Using Augmented Reality and Machine Learning for Captioning in Theatrical Experiences

Olga Chatzifoti[1], Nick Achilleopoulos[1], Spiros Borotis[1], Moonisa Ahsan[2], and Eleni Oikonomou [3]

[1] *Maggioli Spa, Research & Development Department, Athens, Greece*
[2] *Centrum Wiskunde en Informatica (CWI), Amsterdam, The Netherlands*
[3] *Athens-Epidaurus Festival, Athens, Greece*

Corresponding author: olga.chatzifoti@maggioli.gr

Keywords: Augmented Reality, Machine Learning, Theatre, Cultural heritage, Captions

Introduction

In this work, we present an Augmented Reality Theatre System, exemplifying the benefits of using new technologies for enhancing audience engagement and directors' creative practice. Our work aims to improve accessibility for an international and diverse audience through personalized captions and to creatively support a new generation of directors through hybrid multimedia. To achieve that, we design and develop a system using augmented reality (AR) and machine learning (ML) technologies. In this paper, we present the personalized captions feature of the system. This work has been undertaken in the context of VOXReality (www.voxreality.eu), a European Union-funded Research and Innovation Action.

Motivation

Our goal is to make the theatre-going experience more accessible to a wider audience through the provision of accurate, real-time, customizable captions, translated into the viewer's preferred language. This is an especially relevant need for cultural works in languages with small speaker communities, like Greek, and even more important for historical works, such as Ancient Greek plays. To support this goal, we opted to use AR for content delivery following literature recommendations, and ML for the generation, synchronization and translations of the captions to provide automation as well as extended language support over the current practice.

Related Work & Background

The established practice for theatre caption delivery has two formats: open captions, displayed on large screens or projected on stage, and closed captions, delivered on viewers' smartphones or dedicated devices. Closed captions offer improved accessibility because they can support customization options. Closed captions delivered in AR can offer an improved user experience by allowing the user to focus on a single visual frame, potentially reducing mental workload (Rzayev et al., 2018, 2020). Related work in AR captioning for theatre includes the SmartSubs[1] project by Institute of Informatics & Telecommunications (IIT) at NCSR Demokritos, and the Thearto Project[2] which also investigated audiovisual effects in the performance. The Royal National Theatre in London, UK has already established a monocular AR closed captions system as part of their public accessibility initiative[3]. Industry solutions that support ML generated AR captions, like XRAI[4], are also appearing. This work aims to improve closed AR captions for theatrical performances by investigating optimized ML workflows and extended user customization options for AR display.

System Description

Our system uses a server-client architecture adopting a secure WebSocket protocol for communication between server-client and a secure REST API for communication with the ML models, developed in VOXReality (Maniatis et al., 2023). The server directs the audio signal from the stage's microphones to the ML model and distributes the text responses to the XR clients in the audience for AR display. The ML model (Issam et al., 2024; Maka et al., 2024) uses a transformer architecture and performs audio transcription and text matching to provide pre-formatted captions, ensuring the least latency and best accuracy, in addition to executing translation demands. All VOXReality ML models can be found available online, promoting Open Science[5].

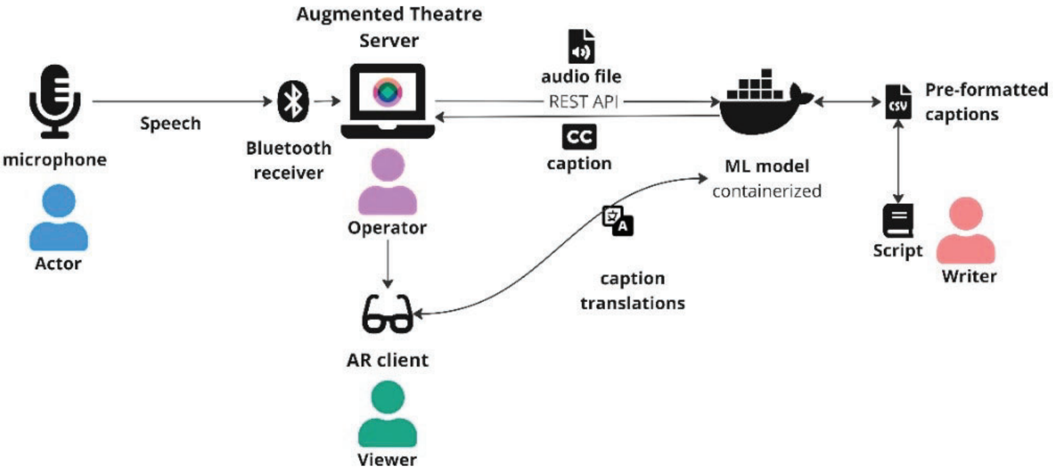

Figure 31. System Architecture for AR Theatre.

Figure 31 describes the system architecture. The AR application allows the user to customize the caption display extensively for improved readability and additionally displays audiovisual effects, as designed and orchestrated by the theatrical director. Figure 32 presents screenshots of the AR application user interface for caption customization and placeholder AVFXs that were used during evaluation.

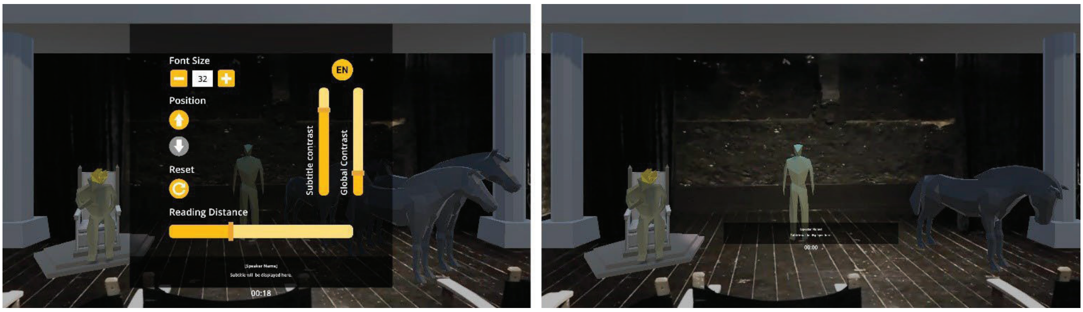

Figure 32. Left: screenshot from AR customization UI. Right: screenshot from caption display

Evaluation

The first user evaluation was completed in May 2024 in Athens, Greece with 12 users recruited through email invitations with users attending a live performance of an excerpt of the Ancient Greek play 'Hippolytus' by Euripides. The focus was on estimating users' behavioural intention toward this system. Alongside that, other parameters of usability (System Usability Scale (Lewis, 2018), NASA Task Load Index (Hart, 1986) and Simulator Sickness Questionnaire (Kennedy et al., 1993)) were evaluated, and a semi-structured interview in pairs of participants was performed to capture more nuanced feedback. Application data logs were used to contextualize user responses and were complemented by facilitator observations. Questionnaire results were analysed using statistical analysis and the interview transcripts using thematic analysis (Clarke & Braun, 2017). The second user evaluation will be delivered in May 2025 and will be advertised for user recruitment on public communication channels.

Results & Discussion

Despite being at an early technical and aesthetic level, the initial evaluation was decisively positive with users stating that they would be interested in attending this form of theatre in real conditions and that they saw practical benefit and artistic merit in the provided features. Negative feedback was focused on the technical performance of the system and the learning curve of the AR application. Table 4 presents a summary of key results focusing on the custom Behavioural Intention Questions - the extended results of the full evaluation are beyond the scope of this paper. Redesign for pilot 2 will address the initial user feedback and provide a renewed evaluation.

Table 4. Pilot 1 results on Behavioural Intention Questions

Question	Mean	Standard Deviation	Median
Recommend AR theatre	4.3	1	5
Prefer dynamic subtitles	3.3	0.9	3.5
Prefer static subtitles	3.8	1	4
Would like more VFX	4.5	0.7	5
General Experience rating	4.3	0.5	4

Conclusion

This work showcases the potential of augmented reality and machine learning to address the needs of diverse audiences and offer a more inclusive augmented theatre experience. The initial evaluation of our system demonstrated promising user engagement and interest, while also revealing areas for technical improvement. The redesign of the system for the next pilot will aim for improvements in latency, accuracy and translation quality for the ML components, and a more user-friendly experience for the AR components. As we move toward the next iteration of user testing, outcomes are expected to advise theatres' guidelines for performances of international appeal, provide tested technical solutions for wider adoption, and future research recommendations for blending cultural heritage with cutting-edge technology.

Acknowledgement

VOXReality (www.voxreality.eu) is funded by the European Union under grant Agreement No. 101070521. Views and opinions expressed are, however, those of the authors 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. More details: https://cordis.europa.eu/project/id/101070521

References

Clarke, V., & Braun, V. (2017). Thematic analysis. J. Posit. Psychol. 12(3), 297–298. https://doi.org/10.1080/17439760.2016.1262613

Hart, S. G. (1986, January). NASA Task Load Index (TLX). https://ntrs.nasa.gov/citations/20000021487

Issam, A., Can Semerci, Y., Scholtes, J., & Spanakis, G. (2024). Fixed and Adaptive Simultaneous Machine Translation Strategies Using Adapters. In E. Salesky, M. Federico, & M. Carpuat (Eds.), Proc. 21st Int. Conf. Spoken Language Translation (IWSLT 2024) (pp. 298–310). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.iwslt-1.36

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. Int. J. Aviation Psychol. 3(3), 203–220. https://doi.org/10.1207/s15327108ijap0303_3

Lewis, J. (2018). The System Usability Scale: Past, Present, and Future. Int. J. Human-Computer Interaction, 1–14. https://doi.org/10.1080/10447318.2018.1455307

Maka, P., Semerci, Y., Scholtes, J., & Spanakis, G. (2024). Sequence Shortening for Context-Aware Machine Translation. In Y. Graham & M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024 (pp. 1874–1894). Assoc. Comput. Ling. https://aclanthology.org/2024.findings-eacl.127

Maniatis, A., Bourou, S., Anastasakis, Z., & Psychogios, K. (2023). VOXReality: Immersive XR Experiences Combining Language and Vision AI Models. Human Interaction and Emerging Technologies (IHIET-AI 2023): Artificial Intelligence and Future Applications, 70(70).

Rzayev, R., Hartl, S., Wittmann, V., Schwind, V., & Henze, N. (2020). Effects of position of real-time translation on AR glasses. Proceedings of Mensch Und Computer 2020, 251–257. https://doi.org/10.1145/3404983.3405523

Rzayev, R., Woźniak, P. W., Dingler, T., & Henze, N. (2018). Reading on Smart Glasses: The Effect of Text Position, Presentation Type and Walking. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–9. https://doi.org/10.1145/3173574.3173619