Journal of Audiovisual Translation

VR 360º Subtitles: Designing a Test Suite with Eye-Tracking Technology

Marta Brescia-Zapata✉
Universitat Autònoma de Barcelona

Krzysztof Krejtz✉
University of Social Sciences and Humanities SWPS

Pilar Orero✉
Universitat Autònoma de Barcelona

Andrew T. Duchowski✉
Clemson University

Chris J. Hughes✉
University of Salford

Abstract

Subtitle production is an increasingly creative accessibility service. New technologies mean subtitles can be placed at any location on the screen in a variety of formats, shapes, typography, font size, and colour. The screen now allows for accessible creativity, with subtitles able to provide novel experiences beyond those offered by traditional language translation. Immersive environments multiply 2D subtitle features to produce new creative viewing modalities. Testing subtitles in eXtended Reality (XR) has expanded existing methods to address user needs and enjoyment of audiovisual content in 360º viewing displays. After an overview of existing subtitle features in XR, the article describes the challenges of generating subtitle stimuli to test meaningful user viewing behaviours, based on eye-tracking technology. The approach for the first experimental setup for implementing creative subtitles in XR using eye-tracking is outlined in line with novel research questions. The choices made regarding sound, duration and storyboard are described.

Citation: Brescia Zapata, M., Krejtz, K., Orero, P., Duchowski, A.T. & Hughes C.J. (2022). VR 360º Subtitles: Designing a Test Suite with Eye-Tracking Technology. *Journal of Audiovisual Translation,* 5(2), 233–258. https://doi.org/10.47476/jat.v5i2.2022.184

Editor(s): M. Carroll & A. Remael

Received: November 16, 2021

Accepted: July 6, 2022

Published: December 21, 2022

✉ marta.brescia@uab.cat, https://orcid.org/0000-0002-2465-0126
✉ kkrejtz@swps.edu.pl, https://orcid.org/0000-0002-9558-3039
✉ pilar.orero@uab.cat, https://orcid.org/0000-0003-0269-1936
✉ aduchow@g.clemson.edu, https://orcid.org/0000-0003-1681-7878
✉ c.j.hughes@salford.ac.uk, https://orcid.org/0000-0002-4468-6660

Acknowledgements: This study is part of the article-based PhD thesis of Marta Brescia-Zapata in the Department of Translation and Interpreting at Universitat Autònoma de Barcelona (UAB) within the PhD program in Translation and Intercultural Studies. This research has been partially funded by the H2020 projects TRACTION (under Grant Agreement 870610) and MEDIAVERSE (under Gran Agreement 957252). The Commission's support for this publication does not constitute an endorsement of the contents, which reflects the views of the authors only, and the Commission cannot be held responsible for any use which may be made of the information contained herein. Marta Brescia-Zapata and Pilar Orero are members of TransMedia Catalonia, an SGR research group funded by "Secretaria d'Universitats I Recerca del Departament d'Empresa. Coneixement de la Generalitat de Catalunya" (2021 SGR 00077).

Conclusions show that testing subtitles in immersive media environments is both a linguistic and an artistic endeavour, which requires an agile framework fostering contrast and comparison of different functionalities. Results of the present, preliminary study shed light on the possibilities for future experimental setups with eye-tracking.

Key words: subtitles, immersive environments, 360º videos, testing, eye-tracking.

1. Introduction

Immersive media technologies such as Virtual Reality (VR) and 360º videos are increasingly prevalent in society. Their potential has placed them in the spotlight of the scientific community for research and education. Industry has also adopted them not only in the entertainment sector, but also for communication, arts, and culture, which has attracted more and mixed audiences (Montagud et al., 2020). At present, these technologies are gaining popularity very fast due to the COVID-19 crisis as they enable interactive, hyper-personalised, and engaging experiences anytime and anywhere. Moreover, 360º videos, also known as immersive or VR360 videos, are a cheap and effective way to provide VR experiences. Specialised multi-camera equipment that can capture a 360º or 180º Field of View (FoV) instead of the limited viewpoint of a standard video recording, is used to produce content. VR360 videos can be enjoyed both via traditional devices (PC, laptops, smartphones) or VR devices (Head-Mounted Displays). They can also be consumed as a CAVE (Cave Automatic Virtual Environment), which uses high-resolution projection screens to deliver 360◦ visual experiences.

Immersive environments (in eXtended Reality, or XR) are generally used as an umbrella term referring to hardware, software, methods, and experience in Augmented Reality (AR) or VR or in general Mixed Reality (MR). The main goal of any immersive content is to make people believe that they are physically present (Slater & Wilbur, 1997). According to Rupp et al. (2016, p. 2108), VR360 videos can allow for "highly immersive experiences that activate a sense of presence that engages the user and allows them to focus on the video's content by making the user feel as if he or she is physically a part of the environment". Immersive videos, however, can also produce negative effects such as motion or simulator sickness, possibly turning people away from VR as a medium (Smith, 2015).

Like any other type of media content, 360º media experiences should be accessible. In almost all media assets, accessibility is added as an afterthought during the postproduction phase, despite many voices asking for accessibility in the creation process (Mével, 2020; Romero-Fresco, 2013). For this research we focus on subtitling, where standardised practices have emerged (Matamala & Orero, 2018), rather than covering different accessibility services. In 2D subtitles, the main aspects to consider are position, character identification, speed, number of lines, and number of characters (Bartoll, 2004; Díaz-Cintas & Remael, 2007; Gottlieb, 1995). Nevertheless, some Audiovisual Translation (AVT) studies have challenged traditional subtitling practices, encouraging more creative and integrated subtitles (Foerster, 2010; Fox, 2018; McClarty, 2012, 2014). The production of creative subtitles requires technology since such subtitles may change any of their paratextual features like the font or size or colour, but also where they are positioned, and more so in immersive environments where 2D features do not apply (Hughes et al., 2015; Lee et al., 2007). The integration of subtitles in XR is yet to be defined, and multiple challenges have emerged. Subtitles should be generated "in an immersive, engaging, emotive and aesthetically pleasing way" (Brown et al., 2017, p. 1), always considering accessibility and usability.

Beyond the challenge of subtitle text creation, XR requires direction to the sound source, as it may be outside the current audience viewpoint. Guiding and readability require the subtitler to preview

and tweak formal aspects (Hughes & Montagud, 2020; Orero et al., 2020). This has led to the design of a new, web-based, prototyped framework that generates subtitles in 360º videos. The present article aims to identify how to display such subtitles for an optimal viewing experience. The framework allows for methods used in existing solutions (Brown & Patterson, 2017; Montagud et al., 2019; Rothe et al., 2018) to be easily contrasted and compared, as well as for the quick implementation of new ideas for user testing. After an overview on subtitle features in XR, the article describes the challenges of generating subtitle stimuli to test meaningful user viewing behaviours, based on eye-tracking technology. The approach for the first experimental set up for implementing creative subtitles in XR using eye-tracking is presented, in line with the stated research questions.

2. An Overview of Subtitles in Immersive Environments

Even though XR media was first introduced in the world of videogames, thanks to the development of 360º recording equipment these technologies are now expanding to videos (Hughes et al., 2020a). There are a few significant differences between content created within 2D and 3D environments. 2D means that the content is rendered in two dimensions (flat), while 3D content has depth and volume which allows a rich visual experience. According to Skult and Smed (2020, p. 451), "the key challenge for XR is that the FoV is limited, and the interactor cannot pay attention to the entire virtual scenery at once." The immersive experience, as in real life, moves from passive to active with the user becoming the centre of the story "creating a greater emotional nexus" (Cantero de Julián et al., 2020, p. 418). In a play or opera, the action takes place on the proscenium. However, another activity somewhere in the theatre may distract from that narrative, such as the noise of a lady unwrapping sweets two rows away. The audience in VR has freedom of movement. They can also determine the time spent in any area of interest or field of vision and decide on where to focus their attention. This freedom affects subtitle reading since the development of the narrative may be random, decided by the viewer. Similarly, in VR, the aim is for immersiveness and the concepts of presence and engagement are central, with the ultimate goal of being a witness to the narrative from a first-person viewpoint. This breaks with the concept of a passive audience that reads subtitles, following what Jenkins et al. (2015) define as *spreadable* reading, meaning that the audience spreads its attention across the image as the linear narrative is displayed. In VR images and sound surround, there is no linear narrative and passive viewing moves towards interaction or *drillable* viewing as in video games or transmedia products. In this context, Mittell (2009) explains that "spreadable media encourages horizontal ripples, accumulating eyeballs without necessarily encouraging more long-term engagement. Drillable media typically engage far fewer people but occupy more of their time and energies in a vertical descent into a text's complexities." These features are theoretical principles that have yet to be tested.

The value of VR lies in its potential to tamper with both time and space; hence the experience relies on the viewer. This has a direct effect on the way subtitles are consumed. A person may be watching one part of the scene while there is a person speaking away from the viewing field. A hearing person may be able to locate the sound source but someone with hearing loss will need to be guided.

Another feature that is different from 2D resides in the way media is accessed. There are two options: using CAVE (the naked eye) or a device such as a Head-Mounted Display (HMD). This is a type of display device or monitor that is worn over the head and allows the user to be immersed in whatever experience the display is meant for. The 360º environment accessed when wearing the HMD may be an animation (such as a video game or an animated movie) or live action (such as a movie or a documentary). Depending on the type of media, the content will be installed on a PC or on the HMD itself or stored on the cloud. As Internet speed improves, media content streamed from the web is becoming increasingly popular.

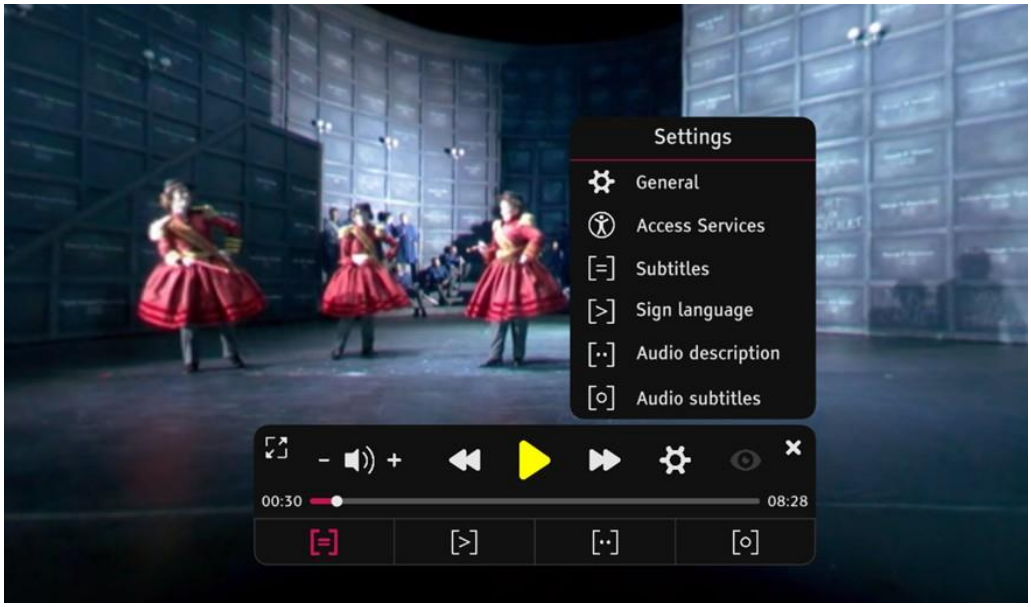
As in traditional 2D media, to create and consume subtitles in XR, a subtitle editor and a subtitle player are needed. Although immersive video players offering the ability to play VR360 video are commercially available, not many of them support accessible services (Brescia-Zapata, 2022). The player needs to be accessible, and the user has to activate the display accessibility. The interface or menu also needs to display the choice of accessibility services available, and finally, the interaction with the terminal or device also needs to be accessible. All these features show the complex ecosystem required for a true XR accessible experience. This, linked to the lack of standardised solutions and guidelines, has led to the development of non-unified solutions, meeting only specific requirements (Hughes & Montagud, 2020). The majority of players seem to have inherited features from the traditional 2D world, instead of addressing the specific features of 360º environments.

This situation served as an inspiration for initiatives like the European H2020 funded Immersive Accessibility (ImAc) project[1] that explored how accessibility services and assistive technologies can be efficiently integrated with immersive media, focusing on VR360 video and spatial audio. Under the umbrella of this project, both an accessible player and a subtitle editor were developed. The accessibility-enabled 360º ImAc player supports audio description, audio subtitles, and sign language along with other features (Montagud et al., 2019) as can be seen in Figure 1.

[1] http://www.imac-project.eu

Figure 1

ImAc Player Settings

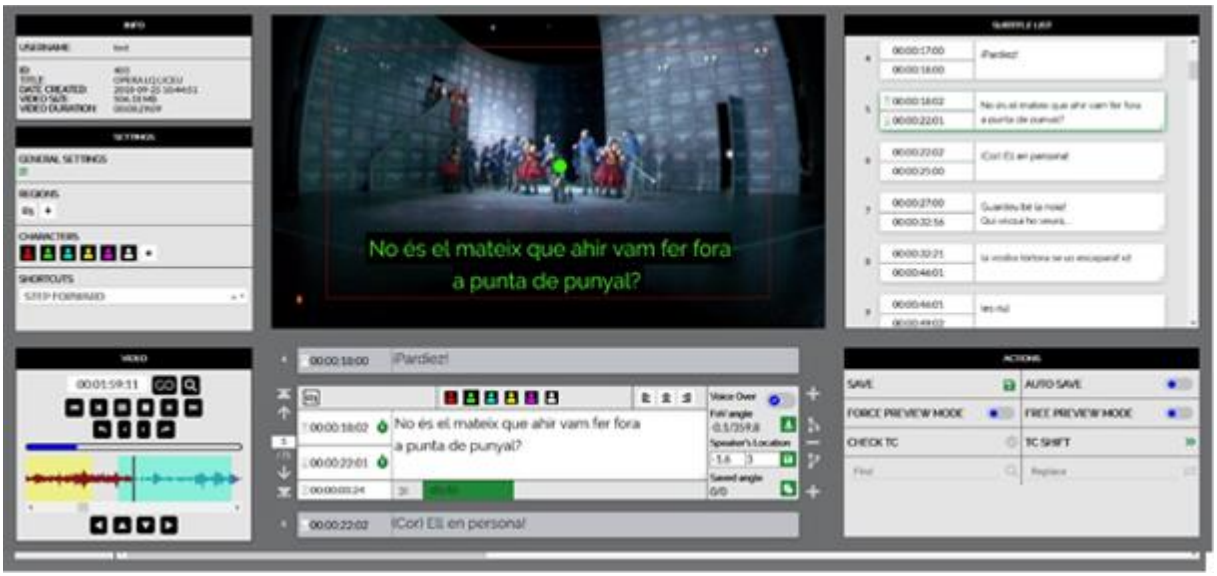


Source: ImAc player screenshot.

In contrast, the ImAc subtitle editor is a commercial web-based editor, and its interface is similar to that of any traditional subtitle editor, as can be seen in Figure 2. The main innovations are related to the FoV in VR360 video, i.e., the extent of observable environment the user is able to see at any given moment. It includes navigation buttons for FoV in spherical space to move up, down, left, and right. There is also a button which moves the FoV to the angle where the speaker of the current subtitle is located. The editor also allows the FoV angle to be changed using the navigation buttons in the video control area or moving the mouse with the left button over the video. By default, the video initially has the current angle as longitude: 0.00º and latitude: 0.00º. In addition, the voiceover option can be marked when there is no speaker in the 360º scene.

Figure 2

Immersive Subtitle Editor Developed in ImAc



Source: ImAc subtitle editor screenshot.

The basic tools to create and consume accessible VR content are now commercially available, e.g., a VR subtitle editor and a VR subtitle player. What is evident is that unless different display modes can be produced, they cannot be tested, and this is one of the shortcomings of the ImAc project which was concluded recently and focused on traditional subtitles projected on immersive environments (Hughes et al., 2020b).

2.1. Related Work

Excluding works that have added (sub)titles at post-editing stages, only three recent studies have focused on investigating subtitles in immersive environments. All the studies followed a user-centric methodology and chose people with hearing loss for testing. Reading skill was not considered within the demographic data.

The British Broadcasting Corporation (BBC) was one of the first research organisations to design subtitles in XR (Brown & Patterson, 2017). The BBC research team first identified the main challenges when developing subtitles for immersive content and, based on these, the following four solutions for subtitle rendering were developed by Brown et al. (2017):

- Evenly spaced: subtitles are placed into the scene in three fixed positions, equally spaced by 120° around the video and slightly below the eye line;

- Follow head immediately: the subtitles are presented as a "head-up display" always in front of you, and slightly below straight ahead. As you turn your head, the subtitle moves with you, always at the same location in the headset display;
- Follow head with lag: the subtitles follow head direction, but only for larger head movements: if you look slightly left or right the subtitle stays in place, but a head movement of a greater amplitude will cause the subtitle to catch up with your head orientation;
- Appear in front, then fixed: each subtitle is placed in the scene in the direction you are looking at the time when it appears and remains fixed in that location in the scene until it disappears.

These four rendering modes were tested with several clips (Brown, 2017), and users reported that while it was easy to locate the evenly spaced captions, they preferred the head-locked options (see Table 1). Head-locked subtitles resemble most traditional ecstatic 2D subtitles, always visible at the bottom of the screen. These results come as no surprise since for years now subtitle testing in Europe has shown that people like what they are used to, even if the data demonstrate that the solution is not ideal, as proved with eye-tracking tests (Mas Manchón & Orero, 2018; Romero-Fresco, 2015).

Table 1

Numbers of People (and Percentages) Who Selected Each Behaviour as Their Favourite or Least Favourite Behaviour. Least Favourite Was Not Specifically Requested, so Was Not Available for All Participants

Behaviour	Favourite	Least favourite
Evenly spaced	1 (4%)	5 (38%)
Follow head immediately	10.5 (44%)	3 (23%)
Follow with lag	7 (29%)	2 (15%)
Appear in front, then fixed	5.5 (23%)	3 (23%)

Source: Authors' own elaboration based on Brown (2017).

The second study (Rothe et al., 2018) compared the two presentation modes: fixed and head-locked subtitles. Although no conclusive results were found, in terms of comfort (i.e., presence, VR sickness, and task load) fixed subtitles led to slightly better results even though fixed captions in general mean that users may not always be able to see the caption as it may be outside their FoV.

The third study, performed under the umbrella of the H2020-funded ImAc project (Hughes et al., 2019), revealed the need to guide users to the sound source of the subtitle (i.e., a sound effect or a character speaking or not speaking). To facilitate this requirement, location within the 3D space information was added to each subtitle (Agulló & Matamala, 2019). This allowed for different modes to be developed which could guide the user to where the speaker was located (Agulló et al., 2019). However, this had the drawback that the location was only specified once per caption, and if a person was moving dynamically, this could affect the exactness of the guiding feature (Hughes et al., 2019).

Nevertheless, the ImAc project designed and developed several guiding mechanisms, and test results showed two preferred methods:

- ImAc Arrow: an arrow positioned left or right directs the user to the target;
- ImAc Radar: a radar circle is shown in the user's view. This identifies both the position of the caption and the relative viewing angle of the user.

In the area of standardisation, a W3C Community Group[2] is focusing on developing new standards for immersive subtitles. They have recently conducted a community survey to gather opinions, but no tests were performed. A small group of users with different hearing levels (Deaf, Hard of Hearing, and Hearing) were asked to evaluate each of the identified approaches for subtitles within immersive environments. Head-locked was clearly identified as the preferred choice. However, it was noted that this was a likely outcome since it replicated the experience that users were familiar with, as indicated above. It was also acknowledged that it was difficult for users to evaluate new methods theoretically without the opportunity to experience them while accessing content. Although all agreed that the head-locked option should be set as default, the respondents maintained that other choices should be made available. Other suggestions included changing the font size and colour and the number of lines (two lines being the default). Consequently, the need to develop a framework enabling delivery of the full experience of each captioning mode, in an environment where an extensive user study could be conducted, would be a priority prior to testing.

3. Methodology for a Pilot Study

Conducting a pilot study before launching a full spectrum study is always desirable. The goal of such a pilot study is not only to try to ensure that the survey questions operate well, but also that the research procedures and measures are adequate and reliable (Bryman, 2004). Especially when research aims to substantiate the validity of a new framework and/or involve the use of novel technology (such as eye-tracking in VR), a pilot study is crucial to ensure that both methodology and the study design are accurate and reliable. The preparation stage for this pilot study involved four main steps: user profile definition, selection of the testing material, implementing the material within the new framework, and design of the test procedure itself.

The procedure followed by the current study consisted of four stages: an introduction, a questionnaire on demographic information, an eye-tracking test using 360º immersive videos, and a focus group. The main aims were (1) to test a new framework for subtitle presentation in 360º videos, (2) to obtain feedback regarding expectations, recommendations, and preferences from users when consuming subtitles, and (3) to explore the visual attention distributions between subtitles and movie scenes while watching videos in VR. To do so, three different subtitle modes were implemented:

[2] https://www.w3.org/community/immersive-captions/

mode 1 (following ImAc results), mode 2 (following Fox's (2018) studies), and mode 3 (fully customised).

Before starting the pilot study and taking the previous work in the field as a reference, the following hypothesis was formulated: Fixed, near to the mouth subtitles will allow viewers to spend more time exploring the image instead of reading the subtitles.

3.1. The Live Web Testing Framework

One of the challenges for testing immersive subtitles is the difficulty of having users evaluate new modalities properly because of the cost and time needed to create new prototype subtitle presentations so that users can experience them. To this end, an XR subtitle web simulator was developed by Hughes et al. (2020b). This web-based simulator was designed for rapid prototyping of subtitles within a 360º space, as can be seen in Figure 3 below.

Figure 3

Open-Source Demo Player Developed as Part of This Study

Source: Chris Hughes' demo player screenshot.

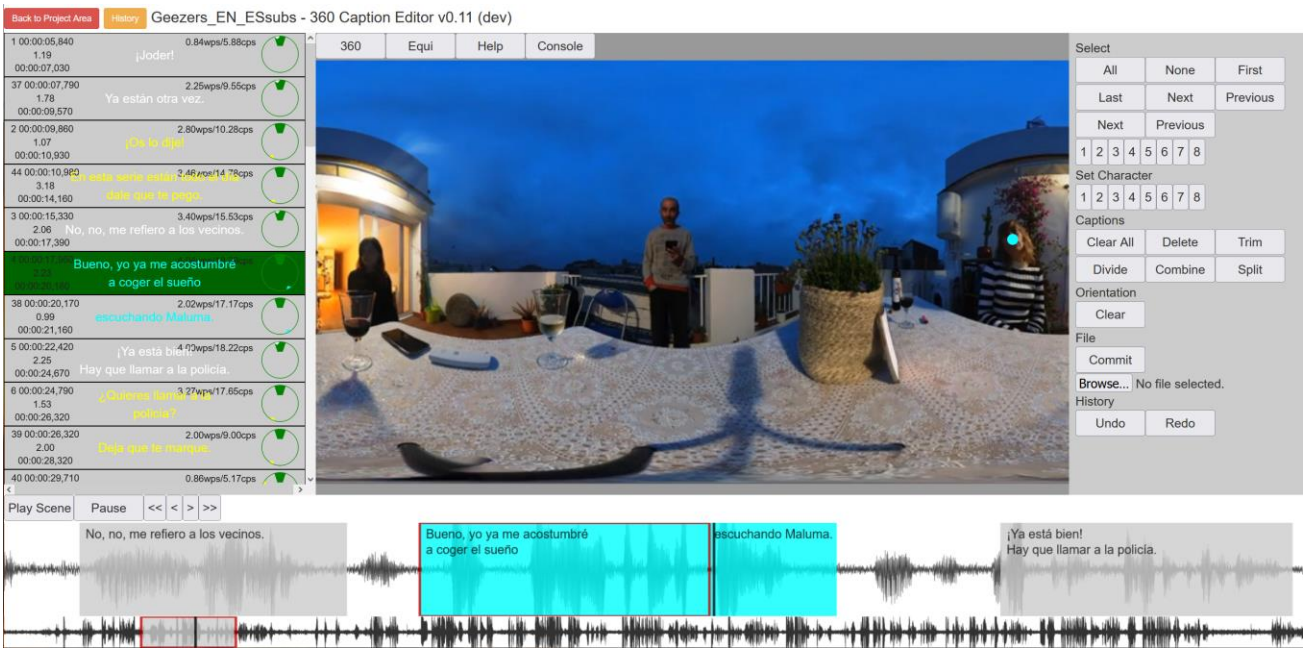
This new framework allows for instant immersive subtitle production in up to nine different modes: four of them are fixed, where the subtitle is rendered relative to a fixed location in the world, generally at the position of the character speaking, and five are head-locked, where the subtitle is rendered relative to the user's viewpoint. The main idea behind this demo player is to allow as much personalisation as possible (i.e., subtitle display, placement, timing, render mode, guiding

mechanism, etc.); this way, any feature may be activated to define and test subtitles within 360º videos.

Along with this XR subtitle simulator, a web-based editor was also developed (see Figure 4), which allows previously created subtitles to be imported in .srt format or subtitles to be created from scratch. On the one hand, each subtitle can be associated with a character ("Set Character" button), and, on the other hand, each subtitle must have an associated position (FoV), i.e., the place in the 360º scene where it should appear.

Figure 4

Open-Source Editor Developed as Part of This Study



Source: Chris Hughes' subtitle editor screenshot.

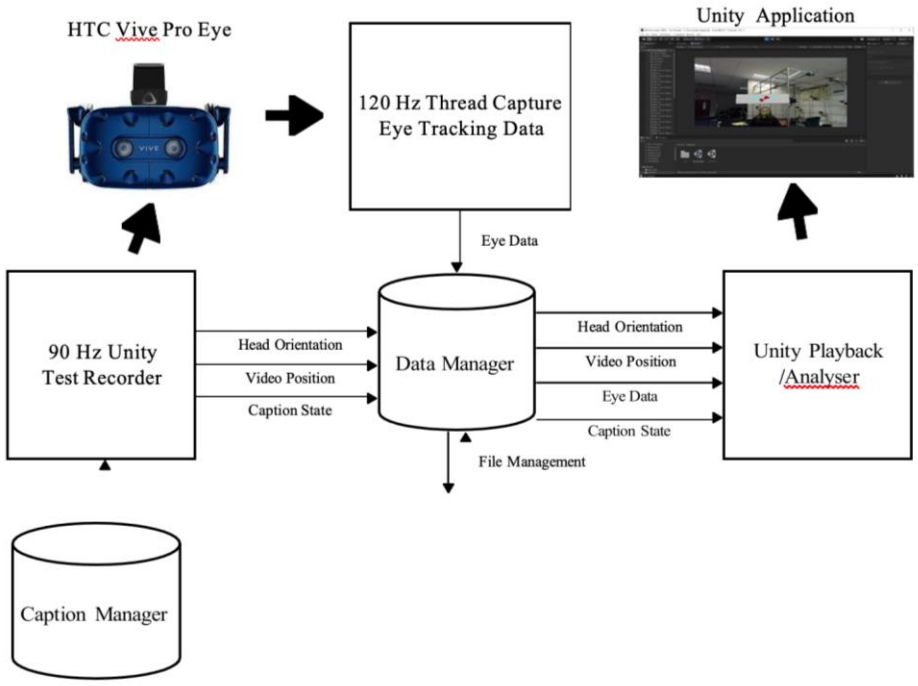
Both the demo player and the editor are open-source and can be accessed from a main project area where all the imported 360º videos are located. These tools take their inspiration from the player and the editor developed in the ImAc project. The main difference between them is that ImAc tools are intended to be used by generic audiences (final users), while the tools used in this study are more focused on research and testing.

3.2. System Architecture

To enable recording of gaze within 360º video, the live web testing framework developed by Hughes et al. (2020a, 2020b) was ported to Unity 3D to allow the display of 360º video content and to capture data from the eye tracker built into the VR device. A new system architecture emerged, as depicted by the schematic in Figure 5.

Figure 5

Eye-Tracking VR System Architecture



Source: Authors' own elaboration.

The system architecture was developed to utilise the HTC Vive Pro Eye, which contains an eye tracker from Tobii built into the display. The application uses two Unity assets, one specifically optimised for recording and the other for playback. At the centre of the architecture is a Data Manager, designed to store all test data. It also handles file management and can generate the output data in a variety of formats as required.

The recording application allows for a specified 360º video to be played with the captions fixed in the scene. During the test, each event and data are logged into the data manager as they become available and timestamped. In order to be able to replay a user viewing session, the system needs to record head orientation, video (frame) position, gaze data (raw and analysed, see below) as well as the subtitle caption state, i.e., which caption from the accompanying subrip format (.srt) file was being displayed and where.

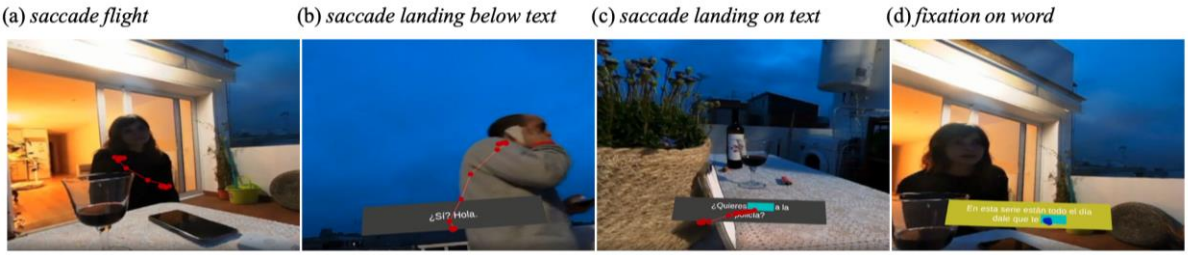
The playback application allows for the data to be retrieved from the Data Manager and the entire test to be replayed. This offers the opportunity to change the analysis process or to include additional Areas of Interest (AOIs) and makes it possible to repeat the analysis. It also permits visual analysis by overlaying the eye data onto the video following capture.

One technical difficulty that had to be overcome was synchronisation of gaze data with video and subtitle data. Gaze data is sampled at 120 Hz while the Unity display refresh rate is 90 Hz. Thus, on average 1.3 gaze samples are expected on any given frame. To enable synchronisation from data

streams of different rates, a separate eye-tracking data thread was created to collect gaze data captured at 120 Hz, ensuring no loss of eye movement samples. System playback can be set to either the speed of video or eye tracker, with gaze data drawn atop the projected video, as shown in Figure 6.

Figure 6

Gaze Recording in VR Showing Varying Elements of Gaze to Subtitle: (a) Saccade in Mid-Flight, (b) Saccade Landing Site With Slight Undershoot, (c) Saccade to Midpoint of Subtitle, (d) Fixation Within a Subtitle



Source: Authors' own elaboration.

3.3. Participants

The size of the group was determined in accordance with the pilot nature of the study (Bryman, 2004, p. 507). In the beginning, 7 participants were expected, but due to complications related to the Covid-19 pandemic, only five appeared (2 male and 3 female). All participants were professionals from the Arts, Sciences, or Humanities fields, staying for a few weeks at the residence Faberllull in Olot. The average age was 40 (*SD* = 8.37) and all of them had completed a postgraduate university degree. All were active professionally (1 social worker, 1 cultural manager, 1 music therapist, 1 pre-doctoral researcher, 1 project manager). All participants spoke Spanish and at least one other language.

All participants were familiar with using computers and mobile devices. Two participants reported having previous experience with VR. Most of the participants declared watching different TV content with subtitles at least occasionally (only one of them claimed that she/he never used subtitles).

3.4. Study Materials

One of the main concerns of the study was to find appropriate material for testing. Due to the difficulty in finding royalty-free material that met the needs of the study, a homemade 360º video was recorded using an Insta360 One X2 camera. The duration was 3 minutes and 45 seconds. The camera was settled in the centre of the action and three characters were positioned around the camera so that the action took place throughout the 360º space. The characters followed a script to

avoid overlaps because if two characters located at different points in the 360º scene spoke at once, it would be almost impossible for the user to read the subtitles.

There were three types of subtitles yielding three experimental conditions:

- Mode 1: following ImAc results. Same font and colour (b&w) for all the characters, with a grey background and head-locked.
- Mode 2: following Fox's (2018) studies. Same font and colour (b&w) for all the characters, without background and near the mouth.
- Mode 3: fully customised. Different font and colour for each character, with a grey background and near the mouth.

Other conditions, for example subtitles for non-human sound sources, will be presented in future tests.

3.5. Procedure

The study included the following stages. First, participants were welcomed by the facilitator, who briefly explained the aim of the project. The session took place in a meeting room divided into two separate spaces. On one side there was a large TV screen, a computer connected to the screen and chairs for the participants. On the other side, an improvised eye-tracker lab was installed with a computer and a pair of HTC Vive Pro HMD. One researcher took notes and summarised the conclusions in real-time. Second, the aim of the focus group was explained to the participants, and they were asked to sign informed consent forms. The third step consisted of filling in a short questionnaire on demographic information. Finally, the session began. To trigger the discussion, the facilitator gave a short introduction to VR and 360º content and explained how subtitles are integrated within 360º content, showing VR glasses to the participants.

The eye-tracking technology was introduced, as it is integrated within the VR glasses and was one of the data collecting methods in the study. The facilitator explained that 360º content can also be accessed on a flat TV screen using a mouse to move around the 360º scene. Different types of subtitles were presented to give users some idea about how creative subtitling can be implemented in immersive content and to stimulate their imagination.

Then, each participant used the HTC Vive Pro HMD to watch a short video with audio in English and subtitled into Spanish. In total there were three rounds, always using the same video but a different subtitle mode each time. The order of the participants was determined randomly. Immediately after each visualisation, participants filled out a short questionnaire with questions on content understanding, subtitling preferences, and the task load index (NASA-TLX).

After the last round the focus group took place. Together with the stimuli, the facilitator used a list of guiding questions grouped under major topics to generate the participants' reactions. A balance

between an open-ended and a structured approach was sought, and the result was a lively discussion in which interesting ideas came up.

4. Pilot Study Results

The data analysis of the study was mainly qualitative accompanied by descriptive statistics of the post-study questionnaire and eye movements captured during the study (see Figure 6).

4.1. Movie Content Understanding

To check the understanding of the stimuli movies, we averaged the accuracy of responses to questions about the content separately for each condition. The highest average accuracy was obtained for the movie with fully customised subtitles (M = 0.64, SD = 0.26). Average accuracy for movies with subtitles in mode 1 (M = 0.52, SD = 0.18) and mode 2 (M = 0.52, SD = 0.36) were the same.

Additionally, when asked about the description of the scenes presented in the movie, participants used, on average, slightly more words after watching the movie in mode 1 (M = 22.2, SD = 12.99) than mode 3 (M = 18, SD = 8.34). The smallest number of words used in the description after watching the movie was in mode 2 (M = 16.20, SD = 9.01).

Qualitative analysis of responses during the focus group interviews showed that some of the participants could not understand the plot until the third visualisation of the clip. This could be related to a learning effect, but also because 3 of the participants had no previous experience with subtitled immersive content. Furthermore, another participant commented that sometimes it was difficult to follow the story because she was distracted exploring the 360º scene. The participant who was the least familiar with new technologies (and the least interested in the immersive format) noted that paying attention to the story stressed her and that she tried to distract herself during the visualisations.

4.2. Subtitle Readability

The participants were asked whether they had been able to read the subtitles after watching each movie. Two responded “yes”, two “no”, and one was “not sure” for mode 1. In mode 2, two responded “yes” and three “no”. The least readable subtitles seemed to be in mode 3. Three participants noted they were not able to read them; only one responded “yes” and one participant was “not sure”. When asked to estimate the percentage of subtitles that they were able to read, the differences were very small: 70% in mode 1, 68% in mode 2, and 67% in mode 3. Both results seem to suggest a slight preference for the subtitles in mode 1 as the most readable ones.

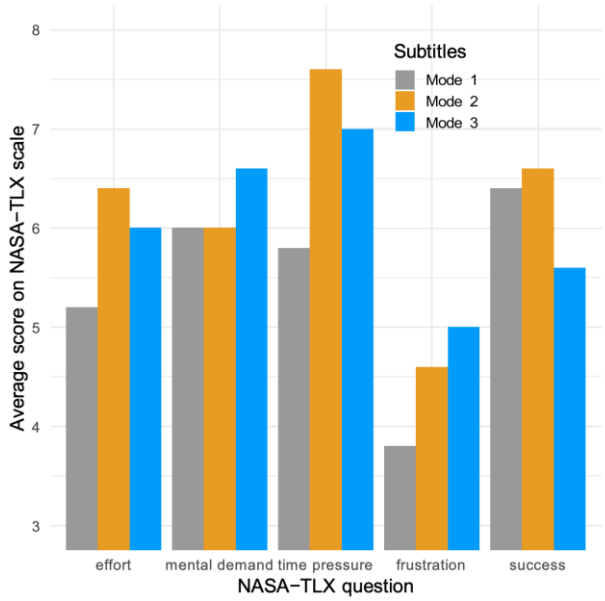
These results comply with the qualitative data extracted from the focus group, since most participants agreed that mode 3 was difficult to read. Only one of the participants noted that she liked the coloured text, and there was a brief discussion about the possibility of customising the subtitles further. Regarding the grey background, there was no consensus: some of the participants found subtitles with no background hard to read, others found them less intrusive. One participant highlighted the reading pace in general, arguing that some captions disappeared "too soon" forcing the user to read faster.

4.3. Self-Reported Task Load

To collect self-reports on the effort elicited by the task of watching stimuli movies in different subtitle modes, the NASA-TLX scale with five questions was analysed (see Figure 7). Subjective evaluation of effort while watching videos in different subtitle modes also suggests a preference for mode 1 (M = 5.20, SD = 2.39). In the participants' opinion, more effort was required to read subtitles in modes 2 (M = 5.80, SD = 2.17) and 3 (M = 6.00, SD = 1.87). However, evaluation of mental load shows a different pattern, namely that modes 1 (M = 6.00, SD = 2.35) and 2 (M = 6.00, SD = 2.35) were equally less demanding than mode 3 (M = 6.60, SD = 2.30). Participants also evaluated reading modes 1 (M = 6.40, SD = 1.52) and 2 (M = 6.60, SD = 2.30) with greater perceived success than mode 3 (M = 5.6, SD = 2.19). These results are not surprising considering the average responses regarding how time-pressured participants felt. Subtitles in mode 2 caused the highest experience of time pressure (M = 7.60, SD = 0.89); this was lower in mode 3 (M = 7.00, SD = 0.71) and lowest in mode 1 (M = 5.80, SD = 2.17). The perceived level of frustration/stress was lowest when watching the video in mode 1 (M = 3.80, SD = 2.17), greater in mode 2 (M = 4.60, SD = 3.21), and greatest in mode 3 (M = 5.00, SD = 2.35).

Figure 7

Task Load Self-Reports With NASA-TLX Scale While Watching Videos in Different Subtitle Modes



Source: Authors' own elaboration.

4.4. Attention Distribution and Cognitive Effort While Reading Captions and Scene Viewing

Gaze was captured as it traversed subtitles when reading the text displayed within the quadrilaterals that contained them. The analysis of the eye movement signal relies on fixation detection, which in turn depends on saccade detection. Fixations are detected within the raw eye movement signal following Nyström and Holmqvist (2010) and by using the Savitzky-Golay filter for velocity-based (I-VT (Salvucci & Goldberg, 2000)) event detection Savitzky and Golay (1964).

The current system architecture allows for detection of fixations falling within arbitrarily defined Areas of Interest (AOIs), including polygons defined over actors and more importantly over quadrilaterals (quads) used to display subtitles as well as quads defined over individual words, see Figure 8 below.

Figure 8

Gaze Recording Showing Fixations Over Areas of Interest: (a) Actor Body, (b) Subtitle Quad, and (c) Individual Word

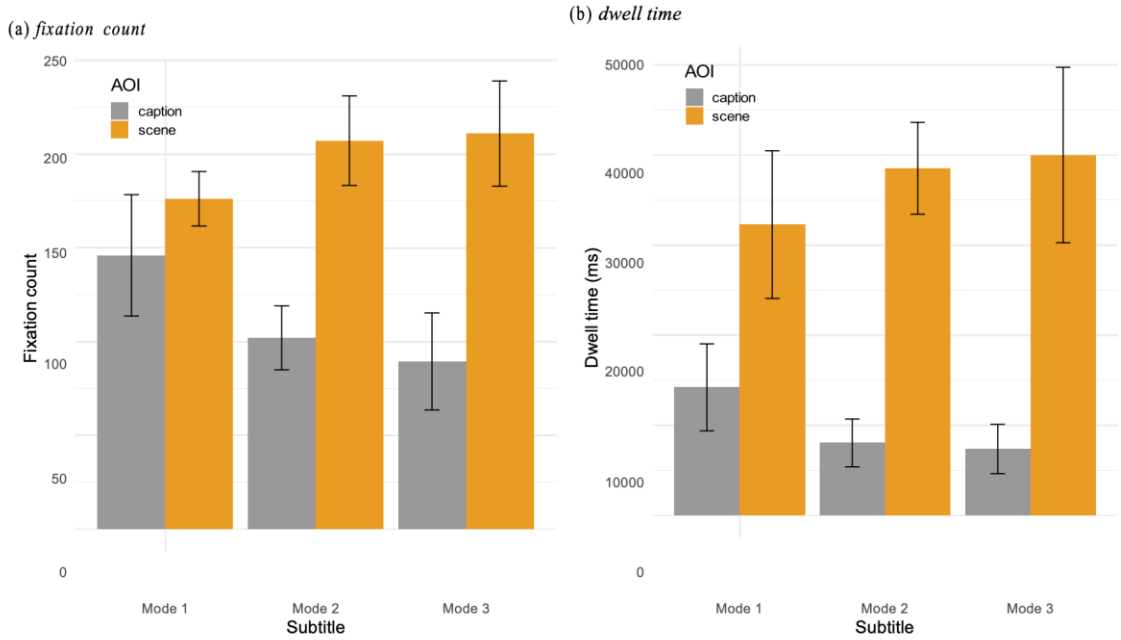
Source: Authors' own elaboration.

The eye movement analysis aimed first at capturing differences in attention to captions and visual scenes in terms of fixation count and dwell time as dependent variables. Descriptive statistics show that in all conditions most fixations were on the visual scene rather than on subtitles. However, the difference is less important for the video in mode 1.

Participants exhibited more fixations on captions (M = 145.8, SD = 32.37) than on the visual scene (M = 176.0, SD = 14.54) in mode 1 compared to modes 2 (for caption M = 101.80, SD = 17.08; for scene M = 207.00, SD = 23.92) and 3 (for caption M = 89.20, SD = 25.90; for scene M = 210.80, SD = 28.05) see Figure 9(a). A similar pattern is observed when analysing dwell time. On average, participants dwelled more on captions than on the visual scene in mode 1 than in modes 2 or 3, see Figure 9(b). Participants appeared to allocate more attention to captions when viewing subtitles in mode 1 than in modes 2 or 3.

Figure 9.

Visual Attention Distribution Over Captions and Visual Scenes While Watching Video With Different Types of Subtitles



Note: Attention distribution is depicted by two metrics: (a) shows fixation counts over captions and visual scene, (b) shows dwell time of captions and visual scene fixating.

Bars height represents mean values and whiskers represent ±1 SD.

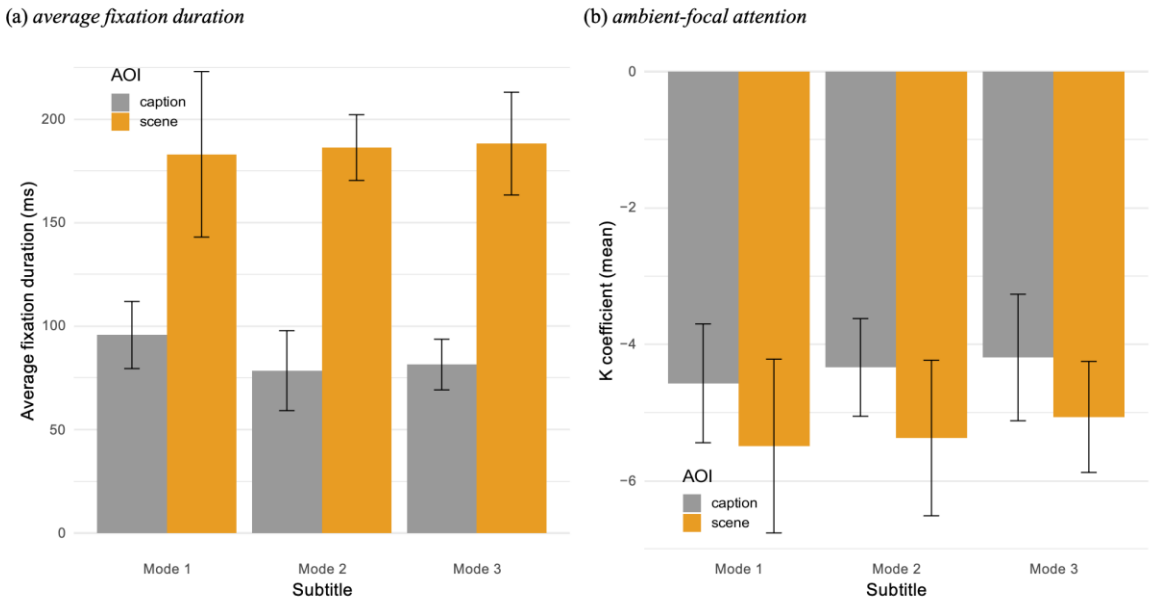
Source: Authors' own elaboration.

We also examined cognitive effort while processing information from captions or visual scene based on average fixation duration (following the *eye-mind assumption* (Just & Carpenter, 1976)) and focus of attention with coefficient *K* (Krejtz et al., 2016), which captures the temporal relation between fixation duration and subsequent saccade amplitude. $K > 0$ indicates focal viewing while $K < 0$ suggests ambient viewing. Focal attention is usually related to higher cognitive effort when processing complex visual or text stimuli (Duchowski et al., 2020; Krejtz et al., 2017; Krejtz et al., 2018). Analysis of descriptive statistics on average fixation duration showed that the visual scene triggered longer average fixation durations than captions in all modes. However, the difference in average fixation durations between the visual scene and the caption is smallest in mode 1.

Moreover, fixation duration on subtitles in mode 1 ($M = 95.73$, $SD = 16.22$) is much longer than on subtitles in either mode 2 ($M = 78.47$, $SD = 19.35$) or mode 3, ($M = 81.43$, $SD = 12.24$), see Figure 10(a). Coefficient *K* showed that viewers were not as focused when reading captions in mode 1 ($M = -4.57$, $SD = 0.87$) compared to mode 2 ($M = -4.34$, $SD = 0.72$) or 3 ($M = 4.19$, $SD = 0.92$), see Figure 10(b). Both fixation duration and coefficient *K* suggest the highest cognitive effort along with decreased focal processing when processing subtitles in mode 1.

Figure 10

Cognitive Processing of Textual and Visual Information From Captions and Visual Scenes While Watching Video With Different Types of Subtitles



Note: (a) shows average fixation duration as a metric for cognitive effort, (b) shows K coefficient as a metric ambient-focal attention.

Bar heights represent mean values and whiskers represent ±1 SD.

Source: Authors' own elaboration.

4.5. Focus Group Insights

A qualitative analysis was carried out on the notes taken during the focus group (the last part of the session, after the participants had watched the video with the different subtitle options). The notes were thoroughly revised and tagged using Atlas.ti. This procedure allowed us to identify three areas that can be associated with the quantitative analysis (subtitle readability, task load, and understanding of the movie content). The analysis also allowed us to define user preferences and identify aspects on which there was consensus among users and issues on which opinions diverged.

In terms of preference, most of the participants agreed that mode 2 was the easiest to read with one participant suggesting adding a colour code to mode 2 (like mode 3). The second preferred option was mode 1 (selected by 2 participants). The main problem in mode 1 seems to have been the difficulty in identifying the character speaking at each given moment.

Regarding creative subtitles, all the participants agreed it would be a great idea to dramatise what is said and that this could add much visual beauty to the content. One of the participants noted that, in some cases, so much creativity distracted her from the content itself.

As in previous studies reviewed earlier, participants highlighted the lack of direction to guide people to the source of the sound (guiding mechanisms). Some of them mentioned that they missed human interaction when watching the immersive content, and that they felt isolated when wearing the HMD for the first time.

5. Discussion

The first hypothesis we wanted to validate was whether fixed, near to the mouth subtitles allow viewers to spend more time exploring the image instead of reading the subtitles when compared to head-locked subtitles. Although the present pilot study cannot yield conclusive evidence, eye movement data together with focus group insights seem to support this hypothesis. Interestingly, eye movement data appear to be consistent with qualitative insights from the focus group, suggesting that participants tend to prefer fixed subtitles located near the mouth of the speaking character (mode 2). These results differ from those obtained in previous studies, in which participants opted for head-locked subtitles.

The results of self-reported cognitive load during movie watching with different subtitle modes suggest a slight preference (less perceived mental effort and higher perceived success in reading captions) for mode 1 (b&w font for all characters, grey background, and head-locked) over modes 2 and 3. However, results carry a large statistical variance and cannot be interpreted decisively. The results may also be biased by a lack of randomisation in the order of presentation and learning effect of the questions during the experimental procedure. Future studies must employ tighter experimental control over stimulus presentation order (e.g., via randomisation or counterbalancing).

Eye movement analysis sheds light on attention allocation (captions vs. scene) and perception. Identification of fixations showed that participants allocated more attention to captions and less to the visual scene when viewing subtitles in mode 1 than in modes 2 or 3. Process measures (average fixation and ambient-focal coefficient) suggest higher cognitive effort paired with less focal processing of subtitles in mode 1.

Subtitles in modes 2 and 3 appear to outperform mode 1 as they may be less distracting from scenes in the movie, but they also seem to require less cognitive effort when focused on reading. We do not know, however, whether mode 2 or 3 is easier to read and less distracting when movie watching. This issue needs to be addressed in a study with more experimental control and a larger sample.

The visualisation of the eye movements analysed, specifically saccades, drawn in red in Figure 11, expose the inadequacy of the velocity-based filtering approach. The I-VT method, while computationally efficient and generally applicable to traditional desktop displays, tends to ignore

head-induced gaze movement when captured in the VR HMD. It is likely that a better model of eye and head coupling is required (Guitton et al., 1990), e.g., a fixation detection algorithm suitable for immersive environments (Llanes-Jurado et al., 2020).

Figure 11

The View and Scanpath for Each Participant

Note: rows: participant 1–5, columns: mode 1-3, 00:22. General observations can be drawn, such as that participant 1, although finding the captions in modes 1 and 3, was lost in mode 2 and can be observed saccading between the mouths of the wrong characters, trying to identify the character speaking. The second participant can be seen fixating on the speaking character's mouth rather than reading the caption in modes 1 and 2. Also Participants 3, 4 and 5 can be observed reading the captions in modes 1 and 2, but not mode 3.

Source: Authors' own elaboration.

6. Conclusions

Immersive environments simulate reality to heighten the immersive experience. Attention should be paid when drawing on results from previous studies on subtitle reading performance in 2D to VR environments. Prior research on evaluation of subtitles has mainly focused on subtitle style, speed of display, and positioning, and has largely been qualitative. In VR we still need to define the challenges faced by subtitle research. Investigating how subtitles are read is one logical quest but finding subtitles when two people are interacting from different fields of vision is also a candidate for testing. The tests carried out in this study have shown that we need some basic understanding of media presentation in VR and user behaviour and habits when consuming media in VR, where the narrative is no longer linear. VR media environments present us with new variables to consider when testing for optimal subtitle presentation. The objective is to find the least disruptive subtitle reading experience for protecting immersivity and the simulation of reality. To understand the visual presentation of subtitles in VR, a framework was developed along three basic presentation modes for 360º videos, which were all piloted. Our contribution is thus two-fold: on the one hand, there is our presentation of the VR subtitle framework and, on the other hand, a new method for triangulation of psycho-physiological (eye movements) self-reports and qualitative (focus group discussions) analyses. To the best of our knowledge, this is the first attempt to advance these two directions when discussing subtitle presentation in VR 360º videos.

Immersive environments need new subtitle presentation modes and reading patterns. This article described the first pilot study using a comprehensive methodological environment to test subtitles in immersive environments. The novel testbed includes a subtitle editor and a VR system designed specifically to collect eye movement data as visual attention is distributed over 360º videos containing subtitles. Both the framework and the methodology tested in this first pilot study can be used to collect quantitative and qualitative behavioural data when viewing subtitled 360º media. Future studies involving more participants are expected to yield new insights and lead to subtitle standardisation in immersive media environments.

References

Agulló, B., & Matamala, A. (2019). Subtitling for the deaf and hard-of-hearing in immersive environments: Results from a focus group. *Journal of Specialised Translation, 32.*

Agulló, B., Montagud, M., & Fraile, I. (2019). Making interaction with virtual reality accessible: Rendering and guiding methods for subtitles. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 33*(4), 416–428. https://doi.org/10.1017/ S0890060419000362

Bartoll, E. (2004). Parameters for the classification of subtitles. In P. Orero (Ed.), *Topics in Audiovisual Translation* (pp. 53–60). John Benjamins. https://doi. org/10.1075/btl.56.08bar

Brescia-Zapata, M. (2022). The present and future of accessibility services in vr360 players. *inTRAlinea, 24.*

Brown, A. (2017, October 26). *User testing subtitles for 360° content*. BBC. https://www.bbc.co.uk/rd/blog/2017-10-subtitles- 360-video-virtual-reality-vr

Brown, A., & Patterson, J. (2017, October 26). *Designing subtitles for 360º content*. BBC. https://www.bbc.co.uk/rd/blog/2017- 03-subtitles-360-video-virtual-reality

Brown, A., Turner, J., Patterson, J., Schmitz, A., Armstrong, M., & Glancy, M. (2017). Subtitles in 360-degree video. *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*. The Netherlands, 3–8. https://doi.org/10.1145/3084289.3089915

Bryman, A. (2004). *Social research methods*. Oxford University Press.

Cantero de Julián, J. I., Calvo Rubio, L. M., & Benedicto Solsona, M. Á. (2020). La tenue apuesta por los vídeos en 360º en las estrategias transmedia de las televisiones autonómicas españolas. [The weak bet on videos in 360º in the transmedia strategies of Spanish autonomic televisions]. *Revista Latina de Comunicación Social*, (75), 415–433. https://doi.org/https://doi.org/10.4185/RLCS-2020-1433

Díaz-Cintas, J., & Remael, A. (2007). *Audiovisual translation: Subtitling*. St. Jerome.

Duchowski, A. T., Krejtz, K., Zurawska, J., & House, D. H. (2020). Using microsaccades to estimate task difficulty during visual search of layered surfaces. *IEEE Transactions on Visualization and Computer Graphics, 26*(9), 2904–2918. https://doi .org/10.1109/TVCG.2019.2901881

Foerster, A. (2010). Towards a creative approach in subtitling: A case study. In J. Díaz-Cintas, A. Matamala, & J. Neves (Eds.), *New insights into audiovisual translation and media accessibility: Media for all 2* (pp. 81–98). Rodopi.

Fox, W. (2018). *Can integrated titles improve the viewing experience? Investigating the impact of subtitling on the reception and enjoyment of film using eye-tracking and questionnaire data.* Language Science Press. https://doi.org/10.5281/zenodo.1180721

Gottlieb, H. (1995). Establishing a framework for a typology of subtitle reading strategies – viewer reactions to deviations from subtitling standards. *Communication Audiovisuelle et Transferts Linguistiques – FIT Newsletter, 14*(3–4), 388–409.

Guitton, D., Munoz, D. P., & Galiana, H. L. (1990). Gaze control in the cat: Studies and 256odelling of the coupling between orienting eye and head movements in different behavioral tasks. *Journal of Neurophysiology, 64*(2), 509–531. https://doi.org/10.1152/jn.1990.64.2.509

Hughes, C., Armstrong, M., Jones, R., & Crabb, M. (2015). Responsive design for personalised subtitles. *Proceedings of the 12th International Web for All Conference*, Italy, 1–4. https://doi.org/10.1145/2745555.2746650

Hughes, C., Brescia-Zapata, M., Johnston, M., & Orero, P. (2020a). Immersive captioning: Developing a framework for evaluating user needs. *IEEE AIVR 2020: 3rd International Conference on Artificial Intelligence & Virtual Reality 2020*. http://usir.salford.ac.uk/id/eprint/58518/

Hughes, C., & Montagud, M. (2020b). Accessibility in 360° video players. *Multimedia Tools and Applications*, 1–28. https://doi.org/10.1007/s11042-020-10088-0

Hughes, C., Montagud Climent, M., & tho Pesch, P. (2019). Disruptive approaches for subtitling in immersive environments. *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video. United Kingdom*, 216–229. https://doi.org/10.1145/3317697.3325123

Jenkins, H., Ford, S., & Green, J. (2015). *Cultura transmedia. La creación de contenido y valor en una cultura en red [Transmedia culture. Content creation and value in a network culture].* Gedisa.

Just, M. A., & Carpenter, P. A. (1976). The role of eye-fixation research in cognitive psychology. *Behavior Research Methods & Instrumentation, 8*(2), 139–143. https://doi.org/10.3758/BF03201761

Krejtz, K., Çöltekin, A., Duchowski, A., & Niedzielska, A. (2017). Using coefficient K to distinguish ambient/focal visual attention during map viewing. *Journal of Eye Movement Research, 10*(2). https://doi.org/10.16910/jemr.10.2.3

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., & Kopacz, A. (2016). Discerning ambient/focal aAt- tention with cCoefficient K. *ACM Transactions on Applied Perception, 13*(3), 11:1–11:20. https://doi. org/10.1145/2896452

Krejtz, K., Wisiecka, K., Krejtz, I., Holas, P., Olszanowski, M., & Duchowski, A. T. (2018). Dynamics of emotional facial expression recognition in individuals with social anxiety. *Proceedings of the 2018 ACM Symposium on Eye-tracking Research & Applications. Warsaw, Poland*, 43:1–43:9. https://doi.org/10.1145/3204493.3204533

Lee, D. G., Fels, D. I., & Udo, J. P. (2007). *Emotive captioning. Comput. Entertain*, *5*(2). https://doi.org/10.1145/1279540.1279551

Llanes-Jurado, J., Marín-Morales, J., Guixeres, J., & Alcañiz, M. (2020). Development and calibration of an eye-tracking fixation identification algorithm for immersive virtual reality. *Sensors, 20*(17). https://doi.org/10.3390/s20174956

Mas Manchón, L., & Orero, P. (2018). Usability tests for personalised subtitles. *Translation spaces, 7*(2), 263–284. https://doi.org/10.1075/ts.18016.man

Matamala, A., & Orero, P. (2018). Standardising accessibility: Transferring knowledge to society. *Journal of Audiovisual Translation, 1,* 139–154. https://doi.org/10.47476/jat.v1i1.49

McClarty, R. (2012). Towards a multidisciplinary approach in creative subtitling. *MonTI: Monografías de Traducción e Interpretación*, 133–153. https://doi.org/10.6035/MonTI.2012.4.6

McClarty, R. (2014). In support of creative subtitling: Contemporary context and theoretical framework. *Perspectives, 22*(4), 592–606. https://doi.org/10.1080/0907676X.2013.842258

Mével, P. A. (2020). Accessible paratext: Actively engaging (with) d/deaf audiences. *Punctum. International Journal of Semiotics, 6*(01). http://doi.org/10.18680/hss.2020.0010

Mittell, J. (2009). *Forensic fandom and the drillable text.* Spreadable media. https://spreadablemedia.org/essays/mittell/index.html#.YgzhV-5_o4g

Montagud, M., Fraile, I., Meyerson, E., Genís, M., & Fernández, S. (2019). Imac player: Enabling a personalized consumption of accessible immersive contents. https://doi.org/10.6084/m9.figshare.9879254.v1

Montagud, M., Orero, P., & Matamala, A. (2020). Culture 4 all: Accessibility-enabled cultural experiences through immersive vr360 content. *Personal and Ubiquitous Computing, 24*(6), 887–905. https://doi.org/10.1007/s00779-019-01357-3

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eye-tracking data. *Behaviour Research Methods, 42*(1), 188–204. https://doi.org/10.3758/BRM.42.1.188

Orero, P., Hughes, C. J., & Brescia-Zapata, M. (2020). Evaluating subtitle readability in media immersive environments. *DSAI 2020: 9th International Conference on Software development and Technologies for Enhancing Accessibility and Fighting Info-exclusion. Online Portugal*, 51–54. https://doi.org/10.1145/3439231.3440602

Romero-Fresco, P. (2013). Accessible filmmaking: Joining the dots between audiovisual translation, accessibility and filmmaking. *The Journal of Specialised Translation, 20*, 201–223.

Romero-Fresco, P. (2015). *The reception of subtitles for the deaf and hard of hearing in Europe.* Peter Lang. https://www.peterlang.com/view/title/36324

Rothe, S., Tran, K., & Hussmann, H. (2018). Positioning of subtitles in cinematic virtual reality. In Bruder, G., Yoshimoto, S., and Cobb, S., editors, *International Conference on Artificial Reality and Telexistence and Eurographics Sysmposium on Virtual Environments*, ICAT-EGVE 2018. The Eurographics Association.

Rupp, M., Kozachuk, J., Michaelis, J., Odette, K., Smither, J., & McConnell, D. (2016). The effects of immersiveness and future vr expectations on subjective – experiences during an educational 360 video. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60,* 2108–2112. https://doi.org/10.1177/1541931213601477

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the 2000 Symposium on Eye-Tracking Research & Applications.* USA, 71–78. https://doi.org/10.1145/355017.355028

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*(8), 1627–1639. http://pubs.acs.org/doi/abs/10.1021/ac60214a047

Skult, N., & Smed, J. (2020). Interactive storytelling in extended reality: Concepts for the design. In B. Bostan (Ed.), *Game user experience and player-centered design* (pp. 449–467). Springer International Publishing. https://doi.org/10.1007/978-3-030-37643-7_21

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments, 6,* London. United Kingdom, 603–616. https://doi.org/10.1162/pres.1997.6.6.603

Smith, W. (2015, November 16). *Stop calling google cardboard's 360-degree video 'vr'*. Wired. https://www.wired.com/_2015/11/360-video-isnt-virtual-reality/