VOXReality

VOICE DRIVEN INTERACTION IN XR SPACES

A guide to integrating voice interactions in XR: insights from VOXReality

Funded by the European Union

Authors	Alberto Casanova (MAG), Greta Ioli (MAG), Olga Chatzifoti (MAG), Spiros Borotis (MAG), Manuel Toledo (VRDAYS), Eleni Oikonomou (AF), Moonisa Ahsan (NWO-I), Pablo Cesar (NWO-I), Leesa Joyce (HOLO), Petros Drakoulis (CERTH), George Nikolakis (SYN), George Kliafas (ADAPT).
Reviewers	Konstantia Zarkogianni (UM), Yusuf Can Semerci (UM), Mateusz Kowacki (F6S)
Abstract	This document details various lessons learned, success stories, best practices and recommendations emerged during the implementation of the VOXReality project. It also presents the Key Exploitable Results identified, the relevant business models and specific pathways for post-project activities.
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

Nature	
RE	Report

Table of Contents

List of Abbreviations & Acronyms

AEF	:	Athens Epidaurus Festival
AI	:	Artificial Intelligence
API	:	Application Programming Interface
AR	:	Augmented Reality
ASR	:	Automatic Speech Recognition
CA	:	Conversational Agents
CAD	:	Computer-Aided Design
CV	:	Computer Vision
GPU	:	Graphics Processing Unit
KPI	:	Key Performance Indicator
LLM	:	Large Language Model
NDI	:	Non-Destructive Inspection Methods
NLP	:	Natural Language Processing
NMT	:	Neural Machine Translation
OC	:	Open Calls
R&D	:	Research & Development
TP	:	Third Party
TRL	:	Technology Readiness Level
TTS	:	Text-to-Speech
UI	:	User Interface
UX	:	User Experience
VFX	:	Visual Effects
VL	:	Vision-Language
VR	:	Virtual Reality
WP	:	Work Package
XR	:	Extended Reality

List of Figures

1. Introduction

1.1. What is VOXReality?

VOXReality is an ambitious project with the goal of facilitating and exploiting the convergence of two major technologies, Natural Language Processing (NLP) and Computer Vision (CV) for improving the user experience of immersive applications. Both technologies are experiencing a performance boost due to the emergence of large-scale Artificial Intelligence (AI) models and carry potential to drive the XR revolution beyond what was possible up to now through natural speech-based interfaces and visual content understanding revolutionizing human-machine interaction. By exploiting the power of large AI models and employing economical approaches to enable widespread utilization, VOXReality project aims to integrate language-based and vision-based AI models into cutting-edge extended reality (XR) applications.

The results of the project are twofold: a set of pretrained next generation AI models combining, in various levels and ways, language and vision modalities to provide more natural immersive experiences that are expected to boost XR adoption and a set of XR applications to showcase the capabilities of these models to demonstrate innovations in various sectors. The technologies mentioned above are validated through three use cases: Augmented Training, Virtual Conference and Augmented Theatre.

1.2. VOXReality core technologies

In VOXReality, while language models offer naturalistic human-human and human-machine communication, vision systems unlock an extra step for realistic interactions by providing additional context into the pipeline. Multimodal experiences such as real-time translation, interactive tutorials, and individualized help are made possible by this fusion which extends the potential of XR technologies in cultural, educational, and professional contexts.

The core technologies developed in VOXReality include Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), Vision-Language (VL) models and Conversational Agents (CA) while these technologies are applied differently across the use cases:

- Virtual Conference uses all four components,
- Augmented Theatre uses ASR, NMT, and VL models,
- Augmented Training relies on ASR and CAs.

The aim is to build intelligent, immersive, and multilingual XR experiences using pretrained models powering demonstrative applications.

1.3. The VOXReality Pilots

In VOXReality, the consortium conducted three pilot studies to validate and demonstrate VOXReality AI components in real world applications. Every pilot was designed to utilize a set of the AI models to answer specific needs in different sectors.

1.3.1. Augmented Training Pilot

In this pilot, the goal is to investigate the effectiveness of AI-powered Augmented Reality (AR) systems for training industrial workers, particularly in assembly tasks. Users interact with holographic 3D Computer-Aided Design (CAD) models while wearing AR glasses, such as the

HoloLens 2 in addition to the speech interaction with the AI assistant. They are guided through the assembly process by a virtual assistant that provides visual cues, real-time support, and performance-based feedback. The goal is to offer a customized and scalable training experience.

1.3.2. Virtual Conference Pilot

With spaces such as a lobby, exhibition stands, social areas, a business meetings area and a main conference hall, this pilot study focuses on a replica of a formal conference environment. Participants are assisted with navigation, conference information, and real-time translation in six languages by a virtual agent (VOXY). The experience includes features such as interactive maps, guided tutorials, and a fully immersive Virtual Reality (VR) environment to test spatial navigation and AI-assisted multilingual interactions.

1.3.3. Augmented Theatre Pilot

This pilot study demonstrates the use of AR captions and visual effects (VFX) in a live theatre production of the ancient Greek play Hippolytus. The AR captions are synchronized in real time with actors' speech using a dedicated AI-based automatic speech recognition (ASR) service that performs online transcription and text alignment. Captions are offered in the original language (Greek), as well as in all other VOXReality languages using pre-generated translations produced with a human-in-the-loop translation methodology (machine translation with human expert proofreads). Users are also offered extensive customization options for improved readability and comfort of the captions. To enable user engagement with said options, users are supported through step-by-step tutorials to familiarize themselves with the AR device and interaction methods. Finally, users can enjoy audiovisual effects triggered automatically in sync with the play's events either through the actors' speech (with the same methodology as the captions), or through physical stage events, such as actors' entrance and exit from the stage, using a custom pipeline built on the VOXReality vision-language services.

1.4. About this guide

The purpose of this guide is to present and analyse the Lessons Learned, Best Practices, and Success Stories gained from the project. By exploring them, readers will not only gain insight into the most significant and impactful aspects of the work that has been done in the project but also understand the challenges faced, the actions taken, and activities that can be approached differently, if given the chance. Furthermore, the guide also delves into the methodology used to identify and extract all these elements. It describes how, in collaboration with all project partners, a structured and systematic process has been developed and implemented to make this exercise as effective and meaningful as possible.

This guide is therefore not intended as a simple report of what has been achieved, rather it seeks to serve as a starting point for others who may find themselves in a similar position. At the end of any project, it is essential to take stock: to reflect on what went well and what did not, on the situations where we performed strongly and those where we could have done better. Yet this is never an easy task, there is no universal recipe or predefined set of steps to follow. Accordingly, this guide does not aim to provide a ready-made checklist or rigid methodology. Instead, it offers a set of reflections and insights that may inspire those who are about to start or are already running a project.

2. Methodology: gathering the insights

2.1. Framing the discussion: Lessons Learned, Best Practices, and Success Stories

Before starting the collection process, how other EU-funded projects had addressed the challenge of identifying, documenting, and sharing project insights were examined. This comparative review was crucial for two reasons: it provided tested methodological references, and it helped anticipate potential pitfalls.

PROCare4Life adopted a qualitative analysis approach based on real-world experiences. The focus was on identifying practical challenges that emerged during project implementation and describing the adaptive solutions developed by the consortium. The methodology emphasized a retrospective lens, looking at both technical complexities (e.g., integration of heterogeneous devices) and organizational aspects (e.g., differences across national health systems).

ERIGrid implemented an experimental validation and structured testing methodology. The goal was to standardize and optimize technological solutions in the field of power systems. This approach relied on rigorous testing procedures, reproducibility, and a holistic validation framework, which made it possible to promote comparability and collaboration across European laboratories.

SHAPE ENERGY used a stakeholder engagement and communication analysis methodology. Its approach placed strong emphasis on dissemination and public engagement. By analysing participation in workshops, events, and online activities, the project was able to identify communication-related lessons, such as the value of video content, flexibility in dissemination formats, and the role of social media as an increasingly central tool compared to traditional press releases.

FORGING introduced a structured methodology inspired by Industry 5.0. It provided a human-centred and needs-driven approach to defining use cases for emerging technologies. Its “three-canvas” method combined problem identification, stakeholder mapping, and resource planning, with a strong emphasis on societal and environmental impact. This represented a best practice in bridging technological innovation with sustainability and human wellbeing.

Search and Rescue stand out as a clear success story. The project addressed interoperability gaps among Europe’s first responders by developing new communication platforms, wearable technologies, and robotics for disaster management. Through large-scale pilots across Europe, it showcased how innovative solutions could have a tangible and durable impact on crisis response, directly demonstrating the value of EU funding.

VIDIS also provided a significant success story. By deploying low-cost sensors for air quality monitoring, the project made environmental data more accessible to citizens, municipalities, and activists. Its innovative calibration methods and strong international partnerships enabled both scientific excellence and practical impact. Importantly, VIDIS also left behind a sustainable infrastructure (the VIDIS Centre) that will outlive the project itself.

These examples highlighted that there is no single methodology to collect project insights: some projects rely on qualitative reflection, others on experimental validation, others on stakeholder engagement or structured frameworks. What they share, however, is a systematic effort to capture knowledge in a way that is valid, applicable, and useful beyond the lifetime of the project.

VOXReality utilizes a hybrid approach by balancing reflection on internal experiences (as in PROCare4Life), structured organization of findings (as in ERIGrid and FORGING), and attention to impact and visibility (as in SHAPE ENERGY, Search and Rescue, and VIDIS). Furthermore, clear definitions of Lessons Learned, Best Practices, and Success Stories are extracted from this comparative analysis.

Lessons Learned are the reflections on experiences gained during the project, both positive and negative. They focus on what worked well and what did not, highlighting the problems faced and the solutions adopted. The goal is to help future projects avoid repeating mistakes and replicate successful approaches. Lessons Learned evaluation criteria include:

- **Significance** à they represent insights with a real impact on project activities,
- **Validity** à they are technically correct and reliable,
- **Applicability** à they should provide concrete actions that can be replicated, avoided, or adapted in future contexts.

Best Practices refer to strategies, methodologies, or tools that proved to be particularly effective and efficient during the project. They are considered replicable models of excellence that can be adapted by other initiatives and not only demonstrate technical soundness but also show efficient use of resources. Best Practices evaluation criteria include:

- **Effectiveness** à they create tangible and measurable results,
- **Efficiency** à they optimize the use of resources (time, funding, personnel),
- **Replicability** à they can be reused or adapted in other projects,
- **Innovation** à they introduce new or better solutions compared to traditional approaches,
- **Sustainability** à they produce benefits that last beyond the end of the project.

Success Stories are concrete achievements that produced a significant and demonstrable long-term impact. They show the added value of EU funding, generate visibility, and inspire replication. They are usually linked to outcomes that reached institutional recognition or broad dissemination. Success Stories evaluation criteria include:

- **Measurable impact** à they produce tangible benefits for society or specific communities,
- **Innovation** à they introduce novel solutions, technologies, or approaches in a given sector,
- **Sustainability** à their results continue to generate value even after the project has ended,
- **Replicability** à they can be used as models for future initiatives.
- **Institutional recognition** à they are promoted at European level (e.g., through CORDIS, DG RTD, or official EU events).

By adopting these definitions and criteria, the VOXReality consortium ensured that all partners had the same framework for contributing insights. This alignment was critical to guarantee that

the information gathered in the following phases could be systematically analysed, compared, and communicated in a way that is valuable not only internally but also to external stakeholders and future projects.

2.2. Collecting findings: the workshop

The second step of the methodology to extrapolate the Lessons Learned, Best Practices and Success Stories was implemented through a participatory workshop designed to gather inputs from the entire VOXReality consortium. Rather than relying on individual reports or fragmented reflections, the workshop created a collective space where partners could share, compare, and challenge their experiences in a structured way. The session was carefully prepared to balance openness (allowing all voices to emerge) with structure (ensuring that contributions could be systematically captured and analysed afterwards). Organisation of the workshop was as follows:

Division into use case groups: partners were grouped according to the three VOXReality use cases, so that discussions could be anchored to concrete examples of implementation. This grouping helped avoid abstract or overly general reflections and instead encouraged participants to ground their contributions in direct project experience. Partners with role in more than one use-case participated alternatively in all relevant group discussions.

Facilitation and time management: each group was supported by a facilitator, chosen by the group's members, responsible for guiding the discussion, ensuring balanced participation, keeping the debate focused on the guiding questions and exposing the findings during the final discussion. Sessions were time-boxed to allow enough space for deep discussion while leaving room for cross-group exchange at the end.

Collaborative tools: Miro boards served as the central collection space for inputs during the workshop. The creation and use of this collaborative board were essential to the process. The board itself (see Figure 1) was structured explicitly to guide the collection process in two main horizontal sections, "USE CASE" (for specific, individual use case contributions) and "COMMON FINDINGS WITH OTHER USE CASES" (for insights that cut across multiple pilots). Both sections were further partitioned vertically into three columns, corresponding to the agreed definitions: "Lessons Learned," "Best Practices", and "Success Stories". This structure ensured that all partners, using virtual post-its, were prompted to categorize their real-time contributions precisely. The Miro board enabled participants to add virtual post-its in real time, making contributions immediately visible to the entire group and ensuring transparency throughout the exercise. The visual functionality of the tool was leveraged to facilitate the subsequent analysis phase, making it easier to cluster similar ideas and identify recurring patterns across different use cases. This combination of real-time input and visual organization ensured that all experiences, both challenges and successes, were systematically captured and made ready for synthesis.

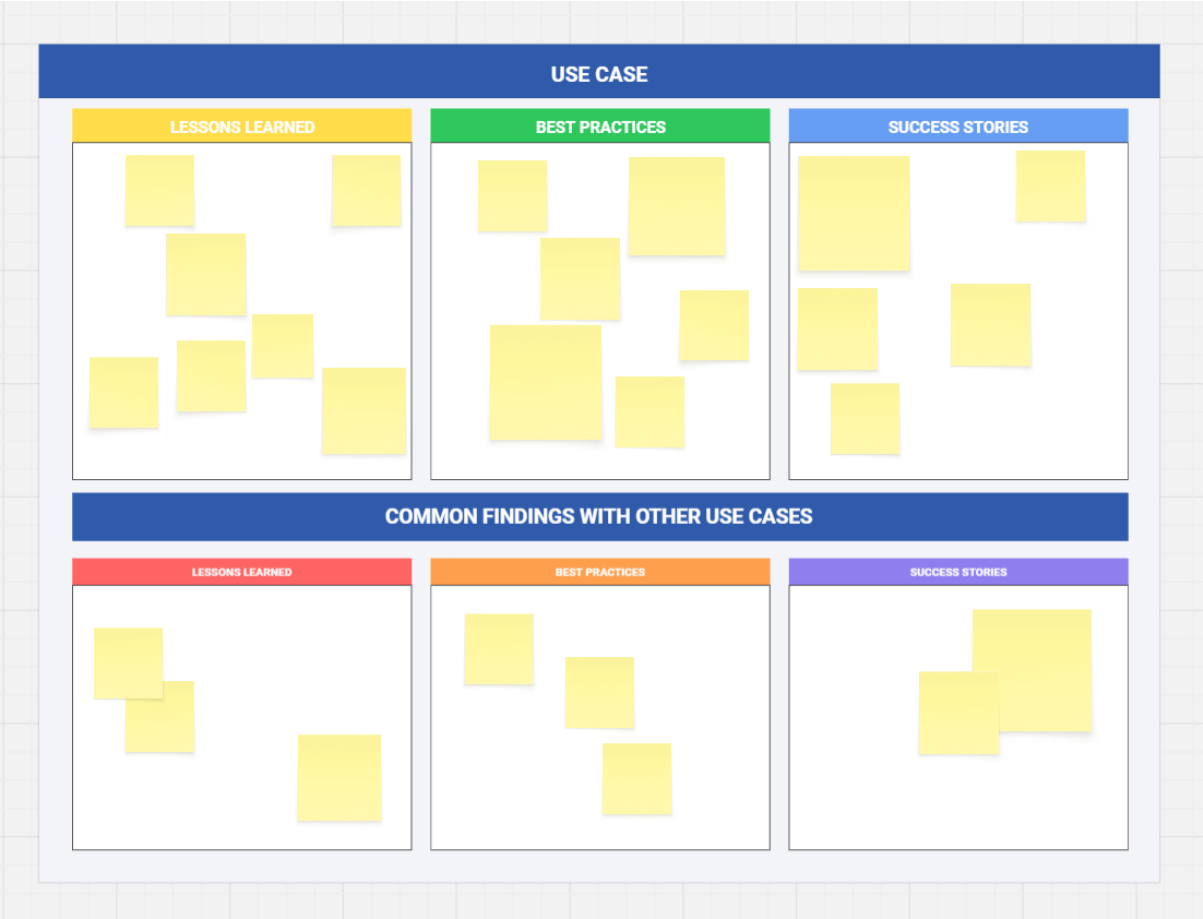

Figure 5. Miro Board

Semi-structured questions: to stimulate reflection and ensure comparability across groups, a set of guiding questions that covered different dimensions of project work were designed, including technical, organizational, managerial, and dissemination aspects. The workshop questions were deliberately designed to be both retrospective (looking back at challenges and solutions) and prospective (considering how insights could be reused or inspire future projects). For instance:

- **Lessons Learned:** What were the main obstacles encountered during project implementation? Were there technical or organizational issues that required changes to the initial strategies? Which solutions proved effective in overcoming these challenges?
- **Best Practices:** Which methodologies or strategies contributed significantly to the success of the project? Which management tools, digital platforms, or dissemination actions proved most impactful?
- **Success Stories:** What were the most significant and innovative results of the project? Did any achievements create long-term impact or recognition beyond the consortium?

The combination of group work, collaborative tools, and guiding questions produced a large body of contributions. Partners appreciated the possibility to reflect not only on their own work but also learn from the perspectives of others, identify common challenges across use cases, which revealed cross-cutting issues such as communication, resource allocation, or user engagement, and highlight both “what worked” and “what did not”, reinforcing the idea that Lessons Learned are as much about difficulties as about achievements.

The workshop thus served a dual purpose: it generated the raw material for this guide and it fostered collective learning within the consortium. By openly sharing experiences, partners were able to better understand each other's perspectives and strengthen the overall cohesion of the project.

2.3. Synthesizing the outcomes

After the workshop, a structured synthesis process to transform raw contributions into the basis for this guide took place with the following three main steps:

Organisation of the findings: all contributions from the Miro boards were collected and categorized under the three agreed labels: Lessons Learned, Best Practices, and Success Stories. Overlaps were consolidated, while recurring themes across use cases were identified to highlight cross-cutting insights.

Creation of the structure of the document: once the findings were organized, they were mapped into the overall structure of this guide. The aim was to provide a clear and logical narrative, where each insight could be contextualized and connected to the broader objectives of VOXReality.

Preparation of the guide: the categorized and structured material was expanded into written text. Short workshop notes and bullet points were transformed into a coherent narrative, ensuring that each insight was explained and illustrated in a way that would be meaningful to both project partners and external readers.

3. Lessons Learned from the VOXReality project

This chapter synthesizes the Lessons Learned, which are reflections on experiences gained during the project, both positive and negative. They focus on what worked well and what did not, highlighting the problems faced and the solutions adopted to help future projects avoid repeating mistakes and replicate successful approaches. The reflections gathered from the workshop with partners identify areas of greatest difficulty and the critical adaptive solutions adopted by the consortium, providing essential knowledge for future complex initiatives involving XR and AI.

3.1. Overcoming technical challenges

The main sources of technical challenges stemmed from the necessity of integrating complex multimodal AI models such as Large Language Models (LLM) and Neural Machine Translation into XR environments characterized by real-time requirements and resource constraints. Key lessons involved both identifying obstacles and developing effective countermeasures.

3.1.1. The cruciality of real-time performance and latency management

A primary technical goal and a significant source of difficulty was "To achieve real time and without delays communication with AI model." This need for instantaneous feedback was particularly pressing in the Augmented Theatre use-case, where multiple stakeholders (creatives, professionals, and audience) requested for “near-instant” responses (sub-second). This was exacerbated by the fact that in this use case the speed of spoken words is dictated by dramaturgy rather than best orating practices, sometimes reaching very high delivery speeds. This in conjunction with high latency risked breaking the semantic relationship between the spoken word and the generated captions causing noise or even worse, confusion. Finally, with the prevalence of captions in cinema, users compared the latency of live generated captions with that of pretimed captions in film, leading to having to meet impossible expectations. In the Augmented Theatre use case, this requirement was addressed by carefully selecting ASR parameters and caption formatting in such a way as to minimize the response time while maintaining accuracy, while keeping deployment on edge running on a Graphics Processing Unit (GPU)-enabled laptop. Serving the same goal (among others), caption translations were pre-generated, instead of being translated in real time, to avoid any further inference-induced delays and out of respect for the integrity of the theatrical process and experience. This also had the secondary benefit of reducing deployment infrastructure needs during the live performances.

Similar challenges were faced in the Augmented Training use case, where the timing of instructions is crucial for the safety and effectiveness of tasks such as industrial assembly. The difficulties were compounded by related challenges, such as the need for "Providing context to LLM" and the specific management of "How to fine tune LLMs utilising data from the user manual." Integrating rich context and detailed training data increased the computational load. It was observed that, in XR environments, a fundamental trade-off exists between the desired complexity of the AI model (richness of context, high accuracy) and the achievable real-time performance. Future projects operating in time-sensitive, critical operational contexts must define absolute latency thresholds before finalizing the model architecture.

3.1.2. Managing reliability of AI responses in real-world settings

Critical issues emerged regarding model reliability in real-world settings, including how to address the mistakes in AI responses and ensuring the model could understand user intent (questions vs commands). Poor understanding makes the AI responses unreliable or confusing, especially in ambiguous dialogue.

The AR Theatre pilot faced the problem of mistakes in ASR responses due to a variety of reasons: 1) the audio sampling duration was constrained as low as possible to minimize latency as mentioned previously, which in response increased the likelihood of poor transcriptions or very short transcriptions leading to misidentified captions, 2) theatrical speech pushes the boundaries of legibility with non-verbal vocalizations (e.g. screams), 3) multiple simultaneous speakers is a frequent occurrence in theatre, and 4) background noise or audio effects can cause interference. The above characteristics are integral to theatre, but other commonly found issues also apply, for example imposed by microphone quality and wireless reception quality and stability. On another front, NMT generated translations were not on par with literary standards, understandably challenged by text heavy with obscure words, poetic syntax, and idiosyncratic linguistic characteristics. Raw NMT translations of the theatrical captions required substantial editing by a human expert before being approved for consumption by the public audience. These challenges were reflected in the workshop comment as: "the integration of real-time translation and captions do not always perform in the desired literary accuracy."

This experience established that for applications requiring high fidelity (such as cultural ones) or high stakes (such as industrial ones), relying solely on real-time generative models is inherently risky. To tackle this problem, the consortium successfully implemented a hybrid validation architecture, where static, human-validated data served as the primary layer of truth, while real-time AI responses provided enrichment and alignment. This approach is fundamental for the responsible deployment of AI in sensitive domains, where specific sector requirements must prevail over mere technological novelty.

3.1.3. Managing hardware constraints and framework dependencies

The Augmented Theatre deployment faced basic infrastructural limitations, including "Charging equipment, internet connection, application serving computer, microphone, scene limitations." Some of these requirements, such as wi-fi and internet connection or charging plugs, are common-place amenities in urban, office environments, but may still present a challenge in the context of ancient theatrical stages in remote locations or experimental stages in former-industrial complexes. These were addressed by assembling a set of accessories and peripherals for the piloting needs. Other requirements though such as a server for edge deployment of the models had to be addressed with the procurement of a GPU-enabled laptop in the spirit of high mobility, despite performance shortcomings compared to a desktop unit. Finally, the AR glasses themselves, being recently released, state-of-the-art devices, presented constraints of their own, especially regarding the ambient heat in the piloting conditions (Athens, summer) which exacerbated their own heat dissipation system, further draining their battery capacity. Dedicated cooling pads were used to mitigate the issue with charging in between performances/rehearsals, as well as power banks to support longer sessions. The battery constraints and limited autonomy was felt even stronger since it prohibited extensive streaming of the AR feed to a locally connected computer for footage

recording and user experience monitoring – both of which were highly important functions. Software development in the XR field is heavily dependent on specific external infrastructure (connectivity, host platforms, specific hardware such as HoloLens 2 in the Augmented Training). The project demonstrated that reliance on potentially unstable external infrastructural components must be identified and mitigated early. Failure to adequately plan for non-technical infrastructure caused performance issues.

In parallel, the Virtual Conference team reported challenges in developing the conference application regarding choice of framework and technology stack. This application was based on integrating AI-based VOXReality services in an implementation of the open source, self-hosted social platform “Mozilla Hubs”. This design was challenged when faced with the announcement that Mozilla hubs would be discontinued on May 31st, 2024[11], meaning that the public servers would be made unavailable (leaving only self-hosting as a deployment option) and the open-source repository would be managed by a new organization (Hubs Foundation[12]). Although this change didn’t affect the pilot application development technically, it did challenge the exploitation and future uptake plans of the use-case. In view of similar news regarding the closing of the public servers of the 8th wall AR platform[13], the combined lesson is to keep in mind the rapid changes and instability of the commercial services landscape and rely on decentralized, open-source frameworks.

3.1.4. Integrating diverse inputs and handling complex interactions

The Augmented Training pilot struggled with managing complex physical actions, noting that "Handing different type of task/interactions (involving tools) was not very straight forward e.g. Screwdriver interaction" Furthermore, linguistic variability posed a "Diverse accent/dialect challenge". The difficulty in handling tool interactions suggests a gap between semantic understanding (what the user said) and contextual and physical understanding (what the user is doing and the required state change of the physical object). Successful voice-commanded XR system implementation requires extremely high fidelity in the integration of vision systems to track fine motor actions. To resolve this difficulty, future work must focus on developing training datasets that explicitly link complex, multi-phase manipulation actions with voice instructions, bridging the multimodal granularity gap.

3.2. Navigating organizational and collaboration hurdles

Organizational challenges focused on resource allocation, defining shared protocols, and ensuring efficient knowledge flow between partners.

3.2.1. The challenge of accurate resource estimation for pilot execution

The Virtual Conference team reported that the original proposal for the Virtual Conference pilot plan did not consider enough resources. Consequently, the consortium had no other option but to rely on voluntary actions of its members to complete the requirements. They confirmed that "Testing & Piloting" required more resources than foreseen. Research projects often underestimate the cost, time, and human effort required for the crucial validation phase in the field. Relying on "voluntary actions" can challenge the consistency of the process and strain partner commitment. The conclusion is that resources dedicated to validation (time and

[11] https://support.mozilla.org/en-US/kb/end-support-mozilla-hubs
[12] https://hubsfoundation.org/
[13] https://www.8thwall.com/blog/post/200208966730/next-chapter

money) must be substantial and explicitly budgeted. Policy guidelines should encourage allocating a larger, protected resource pool specifically for the complexities of real-world trials in integrated multimodal projects.

3.2.2. Establishing mutual agreement on pilot protocols and parameters

The project suffered "delays for mutual agreement on protocol and parameters for the pilot testing and actual execution". Standardizing testing protocols (e.g., data format, environment configuration, success criteria) across different use case owners and technical teams must happen in the early stages of the project, rather than during the pilot planning and execution phases. Delaying this critical alignment creates system-level bottlenecks that hinder overall project progress. The specifications of the media involved (images, sounds, videos) in components' transactions and training datasets, must be fully agreed and defined in advance until the end of a development phase. A single resolution change in a medium could render days of model training useless if changed, further impacting pilot planning, deployment, and execution timelines.

In VOXReality alignment meetings between domain expert partners and technical partners were held in the beginning of the project, which proved to be beneficial during the design and development of the technical components. Furthermore, the pilot protocol alignment meetings between all stakeholders (use case owners, domain experts, and technical partners) in the consortium for all three use cases were also held. Although the consortium invested in such meetings, establishing mutual agreement on the protocols to align all three use cases, on the requirements of the technical components, and on the evaluation criteria and metrics to be used required more time and effort than anticipated. This impacted the pilot planning and execution and required additional effort on all parties included and additional pre-pilot studies to be conducted to successfully complete the piloting project outcomes.

3.2.3. The need for a dedicated technical architect role per use case

The consortium identified a crucial need for a dedicated software architect for each use case, responsible for overseeing the integration of different technologies and facilitating knowledge transfer between new hires and evolving roles. In complex projects with multiple components and multiple use cases, standard work package or task leadership, as well as scientific and technical coordination, are often insufficient to ensure coherent technical integration.

In VOXReality, the scientific and technical coordinator was responsible for monitoring and ensuring the consistency and quality of both scientific outcomes and technological results. The role encompassed coordinating research activities across the consortium, ensuring alignment between technical partners and use-case domain experts, overseeing technical developments in close collaboration with work package leaders, tracking the progress of project Key Performance Indicators (KPIs), and working with the main project coordinator to bridge administrative and technical requirements. Due to the scope of these responsibilities, the role could not engage in hands-on management of knowledge utilization and integration within each individual use-case software.

Consequently, a dedicated integration architect per use case is necessary to maintain the architectural integrity of the software, prevent knowledge fragmentation, particularly in the context of staff turnover, and ensure seamless interaction among the various system components.

3.2.4. Managing expectations: balancing innovation and exploitation demands

The consortium highlighted a common trade-off in Research and Innovation Action projects: "We need to produce good graphics in order to be competitive in XR market. Although this was never intended to be the aim of the project, reviewers still expected this under the umbrella of future exploitation". This highlights a potential conflict between scientific goals (innovative AI models at a lower Technology Readiness Level (TRL)) and market expectations (polished User Experience (UX) at a higher TRL). It is essential for future proposals to clearly define the TRL scope for user interface elements. If high exploitation is targeted, professional UX and graphics design must be adequately funded alongside research and development (R&D) activities.

3.3. The role of end-user feedback

User feedback provided essential data on usability, engagement thresholds, and audience acceptance.

3.3.1. Balancing difficulty levels of XR experiences to incentivize interaction and engagement

In the Augmented Theatre use case, when the assembly process is spoon-fed to the user, the user completes the task quickly, however the knowledge retention is questionable. To incentivize the user to stay engaged, to ask relevant questions, and to understand the assembly process by heart, it was important to make the tasks difficult to make the user utilize the AI assistant to learn more. The effectiveness of the AI assistant can be maximized when the application logic is designed such that the user has to take help from the assistant, yet without frustrating the user. This indicates that the assistant's intervention logic should potentially adapt dynamically based on user performance, considering interaction frequency as a core metric for measuring engagement success.

3.3.2. Tailoring UI and complexity for mixed audiences

The Augmented Theatre pilot faced the challenge of "Dealing with the balance of needs for mixed audiences: theatre goers who might be alienated with technology and people with affinity to technology who might be interested in theatre". This was resolved through "Continuous testing with both audiences to tailor the user interface (UI)." XR implementations, particularly in cultural or public contexts, cannot rely on a single user interface. The solution requires a layered UX design, where users can choose, or the system can default to, an appropriate level of technical complexity, simplifying adoption for the general public while retaining functional depth for technical users.

In VOXReality, the Augmented Theatre system did not implement multiple user interfaces, or a method to profile the current user's skill and needs and adapt the user interface accordingly. Instead, the piloting opted for step-by-step tutorials designed for a fully inexperienced but open-to-learn persona. This persona did not adequately cover the diversity of participants. Among others, the process of familiarizing with the device, its input methods and learning the application's interactions requires a mental space and mood, which not always aligns with users' expectations from a theatre-going experience.

Appropriately dedicating time, space, and human support to allow users to enter an open-to-learn attitude is very important to facilitate successful onboarding. In parallel, system design

should keep requirements for user interaction intensity and complexity as low as possible.

3.3.3. Considering stakeholder sensitivity: the perception of "AI"

The user feedback raised the question whether the word AI should be avoided. This suggests the existence of a psychological barrier to adoption in end-users. Communication strategies must focus on human-centred benefits (e.g., accessibility, enhanced immersion) rather than the underlying, potentially intimidating, technological terms such as "AI."

In the Augmented Theatre use case, extensive effort was dedicated to clarifying the differences between “popular” generative AI tools and VOXReality’s AI tools for ASR, NMT and VL. This step was essential to increase acceptance of the VOXReality technologies, since incorporating generative AI tools is not an established or acceptable part of the creation and production methodology in theatre. For this use case clear boundaries on the function and impact of the AI technology were created: 1) no AI-outcome should reach the audience unmediated, instead, to preserve , 2) AI tools had to offer automation conveniences to the respective human expert: either the literary translator (NMT) or the theatre’s technical operator (ASR and VL), instead of substituting human labour.

3.4. Key takeaways for future XR/AI projects

This array of challenges and successful adaptations across the VOXReality project provides critical guidance for future initiatives in the XR/AI domain. The key takeaways are:

- **Dedicated Integration Leadership:** Projects of this complexity require dedicated integration/software architects per software/use case to oversee the coherence between different technologies and prevent knowledge gaps, especially with staff turnover.
- **Proactive Resource Allocation:** There must be a clear and substantial allocation of resources (money and time) dedicated specifically to the testing and piloting phases, as these activities are consistently underestimated.
- **Early Protocol Standardization:** Mutual agreement on testing protocols, parameters, and development lifecycles should be achieved in the initial phase to prevent project-wide delays during execution.
- **Human-centred AI Architecture:** For high-stakes or high-fidelity applications, implementing a human in the loop technologies is essential for maintaining reliability and accuracy.
- **Pre-defined Performance Thresholds:** Latency thresholds for real-time communication with AI models must be established before finalizing model architecture to prioritize speed in time-critical operational contexts.
- **Manage Stakeholder Expectations:** Clear boundaries should be set regarding market needs versus core research objectives. This is essential to manage expectations and ensure the correct reception of research prototypes as opposed to commercial services.
- **Personalized Experiences for Higher Engagement:** Prioritizing personalized and adaptive technologies in XR experiences is a successful strategy for increasing adherence in training scenarios and overall user engagement.

4. Best Practices and Recommendations for developing and deploying XR/AI solutions

This section outlines the Best Practices, which refer to strategies, methodologies, or tools that proved to be particularly effective and efficient during the project. They are considered replicable models of excellence that demonstrate both technical soundness and efficient use of resources, making them adaptable by other initiatives. The practices delineated here significantly contributed to VOXReality's success, providing concrete guidance for efficient execution in future XR/AI projects. These best practices are recommendations for other projects related to the same field as the one of VOXReality.

4.1. Effective development methodologies and tools

VOXReality leveraged modern DevOps practices and standardization to ensure consistency, scalability, and technical quality of the products.

4.1.1. Leveraging containerization and modern frameworks

The team made extensive use of Docker and containerization practices for software release and developed using well known and current programming languages and frameworks such as PyTorch and Python. The use of containerization is fundamental for multi-partner projects, as it guarantees that the software runs identically across all environments, drastically reducing integration friction and the time dedicated to resolving dependency issues.

4.1.2. Strategic use of public repository and robust service design

The creation and use of public project repositories such as GitLab, DockerHub and HuggingFace were one of the main practices utilized by the VOXReality team, which aligns with EU Open Science mandates and maximizes the visibility of R&D results. This approach proved essential for the Open Call process, allowing applicants to access and test services hands-on. Complementing this, the adoption of tools such as Swagger and Fast Application Programming Interface (API) for API/model serving ensures a robust and scalable service architecture. This standardization guarantees consistency and ease of integration between partner services (such as connecting ASR output to CA input), effectively mitigating technical communication challenges.

4.1.3. Accommodating flexibility in software development

In collaborations between R&D and creative industries, the technology must be presented early as a modular toolkit or sandbox applications. This allows artistic stakeholders to interact with the system and experiment with its potential before final content design and integration, promoting genuine co-creation rather than forced integration. To reduce the risk of detecting incompatibilities or system limitations later in the design process (during final integrations), a suitable complementary method to mitigate this risk is by creating verticals in key, high-risk/high-innovation areas.

4.1.4. Defining and documenting evaluation frameworks

The evaluation framework required the project to define extraneous variables, detect potential biases and measures to mitigate them, and formalize the use of control/experimental groups. This emphasis on academic rigor in an applied R&D project ensures that claims relating to the effectiveness of the VOXY assistant (Success Story 5.1.2) are scientifically validated,

increasing the transparency and reliability of the results. Furthermore, the framework includes creating a report system to automatically track user performance, which is crucial for objective analysis.

4.1.6 Successful mimicry of known system interactions

Adoption of Familiar Interaction Paradigms reduces user friction by leveraging existing habits. The strategy to mimic interactions that people know from other systems (e.g., 'Hey Voxy' similar to 'Hey Google') proved crucial. By integrating familiar interaction cues into the development phase, the system significantly reduced the cognitive barrier to entry, facilitating the adoption of complex new technology.

4.2. Fostering collaboration within the consortium

Successful collaboration extended beyond digital tools, including strategic in-person interactions and leveraging external expertise.

4.2.1. Strategies for close technical communication

The consortium prioritized close collaboration and communication between technical parties during the development process through dedicated work package meetings, broader scientific and technical meetings, and separate use case meetings attended by both the domain experts and the technical partners. This structure of frequent communication supplements formal reporting, allowing for rapid resolution of technical conflicts before they escalate into integration blockers, especially in a multimodal and interconnected environment.

4.2.2. The role of regular (in-person) general assembly meetings

General Assembly meetings were held in person, which proved to be crucial and beneficial. Face-to-face interaction is indispensable for building personal trust, aligning high-level strategy, and resolving complex organizational issues that are difficult to manage purely via digital platforms.

4.2.3. Monitoring progress with online collaborative tools

Progress was efficiently monitored using online Excel tables and Miro boards which were evaluated and discussed via regular telcos that all partners participate. This simple and transparent method ensures administrative oversight and allows all partners to have a synchronized view of task progress, KPI status, technology readiness, and budget utilization.

4.2.4. Leveraging external validation via open call projects

The team recognized that the partnerships with the external parties from the Open Calls enabled easy procurement of opinion on VOXReality technologies and rapid response to their requirements for improvement of the components released. Open Call Third Parties serve as high-value external reviewers, providing early market input and robust feedback on technical performance, stability, and security, essential before broader deployment.

4.3. Engaging stakeholders and end-users

High user engagement and adoption were achieved through human-centred design principles.

4.3.1. Adopting a co-creation and human-centred approach

VOXReality methodology to conducting R&D included co-creation and human-centred approach through constant stakeholder engagement, from focus groups to gather requirements and to multiple user studies to gather feedback. This continuous feedback loop ensures that the technologies are relevant and closer to end-user utilization.

4.3.2. The value of interactive tutorials and contextual suggestions

The efficient user onboarding was ensured through interactive tutorials and user engagement was maintained by providing contextual suggestions to ambiguous user intent. The interactive tutorials were developed to be presented to the users at the beginning of each use case application to ensure user interaction through the experience could be maximized, while contextual suggestions were used to mitigate LLM inaccuracies and intent misinterpretation (Lesson Learned 3.1.2), guiding the user back on track and improving the perceived intelligence and reliability of the system.

4.4. Dissemination and Communication strategies

The VOXReality results were disseminated as soon as they are validated through experiments, peer-review, software tests, or pilot studies. The communication of the project process and its results proved to be beneficial for the project's outcomes.

4.4.1. Leveraging the power of collective network and social media

The team managed to attract a lot of interest via the use of social media (LinkedIn and X) and XR events (Immersive Tech Week, EACL 2024, IHIET AI 2023 Conference), which highlights power of collective network and exposure to different events and people. Active and impactful digital dissemination is essential to maximize project visibility and connect with industry end-users beyond academic or governmental circles.

4.4.2. The importance of disseminating completed results and demonstrations

A key practice was to disseminate completed results and demonstrations through scientific articles, open-source libraries, demo events, and social media as soon as they are ready. Each individual scientific result was published in open-access peer-reviewed conferences and journals to reach the scientific community. The datasets, models, source code that are associated with the VOXReality results were published online for all types of stakeholders for immediate exploitation. It is evident that a fully polished and working demonstration (model, software, or experience) offers the most compelling proof of technical readiness and exploitation potential to potential partners, researchers, developers, or customers.

4.5. Ensuring Sustainability and future relevance

Ensuring the long-term viability and relevance of the project's outputs was a central Best Practice, achieved through strategic choices regarding technology stack, documentation, and external engagement:

- **Open and Standardized Infrastructure:** the creation and use of public project repositories (GitLab, DockerHub, HuggingFace) and the use of current, well-known programming languages and frameworks (PyTorch, Python, Docker) ensure the continuity and reusability of the project's models and software.

- **Robust Documentation and APIs:** utilizing common documentation practices and API/model serving tools such as Swagger and FastAPI guarantees technical interoperability and ease of integration, securing the project's legacy.
- **External Validation Loop:** proactively asking Open Call Third Party partners (OC TPs) for feedback on technology improvements ensures that the outcomes meet external market standards and are robust for future commercial exploitation.
- **Future Collaboration:** fostering close partnerships with OC TPs, even outside the original project scope, creates a network of collaborators committed to the technology's evolution and future joint projects.
- **Demonstration Quality:** prioritizing the dissemination of a completed demonstration showcases high TRL, which is crucial for attracting external interest and ensuring future market uptake.

5. Success Stories: demonstrating impact and innovation

This chapter documents the Success Stories, which are concrete achievements that produced a significant and demonstrable long-term impact. They show the added value of EU funding, generate visibility, and inspire replication, and are usually linked to outcomes that reached institutional recognition or broad dissemination. These achievements highlight VOXReality's innovation, quantifiable impact, and potential for long-term sustainability.

5.1. Key achievements of the VOXReality project

The project produced innovative technical results validated in real-world environments.

5.1.1. Development of novel AI models and optimization strategies for XR

VOXReality had two main goals: advancing natural language processing literature and ensuring these technologies are compatible with XR applications. 16 scientific papers were published. 25 models and 6 datasets were released publicly. 3 optimization tools were developed to convert the models to lighter and efficient versions for XR deployment. The capabilities of these models were demonstrated through three use case applications in addition to the five applications built by the TPs. The pilot studies conducted by the consortium and TPs provided evidence that these achievements are not just academically sound, but also practical and valuable to be used in real-world XR scenarios.

5.1.2. Demonstration of a clear performance advantage of voice assistants

The Augmented Training pilot studies demonstrated a clear advantage of the VOXY assistant compared to the menu assistant. This result validates the project's central hypothesis, that the convergence of language-based and vision-based AI technologies enhances the user experience, providing superior efficiency and ease of use compared to traditional XR interaction methods (controllers and hand gestures).

5.1.3. Empowering unexperienced users and improving training effectiveness

In the Augmented Training, unexperienced users managed the training session with the help of VOXY. This result validates the Augmented Training application as an effective tool for skills development, successfully lowering the entry barrier for complex industrial assembly tasks and increasing training accessibility. Furthermore, it was observed that the training object tooltips enhanced the interactions with the system, fostering engagement.

5.1.4. Use of VOXReality models by Third Parties for product improvement

TPs used the VOXReality models to improve their products with potential substantial social impact. This is clear evidence of early and direct exploitation of the project outcomes. The models did not remain academic prototypes but were integrated into commercial products, accelerating the impact of the research on the market.

5.1.5. Addressing language and geographical barriers to facilitate business expansion

The Virtual Conference use case successfully demonstrated that VOXReality outcomes allows final (business) users to expand their business making activities and tackles language and geographical barriers for business making. The project directly contributed to the EU policy

goal of a digital single market, demonstrating that XR/AI technology can effectively facilitate cross-border business interaction, regardless of participants' physical location or native language.

5.1.6. Higher social impact: benefits for people with disabilities and language learners

Although not aimed at the beginning of the project, people living with disabilities enjoyed and benefited from the Augmented Theatre experience. This result demonstrates the powerful, sometimes fortuitous, social benefits arising from technological innovation. Designing for cultural immersion (AR subtitles) simultaneously created a high-value accessibility tool, indicating that multimodal XR research should proactively explore accessibility benefits, even if not primary objectives. Additionally, people learning the language of the play benefited from the immersion to a cultural production, highlighting a positive impact on language education.

5.2. User engagement and testimonials

Project reached a good amount of stakeholders both for technology validation and for potential future exploitation.

5.2.1. High User Interest and Willingness to Engage

Users reported high engagement during the pilot studies and interest for participation in future immersive experiences. This validates the project's strategic choice to focus on combining language and visual modalities, confirming that this convergence is a desirable path for future XR development. For example, the AR theatre use-case returned high user feedback on a sample of 38 participants regarding behavioural intention, recommendation, enjoyment and perceived value. In specific, 86.8% of the participants stated they would attend future Augmented Theatre performances, rating the experience on average with a 4.47/5 on a Likert scale. This level of user engagement and interest clearly demonstrated the potential of this use-case for future development.

5.2.2. Generating economic impact: job creation

The project generated a measurable economic impact by creating 13 new jobs within Open Call Third Party organizations. This provides a measurable and quantifiable return on the EU investment, demonstrating that R&D projects can directly stimulate job creation in high-technology sectors.

5.2.3. Securing future partnerships and external interest

The consortium members anticipate future collaborations with the Open Calls Third Parties in future projects. The Augmented Theatre solution was also adopted by the Artistic Directorate of Athens Epidaurus Festival (AEF), which is the "actual client" of this solution, encouraging adoption by other organizations. Furthermore, the performance “Hippolytus (in the arms of Aphrodite)" which was developed for the Augmented Theatre use case was accepted in the Re-Fest CultureHub's[14] annual festival that that focuses on creative technologies with the support of AEF. The performance will be supported by the software system that was created for the pilot study, on manual control, without dependence of VOXReality systems. AEF will

[14] https://www.culturehub.org/re-fest

further support endeavours to showcase the output of the use case, recognising the value it presents for cultural organisations. This high level of market pull and the confirmed intention of subsequent collaborations ensure the project's sustainability and its ability to continue generating value well beyond the funding period.

5.2.4. Potential for increased visibility on CORDIS

The consortium considers the results to be suitable for further promotion, where potential interactions with both previous and future EU funded projects could be explored. Leveraging CORDIS maximizes dissemination, fosters synergy with related projects, and ensures that the project's findings are easily discoverable by the wider EU research community. This document also contributes towards this direction.

5.3. Contribution to EU objectives and the XR/AI field

VOXReality delivered strong, measurable contributions aligned with European strategic goals and advanced the state-of-the-art in multimodal XR technologies:

- **Economic Growth and Job Creation:** the Open Calls for Third Parties resulted in the creation of 13 jobs across all use cases, demonstrating a tangible economic return on investment and stimulation of the high-tech sector.
- **Digital Single Market Facilitation:** by successfully tackling language and geographical barriers for business-making in the Virtual Conference pilot, VOXReality directly supported the EU objective of fostering a unified and accessible digital single market.
- **Responsible AI Governance:** the establishment of an Ethical Approval Pipeline by the entire consortium sets a high standard for the responsible deployment of multimodal AI systems, addressing critical EU concerns regarding data privacy and bias.
- **Social Impact and Accessibility:** despite not being a primary objective, the AR Theatre solution provided significant and unforeseen social benefits for people living with disabilities and language learners, demonstrating that XR/AI innovation can inherently drive greater accessibility and cultural inclusion.
- **Knowledge Transfer and Visibility:** the project's results are found to be suitable for promotion on platforms such as CORDIS, suggesting potential interactions with previous and future EU-funded projects and ensuring broad dissemination of the knowledge generated to the wider research community.

6. Insights from the VOXReality Open Call Projects

The VOXReality Open Call was a key strategic initiative designed to amplify the project's impact by extending its cutting-edge multimodal AI and XR technologies to external partners. The initiative successfully funded five third party organizations, who integrated VOXReality models into novel solutions across critical European sectors, including manufacturing, healthcare, and education. The collective experiences of these partners provide a rich repository of practical knowledge, validating the framework's real-world applicability. These projects are:

- **AI-Assisted XR Collaboration and Training with Multilingual Support – AIXTRA:** VR Training Tool focused on VR Training, developing an authoring tool to build multilingual VR applications with AI-based virtual training partners.
- **CrossSense – CS:** Cognitive Support via Smart Glasses explored Healthcare by using smart glasses to aid memory recall for individuals with Dementia, promoting independence.
- **Voice-Activated Augmented Reality Heritage Tours – VAARHeT:** The Voice-Activated AR Heritage Tours applied AI to Cultural Heritage, creating an immersive, voice-activated AR experience for an archaeological park.
- **Enhanced-XR Human-Robot Interactive Collaboration for Welding Operations – WELD-E:** Voice-Driven Industrial Automation targeted industrial settings, building an integrated voice and vision-based AI system for remote expert support and monitoring in robotic welding.
- **XR Career Guidance and Training Assistant – XR-CareerAssist:** Personalized Career Development in XR developed an immersive XR Platform that utilizes the AI models for personalized career path planning and development.

6.1. AI-Assisted XR Collaboration and Training with Multilingual Support - AIXTRA: VR Training Tool

AIXTRA delivered an authoring tool providing capabilities for building VR applications that overcome language barriers, adding AI-based virtual training partners.

6.1.1. Lessons Learned

The team learned that for a tool intended for broader use, the underlying AI components must be highly flexible, and commercial success hinges on strong marketing and realistic planning for regulatory approval timelines. The lessons learned by the team are the following:

- **AI Testing is Time-Consuming:** the validation of AI components requires a significant time investment.
- **Need for API/Component Flexibility:** the initial APIs required modifications and greater flexibility to implement diverse use cases.
- **Dissemination is Essential:** marketing and dissemination activities proved more critical for market engagement than initially expected.
- **Ethics Approval Timeframe:** gaining ethics approval as a company (in Germany) was a long process.

6.1.2. Best Practices

AIXTRA's recommendations were rooted on making the solution technically accessible (documentation, consistent interfaces) while using robust development methodologies (Agile)

to ensure high-quality, reusable code. The best practices suggested by the team are the following:

- **Providing Comprehensive Documentation:** offering code, usage examples, and how-to guides was crucial for developers and end-users.
- **Consistency in Architecture and Interfaces:** maintaining architectural consistency enabled easier integration and use of components.
- **Tool-Supported Agile Development:** agile practices supported by collaboration tools (Jira, Slack) worked well in uncertain project environments.
- **Single Members Deep-Dive:** allowing team members to deep-dive into technical aspects and share insights across the team improved efficiency.

6.1.3. Success Stories

AIXTRA proved that the VOXReality components can be packaged into an easy-to-use authoring tool, making advanced AI/XR training accessible to a wider enterprise market. The success stories reported by the team are the following:

- **Granting Non-Developer Access:** successfully enabling non-developers to access AI capabilities through the authoring tool, broadening the potential user base.
- **Collaboration with Mentor and Consortium:** improved results and enhanced the efficiency of marketing activities.
- **Dockerization of Backend:** allowed flexible and stable deployment of the complex backend system.
- **Innovative Use Case:** demonstrated a compelling, multi-user, voice-chat enabled case to engage with enterprise clients.

6.2. CrossSense – CS: Cognitive Support via Smart Glasses

CrossSense utilized smart glasses to empower people living with Dementia and Mild Cognitive Impairment, focusing on supporting their ability to recall information for independent living.

6.2.1. Lessons Learned

The team learned that legal/ethical processes are a major bottleneck and that effective XR solutions for vulnerable groups should prioritize utility and reliability over feature complexity. The lessons learned by the team are the following:

- **Time-Consuming Ethical Approval:** the ethical approval process took longer than expected.
- **Focus on Core Features:** users indicated that smart glasses only need a few, highly effective features to be useful, driving the team to integrate sparingly.
- **Interpreting User Intent:** accurately interpreting user intent required a fusion of sensory inputs (e.g. visual context) and in-app usage data (e.g. user history), not just pure voice commands.
- **Hardware Variability:** innovative hardware often varies across devices, even with interoperable standards, requiring custom adaptation.

6.2.2. Best Practices

The main suggestions of CrossSense include ethically driven use-case design (person-centred design) and ensuring the highly sensitive edge device remain stable and performant through test early, integrate sparingly practice. The best practices suggested by the team are the following:

- **Person-Centred Design Approach:** co-design workshops were crucial, treating the end-users as experts on their own experience.
- **Test Early and Integrate Sparingly:** testing various models early and ensuring only necessary AI models were loaded to prevent overload.
- **Relevant Partnerships:** collaboration with support organizations (e.g., Leading Lives) granted early access to target users for refining the roadmap.

6.2.3. Success Stories

The project proved that the fusion of VOXReality components can unlock entirely new, valuable capabilities (visual captioning) with a direct, positive impact on vulnerable user groups. The success stories reported by the team are the following:

- **Successful Integration of All Models:** all foreseen VOXReality models were successfully integrated.
- **Unlocking Visual Captioning:** the combination of spatial and semantic data allowed the project to introduce a new capability: visual captioning of the user's history.
- **Significantly Reduced Overall Latency:** latency was improved by using the VOXReality streaming ASR service and asynchronous task management.
- **Transformative Potential:** great enthusiasm from users confirmed the project's potential to significantly impact the lives of people with cognitive impairment.

6.3. Voice-Activated Augmented Reality Heritage Tours – VAARHeT: The immersive heritage tours

The Voice-Activated Augmented Reality Heritage Tours (VAARHET) project developed an AI-powered AR application to create an immersive, transformative visitor experience at the Āraiši ezerpils Archaeological Parks.

6.3.1. Lessons Learned

The team learned that success relies heavily on optimizing the early project phases, specifically by ensuring technical readiness and fostering open communication across partners from the outset since the project contained co-applicant teams. The lessons learned by the team are the following:

- **Encouraging Cooperation:** greater cooperation and exchange between teams should be encouraged to avoid information silos.
- **Design Time Investment:** the user-centred approach requires a bit more time in the design phase to capture needs appropriately.

6.3.2. Best Practices

VAARHET's recommendations focus on leveraging the partnership with the end-user to ensure relevance and utilizing public dissemination to generate further collaboration opportunities. The best practices suggested by the team are the following:

- **Collaboration with End-User Partner:** partnership with the Cultural Heritage Institution (CHI) was crucial for success and ensuring post-project applicability.
- **Regular Short-Form Dissemination:** disseminating project activities regularly (social media, events) helped to create greater collaboration opportunities and public interest.
- **User-Centred Design Focus:** guaranteeing the applicability of the technology through a user-centred design approach.

6.3.3. Success Stories

The project demonstrated that the technology could solve practical challenges for cultural institutions and serve as a showcase for European innovation and cooperation on an international level. The success stories reported by the team are the following:

- **Multilingual Voice Activation:** successfully providing multilingual voice-activated technology, highly useful for museums operating in minority languages.
- **Diplomatic Recognition:** the project was presented to the Irish ambassador in Latvia, who shared it as an example of European cooperation.
- **New Cooperation Opportunities:** successfully opened new avenues for AI cooperation in cultural heritage, including forming a Horizon CL2 consortium.

6.4. Enhanced-XR Human-Robot Interactive Collaboration for Welding Operations – WELD-E: Voice-Driven Industrial Automation

WELD-E implemented a voice and vision-based AI system within an XR environment to facilitate remote support and monitoring for robowelders in harsh industrial settings.

6.4.1. Lessons Learned

The primary lesson for WELD-E centred on deploying multimodal AI in a non-controlled industrial setting requires rigorous planning for hardware constraints, sound isolation, and model fine-tuning to overcome real-world physical friction. The lessons learned by the team are the following:

- **Noise Interference in Industrial Settings:** the high level of ambient noise critically degrades audio input quality.
- **Initial Language Model Shortcomings:** the ASR, NMT, and VL models required refinement due to domain-specific jargon.
- **Resource Bottlenecks on Legacy Hardware:** the complexity of AI models was constrained by the limited capacity of older hardware.
- **Tool Fragility and System Raggedness:** systems must be designed to withstand severe environmental factors (Dust, Heat, Accidents).

6.4.2. Best Practices

WELD-E's suggestions focus on creating a flexible system architecture supported by iterative development and multi-sensory feedback crucial for safety in high-risk environments. The best practices suggested by the team are the following:

- **Modular Voice2Command Pipeline:** developing a modular AI pipeline that allowed independent optimization and supported extensibility.
- **Agile Management and Iterative Sprints:** continuously testing and improving the solution to quickly address industrial bugs.
- **Combined Audio-Visual Alerts:** integrating text-to-speech (TTS) and visual cues in the XR interface to improve usability and safety compliance.
- **Proactive Communication with the Mentor:** essential for resolving technical issues quickly and clarifying requirements.

6.4.3. Success Stories

The project proved that multimodal AI could achieve high technical readiness in demanding environments, delivering tangible time-saving benefits (annotation pipeline) and receiving high praise from domain experts. The success stories reported by the team are the following:

- **Live Demo in Harsh Industrial Environment:** successfully validating the system at TRL7, confirming its industrial readiness.
- **Auto-annotation Pipeline:** reduced the time required to prepare training data for YOLO by 60%.
- **High Impact Report:** expert welders reported the high impact of the solution with the potential to supersede traditional methods.

6.5. XR Career Guidance and Training Assistant – XR-CareerAssist: Personalized Career Development in XR

XR-CareerAssist developed an immersive XR platform utilizing VOXReality's AI to analyse CVs and provide personalized career paths and development plans.

6.5.1. Lessons Learned

The team learned that integrating advanced AI requires accurate initial sizing of GPU resources and that general-purpose AI models must be precisely tuned to understand and explain domain-specific data visualizations. The lessons learned by the team are the following:

- **Underestimated GPU Requirements for Inference:** the platform required an upgrade to 64GB VRAM GPUs to handle the AI workload efficiently.
- **Domain-Specific VL Model Training:** the VL model required further training to accurately explain complex career data and Sankey diagrams.
- **Avatar Compatibility and UI Clarity:** initial avatars lacked sufficient cross-device compatibility, and UI design needed clearer navigation cues.
- **Need for Real-Time Transcription:** qualitative user feedback showed that real-time text transcription alongside audio outputs was necessary for full accessibility.

6.5.2. Best Practices

XR-CareerAssist's recommendations is to embed validation early and continuously through structured pilots and user-centred design within an Agile framework to ensure the platform's performance and educational relevance. The best practices suggested by the team are the following:

- **Structured Pilot Sessions:** using immediate post-session questionnaires and open interviews to structure data relevance and improve KPI alignment.
- **Agile Methodology and Early User Involvement:** ensuring platform robustness and accurate multilingual performance evaluation from the outset.
- **Regular Stakeholder Engagement:** continuous alignment with user needs through consistent interaction with the University of Exeter and other educational bodies.

6.5.3. Success Stories

The project proved that the platform is ready for large-scale deployment in education and that the AI stack can be successfully tailored to provide intelligent insights on complex, domain-specific data. The success stories reported by the team are the following:

- **Successful Validation at the University of Exeter:** confirmed technical readiness and scalability potential in university educational settings.
- **Database Restructuring and Query Optimization:** significant improvement in performance, reducing API response times.
- **VL Model Trained on Sankey Diagrams:** successfully adapting a general-purpose model to interpret and inform on specialized data charts.

Common Thread: the success stories validate Educational Scalability and Customized Intelligence. Common Findings and Suggestions for Third Parties

6.6. Common Findings and Recommendations across all TP teams

The experiences of the five Third Parties, while diverse in sector, revealed powerful, interconnected threads that define the challenges and opportunities of the next generation of multimodal XR systems.

6.6.1. Lessons Learned

The dominant challenge across all TPs was on integration. This challenge manifested in three key areas:

- **Domain-Specific AI Adaptation:** the assumption that "plug-and-play" AI models would be sufficent was proved to be wrong. XR-CareerAssist and WELD-E demonstrated that core models require mandatory domain-specific fine-tuning to handle jargon, noise, or unique data formats.
- **Infrastructure and Environmental Robustness:** legacy hardware limitations and environmental factors (noise, heat) are not trivial; they dictate the performance ceiling. WELD-E, XR-CareerAssist showed that this underscores the need for early, comprehensive infrastructure planning alongside development.
- **Process and Regulatory Overheads:** TPs consistently noted that Ethical Approval and AI Testing consume more time than anticipated. Future projects must allocate a realistic timeline for these critical processes.

6.6.2. Best Practices

The strategies that led to success were universally rooted in Agile principles, user-centric approach, and transparent development. Specifically:

- **Methodological Agility:** the adoption of Agile and iterative sprints (WELD-E, XR-CareerAssist, AIXTRA) provided the necessary speed and flexibility to overcome technical hurdles inherent in cutting-edge R&D.
- **User-Driven Validation:** the mandatory alignment with end-users, whether through person-centred design (CrossSense) or structured pilot sessions (XR-CareerAssist), ensured that development efforts remained focused on practical, impactful features.
- **Architectural Discipline:** practices such as developing modular pipelines (WELD-E) and focusing on consistency and documentation (AIXTRA) proved vital for creating stable, maintainable, and deployable systems that are easy for external parties to integrate.

6.6.3. Success Stories

The combined achievements of the TPs validate VOXReality's impact on scalable and accessible XR/AI Solutions.

- **Validation of Scalability and TRL:** projects demonstrated the technology's readiness across various contexts, from achieving TRL7 in an industrial environment (WELD-E) to platform scalability in education (XR-CareerAssist), confirming the framework's reliability.
- **Multimodal Synergy and New Capabilities:** the integration of voice and vision led to genuinely new, impactful capabilities like visual captioning (CrossSense) and the democratization of AI via authoring tools (AIXTRA).
- **Societal and Cultural Impact:** the most compelling success lies in the transformative potential demonstrated in healthcare (CS) and the successful use of the technology to preserve and disseminate culture in minority languages (VAARHET), positioning multimodal XR as a tool for social good and European collaboration.

7. Conclusion

This guide is not intended as a simple report of activities, but as a practical tool to transfer the knowledge gained through the challenges and successes of the VOXReality project. Through a participatory methodological approach involving the entire consortium and third parties, we have extrapolated a set of Lessons Learned, Best Practices, and Success Stories essential for anyone operating at the intersection of Extended Reality (XR) and Artificial Intelligence (AI).

The experience of the three pilot projects, Virtual Conference, Augmented Theatre, and Augmented Training, highlighted that integrating complex multimodal models (ASR, NMT, VL, CA) into real-time XR environments requires precise engineering. A crucial lesson learned was managing the trade-off between AI model complexity and real-time performance: for critical operational contexts, it is fundamental to define absolute latency thresholds before finalizing the architecture. Furthermore, to ensure reliability in high-fidelity applications such as theatre or industry, the adoption of human-centred approaches, where AI models are enriched by human-validated data, proved to be the winning strategy to mitigate the errors of AI models.

To ensure the scalability and reproducibility of results, VOXReality adopted rigorous DevOps practices, such as containerization (Docker) and the use of public repositories (HuggingFace, GitLab), aligning with EU Open Science mandates. The user-centric approach and the use of interactive tutorials proved essential for lowering cognitive entry barriers and increasing user engagement.

The results confirmed the project's central thesis: voice assistants offer a clear performance advantage over traditional menu-based interactions, improving efficiency and ease of use. An unexpected and significant result was the social impact: solutions designed for cultural immersion proved to be powerful accessibility tools for people with disabilities, demonstrating how XR/AI innovation can inherently drive inclusion. Moreover, the success of the Open Calls validated the technology in diverse sectors, from industrial automation to healthcare, leading to the creation of 13 new jobs and demonstrating a tangible economic return.

In summary, VOXReality leaves a legacy not only of advanced AI models but of a validated framework of architectural and organizational practices, which hopefully will serve as an accelerator for the next generation of European projects.

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.