

VR 360° subtitles: Designing a test suite with eye-tracking technology

 **Marta Brescia-Zapata** 

Universitat Autònoma de Barcelona

Krzysztof Krejtz 

SWPS University of Social Sciences and Humanities

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 allow for placing subtitles at any location of the screen in a variety of formats, shapes, typography, font size, and colour. The screen now affords accessible creativity, with subtitles able to provide novel experiences beyond those offered by traditional language translation. Immersive environments multiply 2D subtitle features to new creative viewing modalities. Testing subtitles in eXtended Reality (XR) has pushed existing methods to address user need and enjoyment of audiovisual content in 360° 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 given, in line with novel research questions. The choices

Citation: Brescia-Zapata, M., K. Krejtz, P. Orero, A. T. Duchowski & C. J. Hughes (2022). VR 360° subtitles: Designing a test suite with eye-tracking technology. *Journal of Audiovisual Translation*, 6(2), XX–XX.

Editor(s):

Received:


Accepted:


Published:


Acknowledgment: 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 Grant 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 i Coneixement de la Generalitat de Catalunya" (2017SGR113).


Copyright: ©2021 Brescia-Zapata, Krejtz, Orero, Duchowski, & Hughes. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/). This allows for unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

 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>

made regarding sound, duration and storyboard are described. 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 future experimental setups with eye-tracking.

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

1. Introduction

Immersive media technologies such as Virtual Reality (VR) and 360° 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 it not only in the entertainment sector, but also for communication, arts, and culture, giving rise to 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. In this context, 360° videos—also known as immersive or VR360 videos—have become very popular as they are a cheap and effective way to provide VR experiences. For the production of content specialised multi-camera equipment is used. These can capture a 360° or 180° Field of View (FoV) instead of the limited viewpoint of a standard video recording. 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 Virtual Reality (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), 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 produce negative effects such as motion or simulator sickness, possibly turning people away from VR as a medium (Smith, 2015).

As for every media content, 360° media experiences should be accessible. Accessibility is an afterthought, despite many voices asking for accessibility in the creation process (Mével, 2020; Romero-Fresco, 2013). We focus on subtitling, where standardised practices have emerged (Matamala & Orero, 2018). 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, 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 360° videos. The present

article aims to identify how to display 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 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 360° 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), “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 center 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. Nevertheless, another activity may distract from that narrative, such as the noise of a lady unwrapping sweets two rows away. The audience has freedom of movement, and the choice of focusing their attention which affects subtitle reading. Similarly, in VR, the concept of presenciality and engagement are central with the ultimate goal of being a witness of the narrative from a first-person viewpoint. This breaks with the concept of passive audience or “spreadable” (Jenkins et al., 2015) and moves towards interaction or “drillable”, as in video games or transmedia products. Jenkins et al. (2015) comments on “the opposition between spreadable and drillable shouldn’t be thought of as a hierarchy, but rather as opposing vectors of cultural engagement. 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 virtual reality 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 different from 2D lies in the way media is accessed. There are two possibilities: using CAVE (naked eye) or using 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 an immersion of the user in whatever experience the display is meant for. The 360° 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, on the HMD itself, or stored on the cloud. As Internet speed improves, media content is increasingly consumed streamed from the web. Video games consoles are finding a confluence with TV and is not unusual to use a PC for the main video game.

As in traditional 2D media, to create and consume subtitles in XR, a subtitle editor and a subtitle player are needed. Although there are commercially available, immersive video players offering the ability to play VR360 video, not many of them support accessible services (Brescia- Zapata, forthcoming). The player needs to be accessible and display accessibility services to be activated by the user. 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 of these features show the complex ecosystem required for a true XR accessible experience. This fact, linked to the lack of standardized 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 from the traditional 2D world, instead of addressing specific features of 360° environments. This scenario served as an inspiration for initiatives like the European H2020 funded Immersive Accessibility (ImAc) project¹ 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. On the one hand, the accessibility-enabled 360° called ImAc player supports audio description, audio subtitles, and sign language, along other features (Montagud et al., 2019) as can be seen in Figure 1.

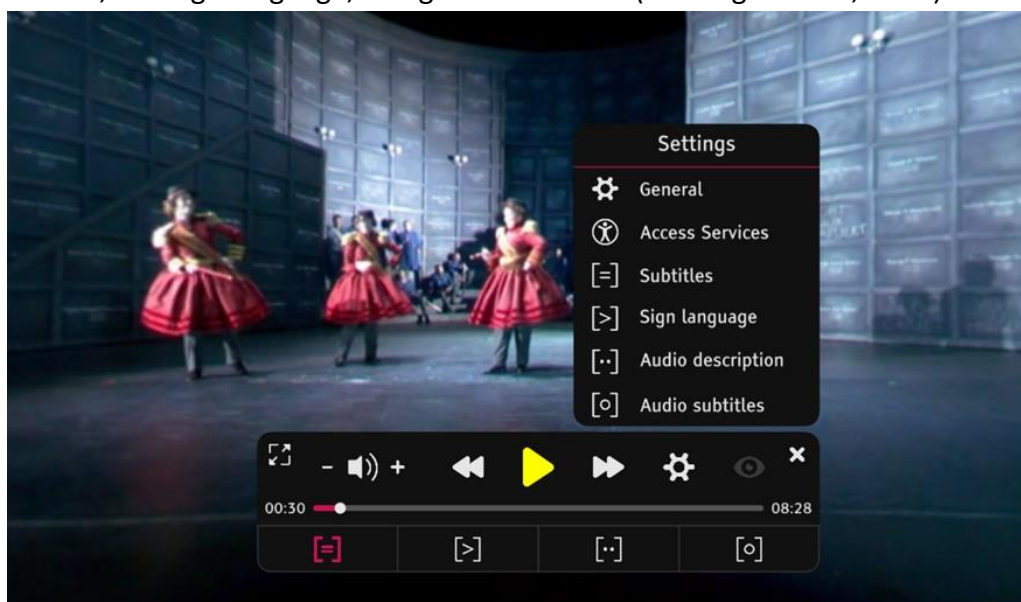


Figure 1.

¹ <http://www.imac-project.eu>

ImAc player settings

On the other hand, the ImAc subtitle editor is a commercial web-based editor, and its interface is similar to 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 to change the FoV angle using the navigation buttons in the video control area or moving the mouse with the left button over the video. By default, at first the video has the current angle as longitude: 0.00° and latitude: 0.00°. Also, the voice over option can be marked when there is no speaker in the 360° scene.

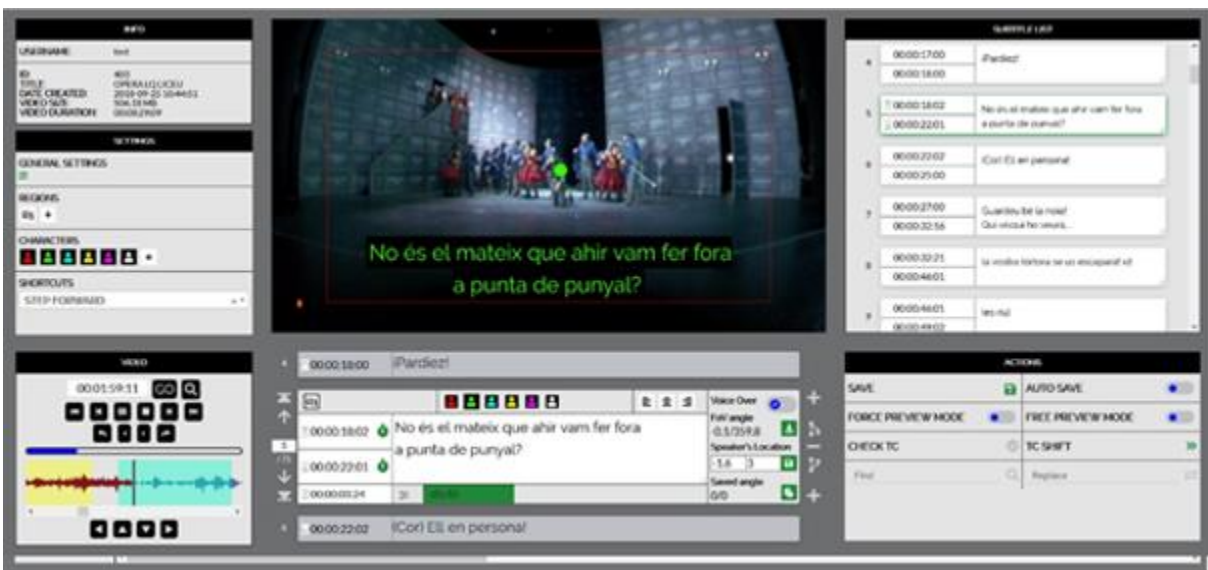


Figure 2.

Immersive subtitle editor developed in ImAc

The basic tools to create and consume accessible VR content are now commercially available, e.g., VR subtitle editor and 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 from the ImAc project which finished recently and focused on traditional subtitles projected on immersive environments (Hughes et al., 2020b).

2.1. Related work

Excluding works that have added (sub)title 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 once 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 (Brown et al., 2017) and based on these they developed the following four solutions for subtitle rendering:

- 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 it 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). These results come to no surprise since for years now, testing subtitles in Europe has always given the same results: people like what they are used to, even if the performance is worse, as demonstrated 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%)

The second study (Rothe et al., 2018) compared the two previous 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 of their FoV.

The third study, performed under the umbrella of the H2020-funded ImAc project (Hughes et al., 2019), obtained similar results but also revealed the need to guide users to the sound source of the subtitle (i.e., a sound effect, or 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 person speaking was located (Agulló et al., 2019). However, this did have the drawback that the location was only specified once per caption, and if a person was moving dynamically during this period, the guide could have been wrong (Hughes et al., 2019). 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 users view. This identifies both the position of the caption, and the relative viewing angle of the user.

In the area of standardization, a W3C Community Group² 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 most likely as it replicated the experience that users were familiar with. It was also acknowledged that it was difficult for users to properly evaluate new methods theoretically without the opportunity and content to enable them to be experienced properly. Although all agreed that head-locked should be set as default, other choices should be made available. Other suggestions were made which included changing the font size and colour and number of lines (two lines being the default number). Multiple captions should also be in different positions, each being near to the speaker. Therefore, the need to develop a framework enabling delivery of the full experience of each captioning mode, in an environment where an extensive user study can be conducted was a priority prior to testing.

3. Methodology for a pilot study

Conducting a pilot study before administering a full spectrum study is always desirable. The goal of piloting such instruments is not only to try to ensure that survey questions operate well, but also to ensure that the research procedures and measures function well (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), the role of the pilot study is crucial to ensure accurate and reliable results. The preparation stage for this pilot study involved four main steps: user profile

² <https://www.w3.org/community/immersive-captions/>

definition, selection of the testing material, implementing the material within the new framework and design of the test procedure itself.

The current pilot study procedure had four stages: an introduction, a questionnaire on demographic information, an eye-tracking test using 360° immersive videos, and a focus group. The main aims of the study were (1) to test a new framework for subtitles presentations in 360° 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: mode 1 (following ImAc results), mode 2 (following Fox, 2018 studies) and mode 3 (fully custom).

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 than head-locked subtitles.

3.1. The live web testing framework

One of the challenges for testing immersive subtitles is the difficulty for users to properly evaluate new modalities. The reasons are the cost and time needed to create new prototype subtitle presentations to enable the users to experience them properly. To allow for visualising creative subtitles, an XR subtitle web simulator has been developed by Hughes et al., (2020b). This web-based simulator was designed for rapid prototyping of subtitles within a 360° space, as can be seen in Figure 3 below.



Figure 3.

Open-source demo player developed as part of this study

This new framework allows for instant immersive subtitle production in up to nine different default 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 and timing, render mode, guiding mechanism, etc.) This way, any feature may be activated to define and test subtitles within 360° videos.

Along with this XR subtitle simulator, a web-based editor was also developed (see Figure 4). This editor allows to import subtitles previously created in .srt format or to create subtitles 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), that is, the place in the 360° scene where it should appear.

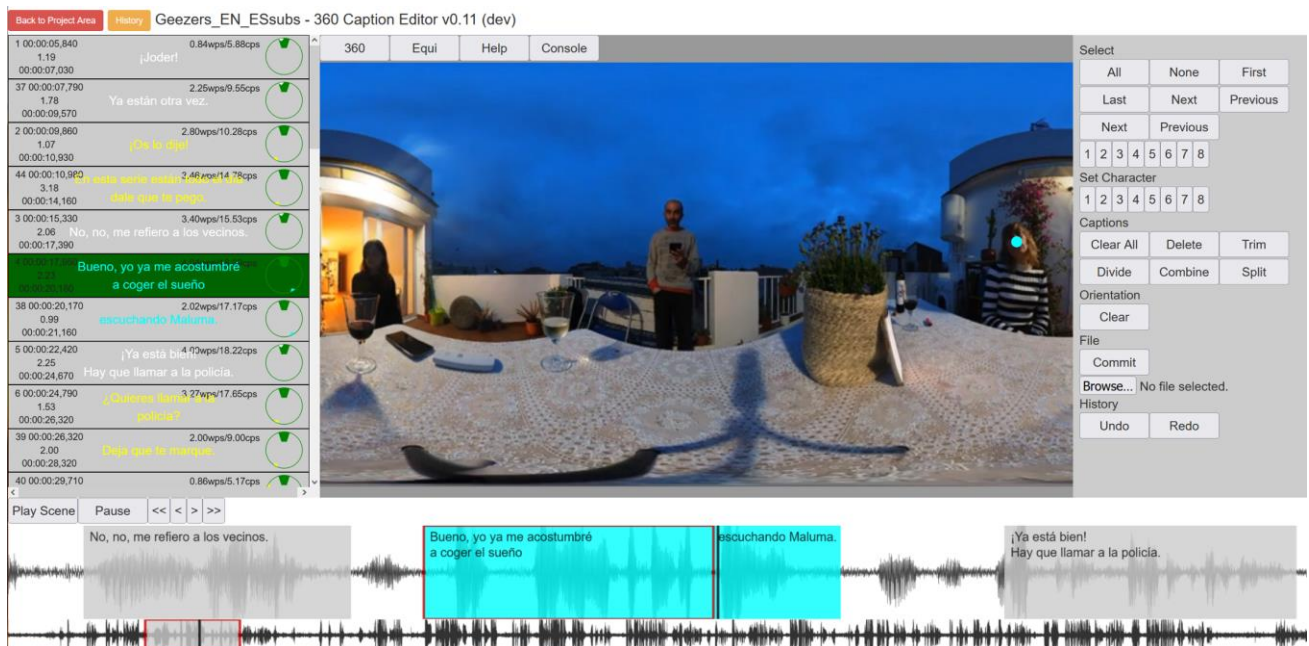


Figure 4.

Open-source editor developed as part of this study

Both the demo player and the editor are open-source and can be accessed from a main project area, where all the imported 360° videos are located. These tools take inspiration from the player and the editor developed by 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 360° video, the live web testing framework developed by Hughes et al. (2020a, 2020b) was ported to Unity 3D to allow display of 360° 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.

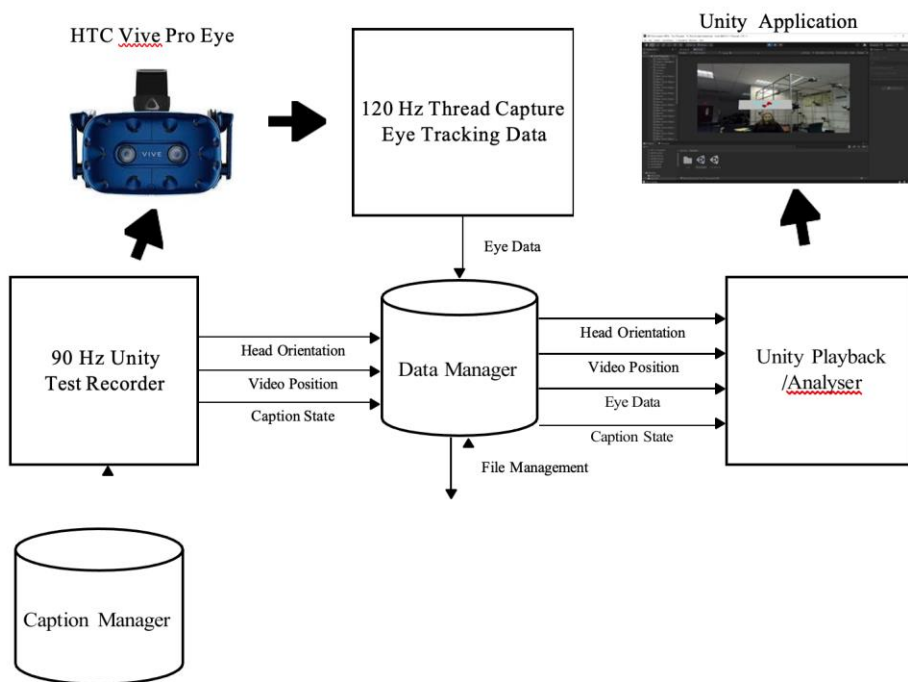


Figure 5.

Eye-tracking VR system architecture

The system architecture was developed to utilize the HTC Vive Pro Eye, which contains an eye tracker from Tobii built into the display. The application uses two Unity assets: one specifically optimized 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 360° video to be played with the captions fixed in the scene. During the test, each event and data is logged into the data manager as it becomes 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 affords the opportunity to change the analysis process or include additional

Areas Of Interest (AOIs) and the analysis repeated. It also allows for visual analysis by overlaying the eye data onto the video following capture.

One technical difficulty that had to be overcome was synchronization 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 synchronization 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 the 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 Virtual Reality 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.

3.3. Participants

The size of the group was decided according to the pilot nature of the study (Bryman, 2004, p. 507). In the beginning, 7 participants were expected, but due to complications derived from 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 completed M.A. university education. All were active professionally (1 accompany to mothers, 1 cultural manager, 1 music therapist, 1 pre-doctoral researcher, 1 project manager). All participants spoke Spanish and at least one of the other modern languages.

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

3.4. Study materials

One of the main concerns of the study was to find appropriate material for testing. Due to the difficulty of finding royalty-free material that meets the needs of the study, a homemade 360° 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 360° space. The characters followed a script to avoid overlaps as if two characters located at different points in the 360° scene speak 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 color (b&w) for all the characters, with a grey background and head-locked.
- Mode 2: following Fox 2018 studies. Same font and color (b&w) for all the characters, without grey back- ground and near to the mouth.
- Mode 3: fully custom. Different font and colour for each character, with a grey background and near to the mouth.

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 different 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-tacker lab was settled with a computer and a pair of HTC Vive Pro HMD. One re- searcher took notes and summarised the conclusions in real-time. Secondly, the aim of the focus group was explained to the participants, and they were asked to sign informed con- sent forms. The third step was 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 360° content and explained how subtitles are integrated within 360° 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 360° content can also be accessed on a flat TV screen using a mouse to move around the 360° 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, one per video. The order of the participant watching each video 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).

Once the last round finished, the focus group took place. Together with the stimuli, the facilitator used a list of guiding questions grouped under major topics to generate 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 pilot 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 custom 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. Also, another participant commented that sometimes it was difficult to follow the story because she was distracted exploring the 360° scene. The participant who was less used to new technologies (and also less 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. Subtitles readability

Participants were asked whether they were able to read subtitles after watching each movie. Two responded 'yes' and two 'no' for mode 1. In mode 2, two responded 'yes' and 3 'no'. The least readable subtitles seemed to be in mode 3. Three participants noted they were not able to read them and only one responded 'yes' and one participant was 'not sure'. When asked to estimate the

percentage of subtitles that they (participants) 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 towards the subtitles in mode 1 as the most readable, possibly because mode 1 subtitles are most similar to what participants were familiar with.

These results comply with the qualitative data extracted from the focus group, since the majority of participants agreed that mode 3 was difficult to read. Just one of the participants noted that she liked the coloured text, and there was a brief discussion about the possibility of customising the subtitles even 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. Another participant highlighted the reading pace, arguing that some captions disappeared 'too soon' forcing the user to read faster.

4.3. Self-reported task load

To collect self-reports on effort evoked 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 preference towards mode 1 ($M = 5.20$, $SD = 2.39$). In their 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 demand 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 with greater perceived success modes 1 ($M = 6.40$, $SD = 1.52$) and 2 ($M = 6.60$, $SD = 2.30$) than mode 3 ($M = 5.6$, $SD = 2.19$). Results are not surprising when looking at average responses regarding how time pressured participants felt. Subtitles in mode 2 evoked the highest time pressure ($M = 7.60$, $SD = 0.89$), lower in mode 3 ($M = 7.00$, $SD = 0.71$), and least in mode 1 ($M = 5.80$, $SD = 2.17$). The perceived level of frustration/stress was lowest when watching 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 ($M = 5.00$, $SD = 2.35$).

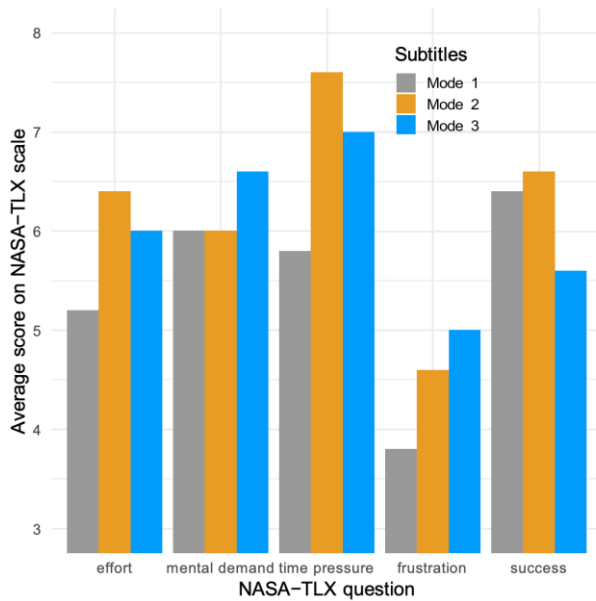


Figure 7.

Task load self-reports with NASA-TLX scale while watching videos in different subtitle modes

4.4. Attention distribution and cognitive effort while reading captions and scene viewing

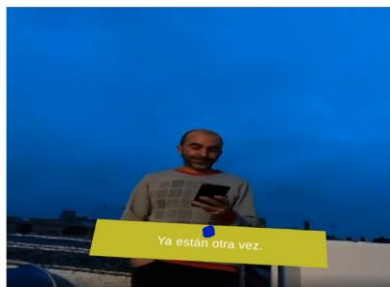
Gaze was captured as it traversed subtitles when reading the text displayed within the quadrilaterals that contained them. Analysis of the eye movement signal relies on fixation detection, which in turn depends on saccade detection. Fixations are detected within the raw eye movement signal following 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.

(a) Actor body AOI



(b) Subtitle quad AOI



(c) Subtitle word AOI

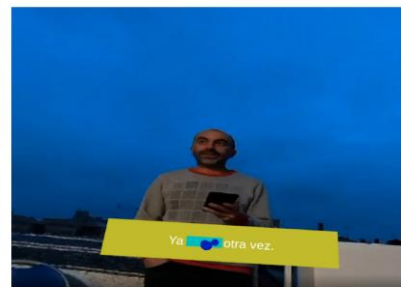


Figure 8.

Gaze recording showing fixations over Areas Of Interest: (a) actor body, (b) subtitle quad, and (c) individual word

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 smallest for 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.

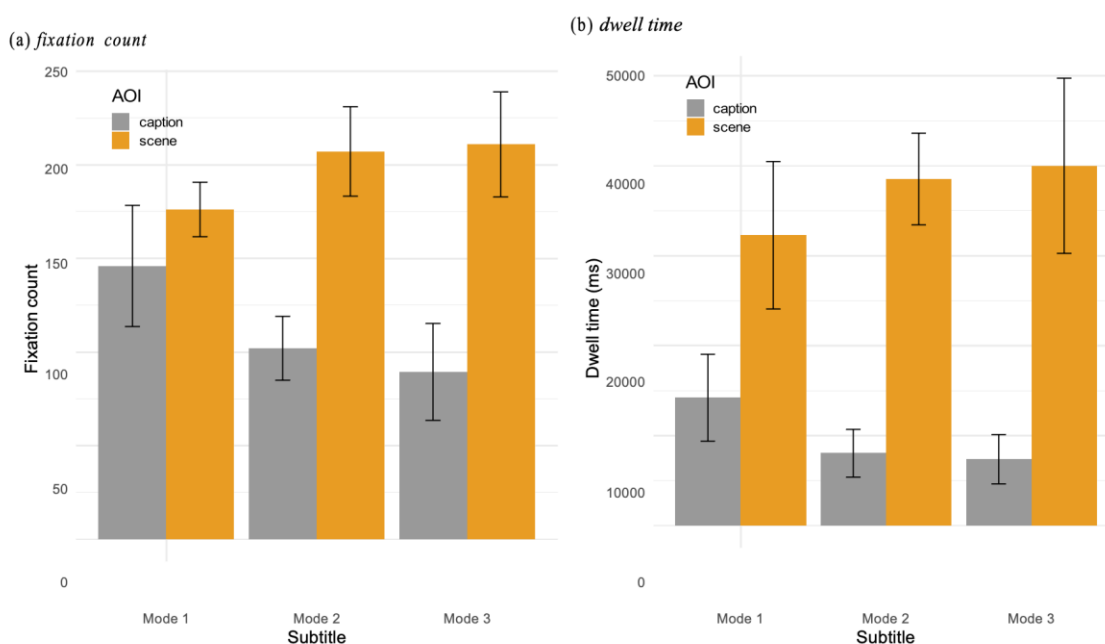


Figure 9.

Visual attention distribution over captions and visual scenes while watching video with different types of subtitles. 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. Note: bars height represents mean values and whiskers represent ± 1 SD

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 visual scene and 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 highest cognitive effort along with the less focal processing when processing subtitles in mode 1.

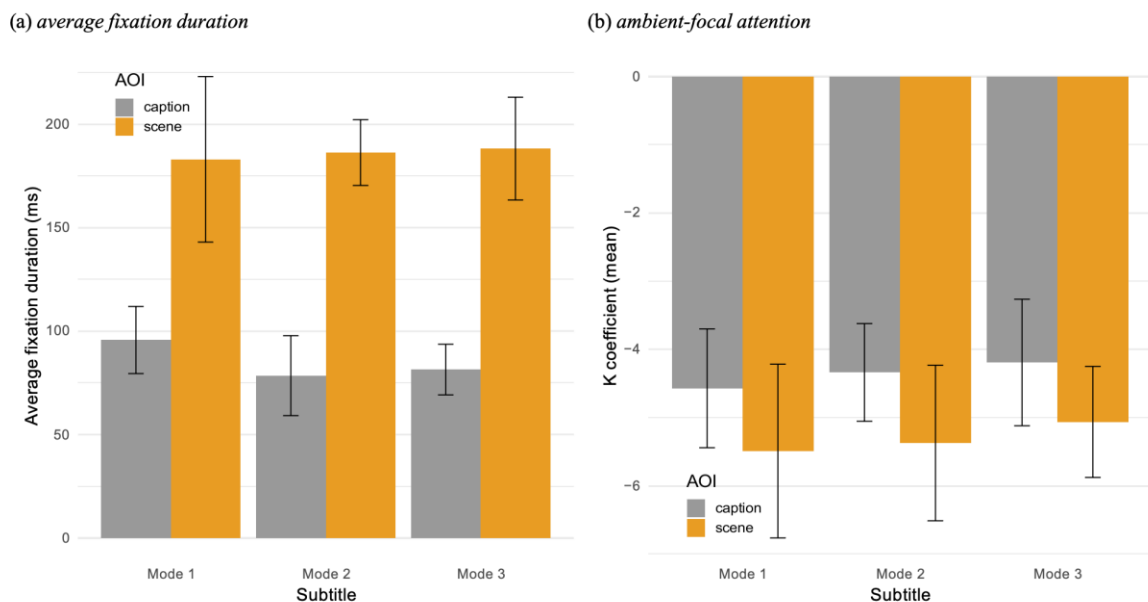


Figure 10.

Cognitive processing of textual and visual information from captions and visual scenes while watching video with different types of subtitles: (a) shows average fixation duration as a metric for cognitive effort, (b) shows K coefficient as a metric ambient-focal attention. Note: bars height represents mean values and whiskers represent ± 1 SD

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 watched the video with the different subtitle options). The notes were thoroughly revised and tagged using Atlas.ti. This procedure allowed to identify three areas that can be associated with the quantitative analysis (subtitle readability, task load, and movie content

understanding). The analysis also allowed defining user preferences, as well as defining aspects in which there was consensus among users and aspects in which opinions diverged.

In terms of preference, most of the participants agreed that mode 2 was the easiest to read. Also, another participant suggested 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 be the difficulty identifying the character speaking in each moment.

Regarding creative subtitles, all the participants agreed that it is a great idea to dramatise what is said and can add plenty of visual beauty to the content. Just 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 also felt isolated when wearing the HMD for the first time.

5. Discussion

The first hypothesis we wanted to validate was if fixed, near to the mouth subtitles allow viewers to spend more time exploring the image instead of reading the subtitles than 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 near to 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 on 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) towards 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. Results may also be biased by a lack of randomization in 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 randomization or counterbalancing).

Eye movement analysis sheds a 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 the less focal processing of subtitles in mode 1.

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

Visualization of analyzed eye movements, 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 capture 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 (rows: participant 1-5, columns: mode 1-3, 00:22). General observations can be drawn, such as participant 1 although finding the captions in mode 1 and 3 was lost in mode 2 and can be observed saccading between the mouths of the wrong characters trying to identify the speaking character. 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 mode 1 and 2, but not mode 3

6. Conclusions

We presented a framework for subtitles construction in 360° movies shown in Virtual Reality along with pilot testing of three subtitling modes. Our contribution is thus two-fold, presentation of the VR subtitle framework and a new method of triangulation of psycho-physiological (eye movements) self-reports and qualitative (focus groups discussions) analyses. To our best knowledge, this is the first attempt to advance these two directions when discussing subtitles construction in VR 360° videos.

Immersive environments are in need of new subtitle presentation modes. 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, a Virtual Reality system designed specifically to collect eye movement data as visual attention is distributed over 360° videos containing subtitles.

A pilot study using the system highlighted features of a methodology that can be used to collect quantitative and qualitative behavioural data when viewing subtitled 360° media. Future studies are expected to yield new insights and lead toward subtitle standardisation.

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.). John Benjamins. <https://doi.org/10.1075/btl.56.08bar>
- Brescia-Zapata, M. (forthcoming). The present and future of accessibility services in vr360 players. *inTRAlinea*.
- Brown, A. (2017). User testing subtitles for 360° content. <https://www.bbc.co.uk/rd/blog/2017-10-subtitles-360-video-virtual-reality-vr>
- Brown, A., & Patterson, J. (2017). Designing subtitles for 360° content. <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*, 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 360° en las estrategias transmedia de las televisiones autonómicas españolas. *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 Modeling 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*, 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*, 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*. 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 Attention with Coefficient 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*, 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>
- 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*, 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. <https://doi.org/10.2312/egve.20181307>
- 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*, 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, 603–616. <https://doi.org/10.1162/pres.1997.6.6.603>
- Smith, W. (2015). Stop calling google cardboard's 360-degree video 'vr'. *Wired*. <https://www.wired.com/2015/11/360-video-isnt-virtual-reality/>