Human Interaction and Emerging Technologies (IHIET 2023), Vol. 70, 2023, 139–148
https://doi.org/10.54941/ahfe1002938

VOXReality: Immersive XR Experiences Combining Language and Vision AI Models

Apostolos Maniatis, Stavroula Bourou, Zacharias Anastasakis, and Kostantinos Psychogyios

Synelixis Solutions S.A., Farmakidou 10, Chalkida 34100, Greece

ABSTRACT

In recent years, Artificial Intelligence (AI) technology has seen significant growth due to advancements in machine learning (ML) and data processing, as well as the availability of large amounts of data. The integration of AI with eXtended Reality (XR) technologies such as Virtual Reality (VR) and Augmented Reality (AR) can create innovative solutions and provide intuitive interactions and immersive experiences across various sectors, including education, entertainment and healthcare. The presented paper describes the innovative Voice-drive interaction in XR spaces (VOXReality)* initiative, funded by the European commission, that integrates language and vision-based AI with unidirectional or bidirectional exchanges to drive AR and VR, allowing for natural human interactions with XR systems and creating multi-modal XR experiences. It aligns Natural Language Processing (NLP) and Computer Vision (CV) parallel progress to design novel models and techniques that integrate language and visual understanding with XR, providing a holistic understanding of goals, environment, and context. VOXReality plans to validate its visionary approaches through three use cases such as a XR personal assistant, real-time verbal communication in virtual conferences, and immersive experience for the audience of theatrical plays.

Keywords: Artificial intelligence, Multimodal artificial intelligence, Extended reality, Human-artificial intelligence interaction

INTRODUCTION

Artificial Intelligence (AI) is a rapidly growing field that involves the development of systems capable of performing tasks which typically require human intelligence, such as visual understating, speech recognition, and language translation. The use of AI is becoming integral part of our everyday lives, and it has the potential to greatly improve the way we live and work. From virtual assistants to self-driving cars, AI is changing the way we interact with technology and the world around us. With the increasing availability of big data and the development of powerful computational resources, AI is constantly evolving and will continue to grow in the coming years. The use of AI has the potential to transform many industries, bringing about significant changes to society. Specifically, the authors in (Yu, et al., 2018) presents the use of AI in healthcare, while the work at (Zhai, et al., 2021) provides a review of AI in finance and education respectively.

Just as AI technology is rapidly advancing, so the Extended Reality (XR) technology is also progressing. XR, which includes virtual reality (VR) and augmented reality (AR), has the potential to change the way we consume media, conduct business, and interact with the world. In recent years, XR technology has made significant advance leading to its use in variety of industries, such as gaming (Tao, et al., 2021) and education (Alnagrat, et al., 2022). The XR technologies are also used for employee training in various domains, like manufacturing (Doolani, et al., 2020). Moreover, study presented at (Gerup, et al., 2020) shows that AR applications outperformed traditional learning methods in the acquisition of anatomy knowledge and needle insertion skills, while the work conducted at (Cha, et al., 2012) demonstrates that VR technology could be a valuable tool for firefighter training.

As both AI and XR continue to evolve, the integration of those two technologies can provide exciting new possibilities. Specifically, this combination has the potential to revolutionize the way we interact with the world as well as to enable the creation of immersive and interactive experiences. AI methods can be used to analyze large amount of data, such as images and speech, to extract meaningful information. This information can then be integrated into XR experiences to provide users with relevant and contextually aware content.

Focusing on the integration of AI models into XR space, we present the Voice-drive interaction in XR spaces (VOXReality) system, which is funded by the European commission. Specifically, VOXReality aligns the parallel progress in Natural Language Processing (NLP) and Computer Vision (CV), which are two colliding fields of AI, to design and develop novel models that merge language and visual understanding with XR, creating multi-modal and immersive XR experiences. Those models can be exploited to address challenges related to human-to-human and human-to-machine interaction. The effectiveness of the framework can be demonstrated through three use cases: XR personal assistant, communication in virtual conferences, and XR experiences for theatrical plays.

VOXReality Vision

VOXReality integrates language- and vision-based AI models, aiming to tackle the challenges associated with human-to-human and human-to-machine interaction in XR space. When it comes to human-to-human interaction challenges, the most common is the language barriers between people who speak different languages or use heavy accent. In this case, people may struggle to communicate effectively, leading to misunderstandings and miscommunications. Regarding the human-to-machine interaction challenges, the usability of complex machine can be overwhelming for some users, resulting to frustration and lack of engagement.

To address the aforementioned challenges, VOXReality utilizes NLP and CV advancements to create robust AI models. The multi-modal information is exchanged between the modalities with unidirectional or bidirectional ways to drive AR and VR, enabling natural human interactions with the

backends of XR systems and creating multi-modal XR experiences from the combination of vision and sound information. Those next-generation models can provide holistic understanding of our goals, surrounding environment and context, while they can improve both human-to-machine and human-to-human XR experiences.

Particularly, powerful multi-tasking AI models that are adjustable to different languages, expressions, and accents, while they are able to consider the surrounding context, are implemented and integrated in user-centric multilingual translation system. Additionally, based on language and visual information, visually grounded language models are built providing valuable information about the surrounding. Finally, by integrating all the aforementioned models and additional knowledge, a context-aware multi-modal agent is developed, which creates well-grounded conversations, provides navigation guidelines and assistance to the user via XR.

By exploiting the VOXReality framework, the language barriers in human-to-human communication can be overcome by providing real-time translation, enabling people to communicate effectively even if they do not use the same language. Simultaneously, VOXReality allows people to easily and user-friendly interact with machine using natural language rather than complex interfaces, addressing with this way the human-to-machine interaction challenge of usability.

Additionally, the developed AI models can be used to create immersive XR experience. Specifically, the language-based AI models can be used to understand natural language commands as well as questions and respond accordingly, allowing people to interact with XR environment in an intuitive way. The vision-based AI models are able to analyze real-world environments, resulting to the generation of accurate XR experiences. Generally, the VOXReality project is set to push the boundaries of what is possible in XR by utilizing AI and bring us one step closer to the future where the physical and digital worlds merge to offer a truly immersive experience.

The framework is aimed to be deployed and validated in the following three use cases. It should be mentioned that the communication can be done in the user's language for each use case.

- **Personal assistant in XR**: The VOXReality framework can be exploited to create robust personal assistant in XR space for various applications, such as instruction assistant, technical support, and navigation guides. The communication can be performed in the user's language.
- **Communication in virtual conferences:** The proposed system can be utilized to improve the real-time communication during conferences in VR. Additionally, it automate the navigation in those spaces with information provided by the agent.
- **XR experiences for theatrical plays**: The VOXReality can create immersive XR experience for a theatre audience. Specifically, with the help of language translation and user-specific audio-visual connections, VOXReality is able to improve theatre experiences.

VOXReality AI Components

This section focuses on the development of the VOXReality's AI components. Each component can be developed to address specific challenges, with the ultimate goal of creating a user-centric, adaptable AI system that can process and understand a wide range of sensory information. The main intention is the evolution of the current state of AI technology by creating intelligent systems that can handle complex tasks such as speech recognition, language translation, and visual grounded language understanding.

Automatic Speech Recognition

Automatic speech recognition (ASR) is a technology that enables computers to recognize and transcribe spoken language in real-time. It's an area of AI that applies Machine Learning (ML) algorithms to recognize patterns in speech and convert them into written text.

Over the past few years, ASR technology has advanced dramatically, driven by improvements in ML algorithms. There are ASR systems using multi-models and end-to-end models. Multi-model approaches aim to solve the problem using multiple models, which are designed to solve sub-tasks (related to the problem) and the targeted task. The use of DNN-Hidden Markov Model (DNN-HMMs) combined with a Language Model obtains impressive results on LibriSpeech (Panayotov, et al., 2015) test-other subset (Lüscher, et al., 2019). In end-to-end approaches, only supervised methods can be used to solve the speech tasks. Authors at (Kim, et al., 2019) present a model based on the Encoder-Decoder architecture using stacked Long Short-Term Memory (LSTM) networks for the encoder and LSTM combined with soft attention for the decoder.

VOXReality intends to create ASR model that is capable of converting spoken words into text supporting five different languages including English, German, Greek, Spanish, and Italian. With the aim of achieving high accuracy and fluency, VOXReality plans to leverage the power of Transformer-based models. These models have proven to be particularly effective in handling special expressions, allowing them to produce text outputs that closely resemble natural language. Additionally, Transformer-based ASR models have shown significant improvements in performance compared to traditional ASR models, particularly in noisy environments where speech recognition accuracy can be challenging. By utilizing these advanced models, VOXReality hopes to provide accurate and reliable speech recognition capabilities for a wide range of applications.

Multilingual Translation

Multilingual Neural Machine Translation (NMT) systems, such as Google's multilingual NMT system (Johnson, et al., 2017), can translate between different languages using a single model, allowing information sharing among related languages. The use of generative pretraining and other techniques, such as multitask learning and pre-training different components of the model, have been suggested to reduce the need for large amounts of paired data (Garcia, et al., 2020).

NMT faces challenges due to the adaptive and generative nature of human language. While progress has been made, semantic contexts, particularly surrounding sentences, need to be considered. Multilingual translation and unsupervised NMT, especially for low-resource languages, are still challenging. Integrating multiple modalities, like images and videos, is also an open issue. Adapting to non-native speech with accents and technical vocabulary in conference presentations, specific theater languages, and audience reading speed and contractions are additional challenges.

The VOXReality project aims to address the challenges faced in NMT and develop multilingual NMT models for high and low-resource languages that can perform task-related translations of high quality. With the focus on creating universal solutions, VOXReality intends to utilize transfer learning and fine-tuning techniques as well as exploring structured knowledge to customize general language models to specific scenarios, such as technical conferences, with the ability to handle an impressive five different languages including English, German, Greek, Spanish, and Italian. The result can be AI models that are capable of adapting to various languages, expressions, and accents while taking also visual context into account and can be conditioned with external information.

Visual Grounded Language Models

VOXReality seeks to develop Visual Grounded Language Models that connect language semantics with visual concepts through spatio-temporal information like videos and 3D representations, rather than static images. This approach considers multiple views of the same objects or scenes, as well as spatial relationships, through the use of structure-from-motion and/or multi-view stereo techniques. The project also intends to incorporate colorful proxies, (Yao, et al., 2021) which could include spatial information and scene graph descriptions. Recently, the field of visual-language processing has seen significant advancements with the emergence of cutting-edge models such as LXMERT (Tan & Bansal, 2019) and UNITER (Chen, et al., 2020), which are considered prominent representatives in the field.

The goal of VOXReality is to develop a vision-language model that captures both spatial and semantic relationships, which is crucial for advancing language-driven XR technology with improved spatial understanding. The outcome is the construction of visual grounded language models that can be used in spatio-temporal tasks like navigation and guidance.

Generative Dialogue System

A Generative Dialogue System is an AI system that is capable of generating human-like responses based on user input, while it can be used as a component of a multi-modal agent to enhance its conversational capabilities. The Generative Dialogue System operates as a task-oriented system with the primary objective of assisting users in completing specific tasks, such as customer service, booking, and virtual assistant services. Task-oriented dialogue systems consist of three individual components. Firstly, the Natural

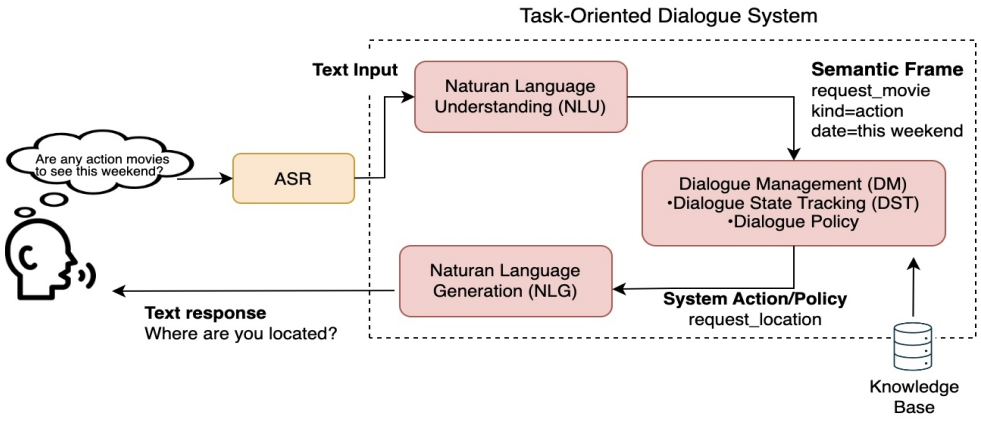

Figure 1: Task-oriented Dialogue system.

Language Understanding (NLU) is responsible for converting the transcribed text from ASR into a machine-readable format, including slot filling and intent detection. Then, the Dialogue Management (DM) system, that consists of Dialogue State Tracking (DST) and Dialogue Policy, is responsible for managing the flow of the conversation and generating appropriate responses. DM can be connected to external knowledge sources, such as a knowledge base, to provide more meaningful answers. Finally, the Natural Language Generation (NLG) can be described as the transformation of a meaning representation into a human-understandable text. Figure 1 presents the Task-oriented dialogue system.

The approach of using pretrained models has been gaining popularity in the field of language generation. MASS, a masked sequence to sequence pretraining method, is one example of this approach that has shown promising results in language generation (Song, et al., 2019). Another method is the unified text-to-text transformer, T5, which has demonstrated its effectiveness in transfer learning (Raffel, et al., 2020). Also, SimpleTOD (Hosseini-Asl, et al., 2020) approach for task-oriented dialogue, which leverages pre-trained models like GPT-2 for transfer learning in the open-domain setting where data is more readily available.

VOXReality's Generative Dialogue System can handle a wide range of tasks, from answering general questions about the environment to providing directions and instructions. With the ability of perceive and process information about the user's surroundings, the developed system could respond in real-time to user requests and provide relevant, personalized assistance. The 3D environments and landmarks plan to be used to train the system to recognize patterns and relationships in the data, allowing it to make informed decisions and respond effectively to user requests. The end goal is to create a highly intuitive and user-friendly system that provides unparalleled levels of convenience and accessibility.

VOXReality System Design

All the AI components described in the previous section are combined to create the VOXReality system. This section explains the way that these

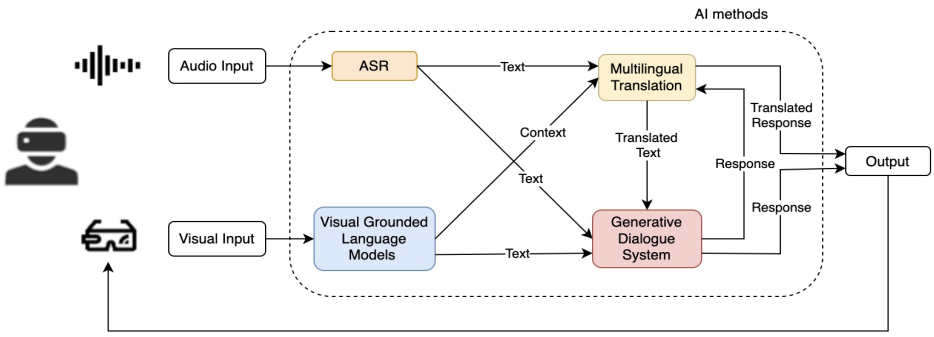

Figure 2: Information flow of VOXReality.

components are linked together as well as how the final output is produced. Figure 2 highlights the interaction between the various components of the VOXReality system, showcasing the flow of information and collaboration between them.

In VOXReality, the combination of two distinct inputs, audio and visual, offers a more complete and accurate representation of the surrounding environment. The audio input captures the user's vocalizations, including any questions and requests, that the user makes to the system. Additionally, the visual input represents the current visual stimuli that the user is experiencing. The input audio is processed and transcribed into written text by the ASR component. Then the transcribed text can be used by the other NLP models. Similarly, Visual Grounded Language Model interprets the meaning of visual inputs and provide text representations.

When translation is the objective, the Multilingual Translation system incorporates both the transcribed text from the ASR component as well as the visual context from Visual Grounded Language component. As a result, the system delivers an accurate translated response to the user. In the case of user's communication with the agent, the system converts the audio and visual inputs into text by the ASR and Visual Grounded Language components respectively. The Generative Dialogue System is responsible to process this multi-modal information and produce meaningful response to the user. Since the Generative Dialogue System is functional for English language, any audio input that is in different language is translated first to English. The generated response can be translated into one of 5 languages by the Multilingual Translation component, before being presented to the user. Finally, the output of AI models can be displayed to the user by using techniques such as AR and VR, enhancing the user's experience of the real world or even creating an entirely new one.

VOXReality Visionary Use Cases

VOXReality is a cutting-edge AI system that is aimed at revolutionizing various aspects of XR technology. The system intents to enhance communication, networking, and cultural experiences through advanced AI models

that understand both audio and visual elements in the environment. This section covers three different use cases that VOXReality aims to address.

Personal Assistant in XR

The first use case scenario that VOXReality plans to build is the personal assistant in XR. Advanced AI models that understand both audio and visual elements of the environment enable language commands and semantic context, both spatially and semantically. This opens a world of new possibilities, such as instructional assistants, technical support with Head-Mounted Display (HMD) devices, and navigation guides. The ultimate goal is to use the cutting-edge vision and language models that can perceive the user's surroundings and guide them to complete tasks through spatially and semantically grounded commands, addressing with this way challenges related to human-to-machine communication.

In comparison to existing image and voice-activated personal assistants, that only treat input modalities as separate entities and do not consider their spatial relationships, VOXReality takes advantage of all the information available to the user. Therefore, VOXReality is expected to provide users with a more comprehensive and intuitive experience. This innovative approach to personal assistants has the potential to revolutionize the way people interact with technology, providing more natural and efficient communication that is tailored to the user's needs.

Communication in Virtual Conferences

The VOXReality system can be utilized to enhance spoken communication within virtual environments. The main advantage of attending conferences and events is the opportunity to network with others who share similar interests. This is the reason why most people attend these types of gatherings. However, successful networking relies on effective communication and the creation of common ground. This can be a challenge in virtual settings as they lack the subtle communication cues that help establish a connection in face-to-face interactions.

VOXReality is going to adopt the developed ASR and multilingual NMT models to enhance real-time verbal communication in VR space. Additionally, the navigation in those spaces can be automated by deployment the Generative Dialogue System of multi-modal agent to the virtual environment, aiming to increase the human-to-machine conversational realism.

XR Experiences for Theatrical Plays

Attending theatre is a wonderful cultural experience with many positive affects to people, such as improved mental health, exposure to different cultures, enhanced creativity, etc. Hundreds of years, theatres have been using many different effects to create a immersive experience for the audience. Already, the theatre industry investigates the use of pioneering technologies, such as VR, AR and XR, to modify the way that the audience immerse in the live

experiences. However, there are several challenges that theatres face in attracting and retaining audiences. The most common one is the language barriers for people going to the theatre, particularly for those attending performances in a language they are not familiar with.

VOXReality creates immersive experience for the audience during theatrical performances, while it can overcome the language barriers challenge. This can be achieved by using the multilingual NMT models and displaying live captions, subtitles, and context-specific information through XR headsets. The captions can be available in the viewer's preferred language, allowing a comprehensive understanding of the play. Also, theatre producers can utilized the ASR models of VOXReality as well as data-driven techniques to enhance their performances with visual effects triggered by predetermined expressions or words. Therefore, VOXReality offers exciting opportunities for theaters to enhance the audience experience, from creating immersive performances to expanding accessibility.

CONCLUSION

VOXReality combines the advancements in NLP and CV to create multimodal AI models that incorporate language and visual understanding of the environment. Those AI models are integrated with XR, leading to immersive XR experiences and various innovative applications across industries. The main components of VOXReality system are Automatic Speech Recognition, Multilingual Translation, Visual Grounded Language Models, and Generative Dialogue Systems of a context-aware multi-modal agent, which interact with each other to take advantage of audio and visual inputs, while produce comprehensive and accurate knowledge for the user. The proposed system can tackle issues and challenges related to interactions between humans and machines, as well as interactions between humans themselves. The functionalities and the added value of the developed system are demonstrated through three use cases: XR personal assistant, communication in virtual conferences, and XR experiences for theatrical plays.

ACKNOWLEDGEMENT

VOXReality is funded by the European Union under grant agreement No. 101070521. 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. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

Alnagrat, A., Ismail, R. C., Idrus, S. Z. S. & Alfaqi, R. M. A., 2022. A Review of Extended Reality (XR) Technologies in the Future of Human Education: Current Trend and Future Opportunity. *Journal of Human Centered Technology,* pp. 81–96.

Cha, M., Han, S., Lee, J. & Choi, B., 2012. A virtual reality based fire training simulator integrated with fire dynamics data. *Fire safety journal,* pp. 12–24.

Chen, Y.-C.et al., 2020. Uniter: Universal image-text representation learning. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*. s.l.:Springer International Publishing., pp. 104–120.

Doolani, S. et al., 2020. A review of extended reality (xr) technologies for manufacturing training. *Technologies,* p. 77.

Garcia, X., Siddhant, A., Firat, O. & Parikh, A. P., 2020. Harnessing multilinguality in unsupervised machine translation for rare languages. *arXiv preprint arXiv:2009.11201*.

Gerup, J., Soerensen, C. B. & Dieckmann, P., 2020. Augmented reality and mixed reality for healthcare education beyond surgery: an integrative review. *International journal of medical education,* p. 1.

Hosseini-Asl, E. et al., 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems,* pp. 20179–20191.

Johnson, M. et al., 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics,* pp. 339–351.

Kim, C., Shin, M., Garg, A. & Gowda, D., 2019. Improved Vocal Tract Length Perturbation for a State-of-the-Art End-to-End Speech Recognition System.. In: *Interspeech*. s.l.:s.n., pp. 739–743.

Lüscher, C. et al., 2019. RWTH ASR Systems for LibriSpeech: Hybrid vs Attention–w/o Data Augmentation. *arXiv preprint arXiv:1905.03072*.

Panayotov, V., Chen, G., Povey, D. & Khudanpur, S., 2015. Librispeech: an asr corpus based on public domain audio books. In: *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. s.l.:s.n., pp. 5206–5210.

Raffel, C. et al., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research,* pp. 5485–5551.

Song, K. et al., 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Tan, H. & Bansal, M., 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Tao, G. et al., 2021. Immersive virtual reality health games: a narrative review of game design. *Journal of NeuroEngineering and Rehabilitation,* pp. 1–21.

Yao, Y. et al., 2021. Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*.

Yu, K.-H., Beam, A. L. & Kohane, I. S., 2018. Artificial intelligence in healthcare. *Nature biomedical engineering,* pp. 719–731.

Zhai, X. et al., 2021. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity,* pp. 1–18.