This would trigger the explanations of the agent on the way to the destination.
- If needed, we can enable the text outputs to be utilized by ready text-to-speech tools on the setup e.g., mobile phone accessibility features.
- Speech synthesis could be technically demonstrated with the VOX pipeline but it is out of scope for the project.
    - Maybe one of the open call applications can address this.

Augmented Theatre:

- Will we collect audio from the headsets or will we introduce a higher quality microphone for the actors and/or the stage?
    - We need to determine the AR equipment for the theatre case so that we can test the quality of the headset microphone
    - AF can test it as soon as they know the model and purchase it.
    - If each actor has a microphone attached to them, we solve the user association problem and we do not need to be concerned about the audio quality.
    - During the pilot study we can use microphones attached to the actors and a single camera on stage and the association could be a part of the research
        - We can say the final version will work with a microphone on stage
- HoloLens is 3.5K Euros. Qualcomm or MagicLeap could be cheaper since they are consumer products.
    - **AP:** HOLO will check this.
- Microphone models should be investigated and the model to be used must be finalized

4th Technical/Use case Meeting

The meeting was held on 14-11-2022 in Teams General Channel

Attendees: *(Names and Surnames hidden for privacy reasons)*

Name	Partner
	UM
	CERTH
	HOLO
	AF
	SYN
	CWI (NWO-I) VRDAYS

Minutes:

General:

- We will organize a WP3/WP4 joint technical meeting next week to align the development/implementation ideas
- Use case designers (HOLO, AF, VRDAYS) will fill the template provided for the use case definitions until the next meeting (21-11-2022 Monday)

Technical:

- CERTH will start the Visual Language models development with existing datasets
    - Papers CERTH is investigating for baseline:
        - LXMERT: Learning Cross-Modality Encoder Representations from Transformers
        - UNITER: UNiversal Image-TExt Representation Learning
        - CPT: Colorful Prompt Tuning for Pre-trained Vision-Language Models
        - Dense Contrastive Visual-Linguistic Pretraining
- Both the context and the communication between components will be in TEXT format
- We need to start working on the technical developments regardless of the completion of use case scenarios, especially we need to start building the baseline tools
- HOLO is the leader of the use case application development task.
    - CERTH, MAG, SYN and ADAPT are the contributors to this task
    - The responsible partner (or joint contribution) for the development of each usecase needs to be defined in a separate usecase development meeting.

Training Assistant:

- HOLO decided on the Assembly Training for this use case
- The scenario will be with AR glasses.
- The scenario will include putting together a machine
- CERTH cannot support motion capture for this case
- HOLO can collect hand movement videos for additional content
- Text-to-speech is probably a requirement for this use case

    - We should not develop new technology for this but a ready tool can be used during the usecase development phase.
    - Agent should generate only TEXT results; it can then be converted into speech using off-the-shelf tools
- Next step for the use case is to create a storyboard.
    - Example:
        - Wear headset
        - “Teach me how to do this”
        - Agent starts the training
        - The agent should collaborate with the user
        - “Am I doing this right?”
        - Agent will take visual information when answering
- NReal is a cheap (Around 400 Euros) alternative to HoloLens for AR equipment.
    - HOLO will check if this device will be supported by their existing SDKs (ISAR)
- The digital agent will use the visual scene as context
    - Additionally, the feasibility of integrating Unity Events will be investigated.

Augmented Theatre:

- We will use a single type of equipment (AR) throughout the theatre use case.
    - Introducing VR as well would increase the number of users for validation and new developments on the technical side
- Equipment budget/price limits the selection of structure of the play
    - AP: We will ask MAG whether it is okay to rent equipment from the project budget for the use cases
- In two weeks, we will hear from AF about possible collaborators and possible plans
- 30-40 minutes sessions including the play, preparation of participants and the post questionnaires.
    - 10-12 minutes play
    - 2 users at the same time
- Scenography + script could be used for context
    - Director’s notes could also be included for sub-text, intention, mood, emotions

5th Technical/Use case Meeting

The meeting was held on 21-11-2022 in Teams General Channel

Attendees: *(Names and Surnames hidden for privacy reasons)*

Name	Partner
	UM
	CERTH
	VRDAYS
	AF
	SYN
	ADAPT
	CWI (NWO-I)

Minutes:

General:

- The meeting will be held in the first week of December
- During the joint WP3 and WP4 meeting we will decide who is responsible for each task.

Requirements:

- During VRDAYS CWI will start the user focused analysis of the usecases that are presented in the meeting next Wednesday by their owners (VRDAYS, HOLO if present)
- Formal focus groups will be started after Christmas by CWI
  (Onsite, if possible, online where necessary)

VR Conference:

- The work on the template is ongoing
- What VR headset to use?
  - Probably Oculus Quest 2
- Usage of the recordings of this year's VRDAYS:
  - Do we have the consent of the participants/speakers for data processing?
  - Potential usage:
    - To fine-tune translation or speech recognition models
    - As a test set for the application
- Important points to consider in the use case:
  - What is context?
  - What the digital agent should do?
  - How should navigation look like?
  - What cues the agent should respond to?
  - What is the story of the user in this use case

Augmented Theatre:

- Photos of the real-world location of the pilot is necessary for the requirements.
- To decide: What will be the number of AR headsets during plays?

    - Sharing equipment between partners may be possible
- The play has not yet been decided
- What to consider:
    - What is context?
    - In what format will the context be?

6th Technical/Use case Meeting

The meeting was held on 05-12-2022 in Teams General Channel

Attendees: *(Names and Surnames hidden for privacy reasons)*

Name	Partner
	UM
	CERTH
	HOLO
	AF
	SYN
	ADAPT
	CWI (NWO-I)
	MAGGIOLI

Minutes:

General:

- The plural of Pilots in proposal means the 3 sub-pilot (usecases).
- We can, internally, perform 2 phases of the pilot study.
    - Phase 1: small group of users – evaluation of first version of the components
    - Phase 2: relatively larger group – evaluation of the final versions
- HOLO is the leader of the development of the applications task, not the sole developer.

Technical:

- MAG has a new technical team that has 3 people who will work on the developments required from MAG.
- NReal is not planned to be included into the SDK of HOLO in 2023.
    - Carina can push it internally for the project if we decide to move forwards with this equipment
- The equipment purchases:
    - Everybody purchases for themselves, one partner purchasing for everyone does not work.
    - We can use the equipment purchased by other partners during the usecase evaluations (pilots) by bringing them to the usecase designer facilities during the pilot (burrowing it).
    - The equipment should be purchased latest next spring

Requirements:

- CWI to prepare and share the D2.1 (M6) ToC draft.
- CERTH will organize the technical requirements calls as soon as possible

- Before Christmas CWI will share the timeline for focus groups and these meetings will start from 16th of January 2023

Augmented Theatre:

- Avant-garde theatre is more suitable for AR technology
- Play is planned to be ancient tragedy
- Original text, the synopsis and character specifications will arrive before Christmas
    - The translation of the original text will come at a later date when the director is chosen
    - Director's notes will be available as context as well
- Data collection: Spring 2024 the earliest
- Other hardware that will be used needs to be finalized:
    - Microphones, special lights (maybe)
- The testing of the project should not influence the normal production of the theatre play
    - The testing should be done "on the side" of the normal production
- The question of how to finance the compensation to the theatre, etc. will be discussed in the future
- AF saw a similar project (manual subtitles and visual effects with depth cameras)
    - Possible meeting with the technical director of the other project can be organized (CERTH)
    - Problems she identified:
        - Delays on the VFX and the subtitles
        - Setup time of the testing
        - 10 minutes with glasses, rest without glasses
            - People who were not wearing the glasses were all wondering what was happening that they were not seeing, which is not desired.
        - Visual effects
            - Too cartoonish
            - Sometimes outside of the stage which made users to move their heads outside of the stage and miss other things while looking at them.
    - Our visual effects should be less ambitious, artistically relevant and delivered in a good quality rather than too ambitious and delivered poorly.

Training:

- User asks for instructions to assemble a machine/object and selects a difficulty mode
- We have the real (physical) gray base the parts will be assembled on.
    - Tools, screws, objects, parts are virtual – easier to track the progress of the user
- Easy mode: 3D model always fits correctly to its location when in the correct place
- Harder modes: user needs to put the parts into the correct location and align the part's orientation.
- HOLO already has the app/components for the virtual assets and the actions/events related to them

During VR Days meeting:

- Visited exhibitions and presentations at the conference
- Presented the general idea of the project

- Including the Open Calls
- Technical discussions:
  - What is context and how to get it?
  - What can we expect from the virtual agent?
  - What will be the experience of the user in the conference use case?
  - Machine Translation with context information.
    - Context is summarized by the textual description provided by the visual language models
  - Virtual Agent
    - Agent will act as the dialogue generator and the component that communicates with the user.
    - 2 modules will be developed separately that will feed the digital agent: navigation and instruction models
    - Path generation (navigation in Conference UC)
      - Using contextual visual information
      - Voice communication
    - Instruction generation (instructions for Training UC)
      - Generate the next step based on the dialog (text) and visual information
      - Output in English, but could be translated by our MT component
      - Off-the-shelf ASR module will be used as the first step (audio -> text)
- Discussed when and how the development of various parts has (will) began?

7th Technical/Use case Meeting

The meeting was held on 12-12-2022 in Teams General Channel

Attendees: *(Names and Surnames hidden for privacy reasons)*

Name	Partner
	UM
	CERTH
	HOLO
	AF
	SYN
	ADAPT
	CWI (NWO-I)
	MAGGIOLI

Augmented Theatre:

- The play and the scene will be decided/selected in December.
- The director will be decided in January
- The specification/selection of the equipment is required.
    - This will be decided in the future (when the computing infrastructure will be known)

Training:

- Headset has been decided: HoloLens 2.
- Navigation is not part of this use case.
- The scenario was presented during the meeting.
    - The higher resolution of the assembly task is needed.
        - What is the machine/task?
        - What steps will be performed by the user?
    - How is the progress of the user monitored?
- What kind of data will be available to train the model?
- AP: More detailed version of the presented scenario will be uploaded tomorrow.

Use Case Templates:

- Technical partners will help with filling the technical parts.
- CWI will help with specifying the assessment protocol.
- The feedback will be put in the uploaded files by other partners.

Deliverable: User and technical requirement

- Should be “per-use case”
- Use cases need to be detailed

Technical meeting

- AP: Technical partners should fill in their availability in the doodle provided by SYN.

VOXReality

Voice driven
interaction in XR spaces

Funded by
the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Directorate-General for Communications Networks, Content and Technology (DG CNECT). Neither the European Union nor the granting authority can be held responsible for them.